# Statistical Modelling

Proceedings of the

15th International Workshop on Statistical Modelling

New Trends in Statistical Modelling

eman ta zabal zazu

**Universidad del País Vasco** **Euskal Herriko Unibertsitatea**

Bilbao, Spain, July 17-21, 2000

V. Núñez-Antón, E. Ferreira
(Editors)

Vicente Núñez-Antón, Eva Ferreira (Eds.)

# Statistical Modelling

Proceedings of the
15th International Workshop on Statistical Modelling
Bilbao, Spain, July 17-21, 2000

## Editors:

Vicente Núñez-Antón
Departamento de Econometría y Estadística (E.A. III)
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avenida Lehendakari Aguirre, 83
48015 Bilbao
Spain

Eva Ferreira
Departamento de Econometría y Estadística (E.A. III)
Facultad de Ciencias Económicas y Empresariales
Universidad del País Vasco
Avenida Lehendakari Aguirre, 83
48015 Bilbao
Spain

# Preface

This proceedings volume contains the papers presented at the 15th International Workshop on Statistical Modelling held in Bilbao, Spain, July 17-21, 2000.

For 15 years the International Workshop on Statistical Modelling has brought together statisticians from all corners of the globe. Originally founded in Innsbruck, Austria, in 1986, the workshop took place in Perugia, Italy (1987), Vienna, Austria (1988), Trento, Italy (1989), Toulouse, France (1990), Utrecht, Netherlands (1991), Munich, Germany (1992), Leuven, Belgium (1993), Exeter, UK (1994), Innsbruck (1995), Orvieto, Italy (1996) and Biel/Bienne, Switzerland (1997). In 1998, the workshop location left Europe and went to New Orleans, USA. Last year's location (Graz, Austria) continued the European tradition of the workshop without losing its international orientation. This year, the workshop comes to Spain for the first time and takes place in Bilbao, a city in the north part of Spain in the region of the Basque Country. The number of participants has grown continuously in the past years. Starting from 40 in 1986 it reached about 150 participants in the workshop celebrated in 1998 and it broke a new frontier by attracting more than 200 participants in the last workshop. This year we have about 160 participants, and around 100 scientists from nearly 30 different countries are presenting their work.

The workshop provides an interesting and stimulating scientific forum combined with an informal atmosphere that allows for discussion and exchange of ideas, paying special attention to the possibility of interaction between well known scientists and young ones. As in previous years, the meeting focuses on various aspects of statistical modelling including theoretical developments, applications and computational methods. It is the tradition of the workshop to have no parallel sessions, which guarantees that the audience is not split. This stimulates a family-like atmosphere and encourages people to return to the workshop every year. However, with the large number of excellent papers submitted this year the Scientific Committee had a difficult job to select 31 of them for oral presentation. Another focus of the workshop is to motivate young statisticians to present their work in session specially designed for students. This year the members of the Scientific Committee were deeply impressed by the solid and valuable work submitted by students. Five papers were selected for oral presentation, namely those by David Allcroft, María Jesús Bárcena, Rami Bustami, Renato Flores and Stefan Lang.

The organizers of the workshop are proud and grateful for attracting 9 invited speakers to come to the meeting. The topics discussed by the invited speakers further develop and extend the area of statistical modelling. Wenceslao González Manteiga gives a tutorial on goodness of fit tests for regression models. Joel Horowitz considers semiparametric and nonparametric estimation in Econometrics.

Mark Steel (joint work with Carmen Fernández and Gary Koop) investigates problems with modelling production with undesirable outputs. Dale L. Zimmerman deals with the issue of aids for modelling the covariance structure of longitudinal data: alternative specifications and graphical diagnostics. Eric Renault (Iterative and recursive inference for options price data) and Winfried Stute (Jump diffusion processes with shot noise effects and their applications to Finance) introduce the issue of statistical modelling in Finance for the first time in the workshop. Johannes Ledolter brings out the topic of statistical techniques for quality improvement: improving the manufacture of viscose fibre. James Zidek investigates the issue of combining statistical and computer models for risk assesment (risk exposure). Christopher Bishop (joint work with Michael Tipping) deals with the issue of variational relevance vector machines.

The editors are indebted to Ludwig Fahrmeir, John Hinde, Michel Mouchart, Jean Opsomer, Juan Romo, Esther Ruiz-Ortega and Bill Venables for their effort and reliability as members of the Scientific Committee. Our special thanks, however, are dedicated to all authors included in this volume for carefully preparing their manuscript. We will also like to thank Fernando Tusell Palmer and Juan Ignacio Modroño Herrán for keeping the web page for the workshop up to date. We must also thank Fernando Tusell Palmer for allowing us to use his student's Alpha machine and for all the help he gave us when we had problems with LaTeX. It is also our duty to thank both Herwig Friedl and Göran Kauermann, the previous organizers of the workshop, for all the help, support and hints they have been giving us during this past year. We must also mention that it has been our pleasure to work together with the secretary of the workshop, John Hinde, and we thank him for being so patient with us and for being so easy to communicate with.

Finally we hope, that all participants and readers can join us with our enthusiasm and delight at the workshop.

Vicente and Eva
Bilbao, May 25, 2000

# Contents

# Variational Relevance Vector Machines

Christopher M. Bishop and Michael E. Tipping

[1] Microsoft Research, St. George House, 1 Guildhall Street, Cambridge CB2 3NH, U.K. , {cmbishop,mtipping}@microsoft.com

**Abstract:** The Support Vector Machine (SVM) of Vapnik (1998) has become widely established as one of the leading approaches to pattern recognition and machine learning. It expresses predictions in terms of a linear combination of kernel functions centred on a subset of the training data, known as support vectors.
Despite its widespread success, the SVM suffers from some important limitations, one of the most significant being that it makes point predictions rather than generating predictive distributions. Recently Tipping (2000) has formulated the Relevance Vector Machine (RVM), a probabilistic model whose functional form is equivalent to the SVM. It achieves comparable recognition accuracy to the SVM, yet provides a full predictive distribution, and also requires substantially fewer kernel functions.
The original treatment of the RVM relied on the use of type II maximum likelihood (the 'evidence framework') to provide point estimates of the hyperparameters which govern model sparsity. In this paper we show how the RVM can be formulated and solved within a completely Bayesian paradigm through the use of variational inference, thereby giving a posterior distribution over both parameters and hyperparameters. We demonstrate the practicality and performance of the variational RVM using both synthetic and real world examples.

## 1 Relevance Vectors

Many problems in machine learning fall under the heading of supervized learning, in which we are given a set of input vectors $X = \{\mathbf{x}_n\}_{n=1}^N$ together with corresponding target values $T = \{t_n\}_{n=1}^N$. The goal is to use this training data, together with any pertinent prior knowledge, to make predictions of $t$ for new values of $\mathbf{x}$. We can distinguish two distinct cases: *regression*, in which $t$ is a continuous variable, and *classification*, in which $t$ belongs to a discrete set.
Here we consider models in which the prediction $y(\mathbf{x}, \mathbf{w})$ is expressed as a linear combination of basis functions $\phi_m(\mathbf{x})$ of the form

$$y(\mathbf{x}, \mathbf{w}) = \sum_{m=0}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi} \tag{1}$$

where the $\{w_m\}$ are the parameters of the model, and are generally called *weights*.

One of the most popular approaches to machine learning to emerge in recent years is the Support Vector Machine (SVM) of Vapnik (1998). The SVM uses a particular specialization of (1) in which the basis functions take the form of *kernel* functions, one for each data point $\mathbf{x}_m$ in the training set, so that $\phi_m(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_m)$, where $K(\cdot, \cdot)$ is the kernel function. The framework which we develop in this paper is much more general and applies to any model of the form (1). However, in order to facilitate direct comparisions with the SVM, we focus primarily on the use of kernels as the basis functions.

Point estimates for the weights are determined in the SVM by optimization of a criterion which simultaneously attempts to fit the training data while at the same time minimizing the 'complexity' of the function $y(\mathbf{x}, \mathbf{w})$. The result is that some proportion of the weights are set to zero, leading to a sparse model in which predictions, governed by (1), depend only on a subset of the kernel functions.

The SVM framework is found to yield good predictive performance for a broad range of practical applications, and is widely regarded as the state of the art in pattern recognition. However, the SVM suffers from some important drawbacks. Perhaps the most significant of these is that it is a non-Bayesian approach which makes explicit classifications (or point predictions in the case of regression) for new inputs. As is well known, there are numerous advantages to predicting the posterior probability of class membership (or a predictive conditional distribution in the case of regression). These include the optimal compensation for skewed loss matrices or unequal class distributions, the opportunity to improve performance by rejection of the more ambiguous examples, and the fusion of outputs with other probabilistic sources information before applying decision criteria.

Recently Tipping (2000) introduced the *Relevance Vector Machine* (RVM) which makes probabilistic predictions and yet which retains the excellent predictive performance of the support vector machine. It also preserves the sparseness property of the SVM. Indeed, for a wide variety of test problems it actually leads to models which are dramatically sparser than the corresponding SVM, while sacrificing little if anything in the accuracy of prediction.

For regression problems, the RVM models the conditional distribution of the target variable, given an input vector $\mathbf{x}$, as a Gaussian distribution of the form

$$P(t|\mathbf{x}, \mathbf{w}, \tau) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \tau^{-1}) \tag{2}$$

where we use $\mathcal{N}(\mathbf{z}|\mathbf{m}, \mathbf{S})$ to denote a multi-variate Gaussian distribution over $\mathbf{z}$ with mean $\mathbf{m}$ and covariance $\mathbf{S}$. In (2) $\tau$ is the inverse 'noise' parameter, and the conditional mean $y(\mathbf{x}, \mathbf{w})$ is given by (1). Assuming an

independent, identically distributed data set $X = \{\mathbf{x}_n\}$, $T = \{t_n\}$ the likelihood function can be written

$$P(T|X, \mathbf{w}, \tau) = \prod_{n=1}^{N} P(t_n|\mathbf{x}_n, \mathbf{w}, \tau). \tag{3}$$

The parameters $\mathbf{w}$ are given a Gaussian prior

$$P(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{m=0}^{N} \mathcal{N}(w_m|0, \alpha_m^{-1}) \tag{4}$$

where $\boldsymbol{\alpha} = \{\alpha_m\}$ is a vector of hyperparameters, with one hyperparameter $\alpha_m$ assigned to each model parameter $w_m$. In the original RVM of Tipping (2000) values for these hyperparameters are estimated using the framework of type-II maximum likelihood, Berger (1985), in which the marginal likelihood $P(T|X, \boldsymbol{\alpha}, \tau)$ is maximized with respect to $\boldsymbol{\alpha}$ and $\tau$. Evaluation of this marginal likelihood requires integration over the model parameters

$$P(T|X, \boldsymbol{\alpha}, \tau) = \int P(T|X, \mathbf{w}, \tau) P(\mathbf{w}|\boldsymbol{\alpha}) \mathrm{d}\mathbf{w}. \tag{5}$$

Since this involves the convolution of two exponential-quadratic functions the integration can be performed analytically, giving

$$P(T|X, \boldsymbol{\alpha}, \tau) = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{S}) \tag{6}$$

where $\mathbf{t} = (t_1, \ldots, t_N)$ and

$$\mathbf{S} = \tau^{-1}\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T \tag{7}$$

in which $\mathbf{I}$ is the $N \times N$ unit matrix, $\mathbf{A} = \mathrm{diag}(\alpha_m)$, and $\boldsymbol{\Phi}$ is the $N \times (N+1)$ *design matrix* with columns $\boldsymbol{\phi}_m$, so that $(\boldsymbol{\Phi})_{nm} = \phi(\mathbf{x}_n; \mathbf{x}_m)$. Maximization of (6) with respect to the $\{\alpha_m\}$ can be performed efficiently using an iterative re-estimation procedure obtained by setting the derivatives of the marginal log likelihood to zero. During the process of this optimization many of the $\alpha_m$ are driven to large values, so that the corresponding model parameters $w_m$ are effectively pruned out. The corresponding terms can be omitted from the trained model represented by (1), with the training data vectors $\mathbf{x}_n$ associated with the remaining kernel functions being termed 'relevance vectors'. Insight into this pruning process is given in Section 3. A similar re-estimation procedure is used to optimize $\tau$ simultaneously with the $\alpha_m$ parameters.

In the classification version of the relevance vector machine the conditional distribution of targets is given by

$$P(t|\mathbf{x}, \mathbf{w}) = \sigma(y)^t [1 - \sigma(y)]^{1-t} \tag{8}$$

where $\sigma(y) = (1 + \exp(-y))^{-1}$ and $y(\mathbf{x}, \mathbf{w})$ is given by (1). Here we confine attention to the case $t \in \{0, 1\}$. Assuming independent, identically distributed data, we obtain the likelihood function in the form

$$P(T|X, \mathbf{w}) = \prod_{n=1}^{N} \sigma(y_n)^{t_n} [1 - \sigma(y_n)]^{1-t_n}. \tag{9}$$

As before, the prior over the weights takes the form (4). However, the integration required by (5) in order to evaluate the marginal likelihood can no longer be performed analytically. Tipping (2000) therefore used a local Gaussian approximation to the posterior distribution of the weights. Optimization of the hyperparameters can then be performed using a re-estimation framework, alternating with re-evaluation of the mode of the posterior, until convergence.

As we have seen, the standard relevance vector machine of Tipping (2000) estimates point values for the hyperparameters. In this paper we seek a more complete Bayesian treatment of the RVM through exploitation of variational methods.

## 2 VARIATIONAL INFERENCE

In a general probabilistic model we can partition the stochastic variables into those corresponding to the observed data, denoted $D$, and the remaining unobserved variables denoted $\boldsymbol{\theta}$. The marginal probability of the observed data (the model 'evidence') is obtained by integrating over $\boldsymbol{\theta}$

$$P(D) = \int P(D, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}. \tag{10}$$

This integration will, for almost any non-trivial model, be analytically intractable. Variational methods, Jordan (1998), address this problem by introducing a distribution $Q(\boldsymbol{\theta})$, which (for arbitrary choice of $Q$) allows the marginal log likelihood to be decomposed into two terms, see Neal (1998).

$$\ln P(D) = \mathcal{L}(Q) + \mathrm{KL}(Q \| P) \tag{11}$$

where

$$\mathcal{L} = \int Q(\boldsymbol{\theta}) \ln \frac{P(D, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \tag{12}$$

and $\mathrm{KL}(Q \| P)$ is the Kullback-Leibler divergence between $Q(\boldsymbol{\theta})$ and the posterior distribution $P(\boldsymbol{\theta}|D)$, and is given by

$$\mathrm{KL}(Q \| P) = -\int Q(\boldsymbol{\theta}) \ln \frac{P(\boldsymbol{\theta}|D)}{Q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta}. \tag{13}$$

Since $\mathrm{KL}(Q\|P) \geq 0$, it follows that $\mathcal{L}(Q)$ is a rigorous lower bound on $\ln P(D)$. Furthermore, since the left hand side of (11) is independent of $Q$, maximizing $\mathcal{L}(Q)$ is equivalent to minimizing $\mathrm{KL}(Q\|P)$, and therefore $Q(\boldsymbol{\theta})$ represents an approximation to the posterior distribution $P(\boldsymbol{\theta}|D)$.
The significance of this transformation is that, for a suitable choice for the $Q$ distribution, the quantity $\mathcal{L}(Q)$ may be tractable to compute, even though the original model evidence function is not. The goal in a variational approach is therefore to choose a suitable form for $Q(\boldsymbol{\theta})$ which is sufficiently simple that the lower bound $\mathcal{L}(Q)$ can readily be evaluated and yet which is sufficiently flexible that the bound is reasonably tight. In practice we choose some family of $Q$ distributions and then seek the best approximation within this family by maximizing the lower bound with respect to $Q$. One approach would be to assume some specific parameterized functional form for $Q$ and then to optimize $\mathcal{L}$ with respect to the parameters of the distribution. Here we adopt an alternative procedure, following Waterhouse (1996), and consider a factorized form over the component variables $\{\theta_i\}$ in $\boldsymbol{\theta}$, so that

$$Q(\boldsymbol{\theta}) = \prod_i Q_i(\theta_i). \tag{14}$$

The lower bound can then be maximized over all possible factorial distributions by performing a *free-form* maximization over the $Q_i$, leading to the following result

$$Q_i(\theta_i) = \frac{\exp\left\langle \ln P(D,\boldsymbol{\theta})\right\rangle_{k\neq i}}{\int \exp\left\langle \ln P(D,\boldsymbol{\theta})\right\rangle_{k\neq i} \mathrm{d}\theta_i} \tag{15}$$

where $\langle\ \cdot\ \rangle_{k\neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k \neq i$. It is easily shown that, if the probabilistic model is expressed as a directed acyclic graph with a node for each of the factors $Q_i(\theta_i)$, then the solution for $Q_i(\theta_i)$ depends only on the $Q$ distributions for variables which are in the Markov blanket of the node $i$ in the graph.
Note that (15) represents an implicit solution for the factors $Q_i(\theta_i)$ since the right hand side depends on moments with respect to the $Q_{k\neq i}$. For conjugate conditional distributions (e.g. linear-Gaussian models with Gamma priors, in the case of continuous variables) this leads to standard distributions for which the required moments are easily evaluated. We can then find a solution iteratively by initializing the moments and then cycling through the variables updating each distribution in turn using (15).

## 3 CONTROLLING COMPLEXITY

The Relevance Vector framework provides a means for solving regression and classification problems in which we seek models which are highly sparse by selecting a subset from a larger pool of candidate kernel functions (one

for each example in the training set). A key concept is the use of continuous hyperparameters to govern model complexity and thereby avoid the intractable problem of searching over an exponentially large discrete space of model structures. This approach, based on a hierarchical prior, was successfully used to find the optimal number of principal components in a Bayesian treatment of PCA, see Bishop (1999).
A conventional way to remove superfluous parameters is to use a 'pruning' prior given by a Laplace distribution of the form

$$P(w) = \lambda \exp(-\lambda |w|). \tag{16}$$

Unfortunately, such a choice of prior does not lead to a tractable variational treatment, since the corresponding variational solution given by (15) cannot be evaluated analytically.
Here we propose an alternative framework based on a hierarchical prior of the form

$$P(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}) \tag{17}$$

as discussed previously, in which we use a hyperprior given by

$$P(\alpha) = \Gamma(\alpha|a, b) \equiv b^a \alpha^{a-1} e^{-b\alpha} / \Gamma(a) \tag{18}$$

where $\Gamma(a)$ is the Gamma function. The distribution (18) has the useful properties

$$\langle \alpha \rangle = a/b, \qquad \langle \alpha^2 \rangle - \langle \alpha \rangle^2 = a/b^2. \tag{19}$$

The marginal distribution of $w$ (a t-distribution) is then obtained by integrating over $\alpha$. A comparison of this marginal distribution, for $a = b = 1$, with the Laplace distribution (16) is shown in Figure 1. The key observation is that the variational framework can be rendered tractable by working not directly with the marginal distribution $P(w)$ but instead leaving the hierarchical conjugate form explicit and introducing a factorial representation given by $Q(w, \alpha) = Q(w)Q(\alpha)$. A further advantage of this approach is that it becomes possible to evaluate the lower bound $\mathcal{L}$ as a closed-form analytic expression. This is useful for monitoring the convergence of the iterative optimization and also for checking the accuracy of the software implementation (by verifying that none of the updates to the variational distributions lead to a decrease the value of $\mathcal{L}$). It can also be used to compare models (without resorting to a separate validation set) since it represents an approximation to the model evidence. We now exploit these ideas in the context of the Relevance Vector Machine.

## 4 RVM Regression

Following the concepts developed in the previous section, we augment the standard relevance vector machine by the introduction of hyperpriors given by a separate distribution for each hyperparameter $\alpha_m$ of the

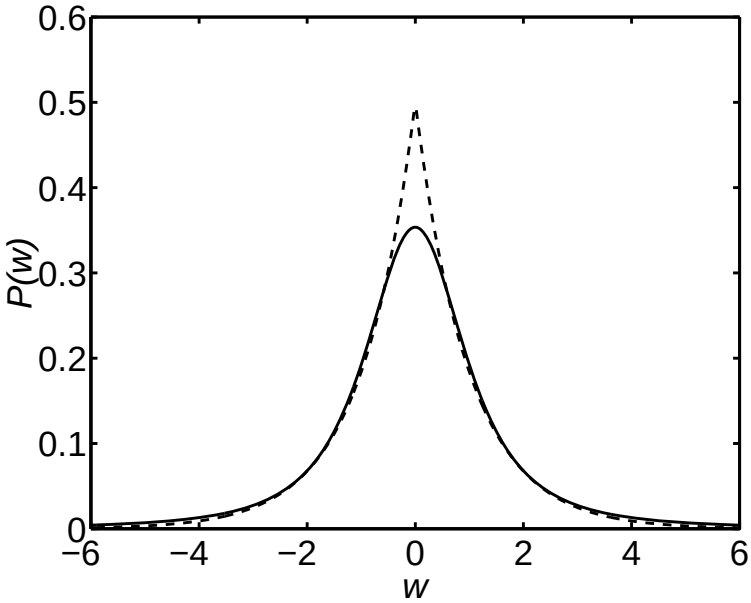



FIGURE 1. Comparison of the marginal distribution defined by the hierarchical model $P(w) = \int P(w|\alpha)P(\alpha)\mathrm{d}\alpha$ (solid line), compared to the Laplace distribution (dotted line).

form $P(\alpha_m) = \Gamma(\alpha_m|a,b)$. Similarly, we introduce a prior over the inverse noise variance $\tau$ given by $P(\tau) = \Gamma(\tau|c,d)$. We obtain broad hyperpriors by setting $a = b = c = d = 10^{-6}$. Together with the likelihood function (3) and the weight prior (4) we now have a complete probabilistic specification of the model. The probabilistic model can also be represented as a directed graph, as shown in Figure 2. Next we consider a factorial approximation to the posterior distribution $P(\mathbf{w}, \boldsymbol{\alpha}, \tau|X,T)$ given by $Q(\mathbf{w}, \boldsymbol{\alpha}, \tau) = Q_{\mathbf{w}}(\mathbf{w})Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})Q_{\tau}(\tau)$. Due to the conjugacy properties of the chosen distributions we can evaluate the general solution (15) analytically, giving

$$Q_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{\mathbf{w}}, \boldsymbol{\Sigma}_{\mathbf{w}}) \tag{20}$$

$$Q_{\tau}(\tau) = \Gamma(\tau|\widetilde{c}, \widetilde{d}) \tag{21}$$

$$Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{m=0}^{N} \Gamma(\alpha_m|\widetilde{a}_m, \widetilde{b}_m) \tag{22}$$

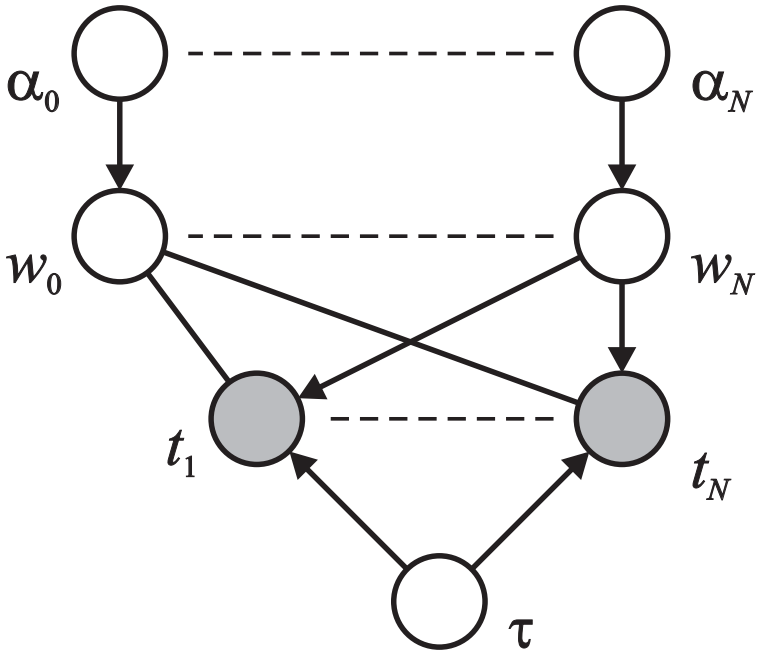



FIGURE 2. Directed acyclic graph representing the variational RVM as used for regression. The classification version is the same, with the omission of the $\tau$ node.

where

$$\boldsymbol{\Sigma}_{\mathbf{w}} = \left( \mathrm{diag}\langle \alpha_m \rangle + \langle \tau \rangle \sum_{n=1}^{N} \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T \right)^{-1} \tag{23}$$

$$\boldsymbol{\mu}_{\mathbf{w}} = \langle \tau \rangle \boldsymbol{\Sigma}_{\mathbf{w}} \sum_{n=1}^{N} \boldsymbol{\phi}_n t_n \tag{24}$$

$$\widetilde{a}_m = a + 1/2 \qquad \widetilde{b}_m = b + \langle w_m^2 \rangle / 2 \tag{25}$$

$$\widetilde{c} = c + (N+1)/2 \tag{26}$$

$$\begin{aligned} \widetilde{d} &= d + \frac{1}{2} \sum_{n=1}^{N} t_n^2 - \langle \mathbf{w} \rangle^T \sum_{n=1}^{N} \boldsymbol{\phi}_n t_n \\ &\quad + \frac{1}{2} \sum_{n=1}^{N} \boldsymbol{\phi}_n^T \langle \mathbf{w}\mathbf{w}^T \rangle \boldsymbol{\phi}_n . \end{aligned} \tag{27}$$

The required moments are easily evaluated using the following results

$$\begin{aligned}
\langle \mathbf{w} \rangle &= \boldsymbol{\mu}_{\mathbf{w}} && (28)\\
\langle \mathbf{w}\mathbf{w}^T \rangle &= \boldsymbol{\Sigma}_{\mathbf{w}} + \boldsymbol{\mu}_{\mathbf{w}}\boldsymbol{\mu}_{\mathbf{w}}^T && (29)\\
\langle \alpha_m \rangle &= \widetilde{a}_m / \widetilde{b}_m && (30)\\
\langle \ln \alpha_m \rangle &= \psi(\widetilde{a}_m) - \ln \widetilde{b}_m && (31)\\
\langle \tau \rangle &= \widetilde{c}/\widetilde{d} && (32)\\
\langle \ln \tau \rangle &= \psi(\widetilde{c}) - \ln \widetilde{d} && (33)
\end{aligned}$$

where the $\psi$ function is defined by

$$\psi(a) = \frac{d}{da} \ln \Gamma(a). \tag{34}$$

The full predictive distribution $P(t|\mathbf{x}, X, T)$ is given by

$$P(t|\mathbf{x}, X, T) = \int\int P(t|\mathbf{x}, \mathbf{w}, \tau) P(\mathbf{w}, \tau|X, T) \mathrm{d}\mathbf{w}\mathrm{d}\tau. \tag{35}$$

In the variational framework we replace the true posterior $P(\mathbf{w}, \tau|X, T)$ by its variational approximation $Q_{\mathbf{w}}(\mathbf{w})Q_\tau(\tau)$. Integration over both $\mathbf{w}$ and $\tau$ is intractable. However, as the number of data points increases the distribution of $\tau$ becomes tightly concentrated around its mean value. To see this we note that the variance of $\tau$ is given, from (19), by $\langle \tau^2 \rangle - \langle \tau \rangle^2 = \widetilde{c}/\widetilde{d}^2 \sim O(1/N)$ for large $N$. Thus we can approximate the predictive distribution using

$$P(t|\mathbf{x}, X, T) = \int P(t|\mathbf{x}, \mathbf{w}, \langle \tau \rangle) Q_{\mathbf{w}}(\mathbf{w}) \mathrm{d}\mathbf{w} \tag{36}$$

which is the convolution of two Gaussian distributions. Using (2) and (20) we then obtain

$$P(t|\mathbf{x}, X, T) = \mathcal{N}(t|\boldsymbol{\mu}_{\mathbf{w}}^T \boldsymbol{\phi}(\mathbf{x}), \sigma^2) \tag{37}$$

where the input-dependent variance is given by

$$\sigma^2(\mathbf{x}) = \frac{1}{\langle \tau \rangle} + \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\Sigma}_{\mathbf{w}} \boldsymbol{\phi}(\mathbf{x}). \tag{38}$$

We can also evaluate the lower bound $\mathcal{L}$, given by (12), which in this case takes the form

$$\begin{aligned}
\mathcal{L} = \; & \langle \ln P(T|X, \mathbf{w}, \tau) \rangle + \langle \ln P(\mathbf{w}|\boldsymbol{\alpha}) \rangle \\
& + \langle \ln P(\boldsymbol{\alpha}) \rangle + \langle \ln P(\tau) \rangle - \langle \ln Q_{\mathbf{w}}(\mathbf{w}) \rangle \\
& - \langle \ln Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \rangle - \langle \ln Q_\tau(\tau) \rangle \qquad (39)
\end{aligned}$$

in which

$$\langle \ln P(T|X, \mathbf{w}, \tau) \rangle = \frac{N}{2} \langle \ln \tau \rangle - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \langle \tau \rangle \left\{ \sum_{n=1}^{N} t_n^2 - 2 \langle \mathbf{w} \rangle^T \sum_{n=1}^{N} \boldsymbol{\phi}_n t_n + \sum_{n=1}^{N} \boldsymbol{\phi}_n^T \langle \mathbf{w}\mathbf{w}^T \rangle \boldsymbol{\phi}_n \right\} \tag{40}$$

$$\begin{aligned} \langle \ln P(\mathbf{w}|\boldsymbol{\alpha}) \rangle &= -\frac{N+1}{2} \ln(2\pi) - \frac{1}{2} \sum_{m=0}^{N} \langle \ln \alpha_m \rangle \\ &\quad -\frac{1}{2} \sum_{m=0}^{N} \langle \alpha_m \rangle \langle w_m^2 \rangle \end{aligned} \tag{41}$$

$$\begin{aligned} \langle \ln P(\boldsymbol{\alpha}) \rangle &= (N+1) a \ln b + (a-1) \sum_{m=0}^{N} \langle \ln \alpha_m \rangle \\ &\quad -b \sum_{m=0}^{N} \langle \alpha_m \rangle - (N+1) \ln \Gamma(a) \end{aligned} \tag{42}$$

$$\begin{aligned} \langle \ln P(\tau) \rangle &= c \ln d + (c-1) \langle \ln \tau \rangle \\ &\quad -d \langle \tau \rangle - \ln \Gamma(c) \end{aligned} \tag{43}$$

$$\begin{aligned} -\langle \ln Q_{\mathbf{w}} \rangle &= (N+1)(1 + \ln(2\pi))/2 \\ &\quad + \ln |\boldsymbol{\Sigma}_{\mathbf{w}}| /2 \end{aligned} \tag{44}$$

$$\begin{aligned} -\langle \ln Q_{\boldsymbol{\alpha}} \rangle &= \sum_{m=0}^{N} \Big\{ \widetilde{a}_m \ln \widetilde{b}_m + (\widetilde{a}_m - 1) \langle \ln \alpha_m \rangle \\ &\quad -\widetilde{b}_m \langle \alpha_m \rangle - \ln \Gamma(\widetilde{a}_m) \Big\} \end{aligned} \tag{45}$$

$$\begin{aligned} -\langle \ln Q_{\tau} \rangle &= \widetilde{c} \ln \widetilde{d} + (\widetilde{c} - 1) \langle \ln \tau \rangle \\ &\quad -\widetilde{d} \langle \tau \rangle - \ln \Gamma(\widetilde{c}). \end{aligned} \tag{46}$$

Experimental results in which this framework is applied to synthetic and real data sets are given in Section 6.

## 5 RVM Classification

The classification case is somewhat more complex than the regression case since we no longer have a fully conjugate heirarchical structure. To see how

to resolve this, consider again the log marginal probability of the target data, given the input data, which can be written

$$\ln P(T|X) = \ln \int \int P(T|X, \mathbf{w}) P(\mathbf{w}|\boldsymbol{\alpha}) P(\boldsymbol{\alpha}) \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\alpha}. \tag{47}$$

As before we introduce a factorized variational posterior distribution of the form $Q_{\mathbf{w}}(\mathbf{w}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$, and obtain the following lower bound on the log marginal probability

$$\begin{aligned} \ln P(T|X) \geq & \int \int Q_{\mathbf{w}}(\mathbf{w}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \\ & \ln \left\{ \frac{P(T|X, \mathbf{w}) P(\mathbf{w}|\boldsymbol{\alpha}) P(\boldsymbol{\alpha})}{Q_{\mathbf{w}}(\mathbf{w}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})} \right\} \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\alpha}. \end{aligned} \tag{48}$$

Now, however, the right hand side of (48) is intractable. We therefore follow Jaakkola and Jordan (1998) and introduce a further bound using the inequality

$$\sigma(y)^t [1 - \sigma(y)]^{1-t} = \sigma(z) \tag{49}$$

$$\geq \quad \sigma(\xi) \exp \left( \frac{z - \xi}{2} - \lambda(\xi)(z^2 - \xi^2) \right) \tag{50}$$

where $z = (2t - 1)y$ and $\lambda(\xi) = (1/4\xi) \tanh(\xi/2)$. Here $\xi$ is a variational parameter, such that equality is achieved for $\xi = z$. Thus we have

$$\begin{aligned} P(T|X, \mathbf{w}) \geq F(T, X, \mathbf{w}, \boldsymbol{\xi}) = & \prod_{n=1}^{N} \sigma(\xi_n) \\ & \exp \left( \frac{z_n - \xi_n}{2} - \lambda(\xi_n)(z_n^2 - \xi_n^2) \right) \end{aligned} \tag{51}$$

where $z_n = (2t_n - 1)\mathbf{w}^T \boldsymbol{\phi}_n$. Substituting the result (51) into (48), and noting that $P(T|X, \mathbf{w}) \geq F(T, X, \mathbf{w}, \boldsymbol{\xi})$ implies $\ln P(T|X, \mathbf{w})/F(T, X, \mathbf{w}, \boldsymbol{\xi}) \geq 0$, we obtain a lower bound on the original lower bound, and hence we have

$$\begin{aligned} \ln P(T|X) \geq \mathcal{L} = & \int \int \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\alpha} Q_{\mathbf{w}}(\mathbf{w}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \\ & \ln \left\{ \frac{F(T, X, \mathbf{w}) P(\mathbf{w}|\boldsymbol{\alpha}) P(\boldsymbol{\alpha})}{Q_{\mathbf{w}}(\mathbf{w}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})} \right\}. \end{aligned} \tag{52}$$

We now optimize the right hand side of (52) with respect to the functions $Q_{\mathbf{w}}(\mathbf{w})$ and $Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ as well as with respect to the parameters $\boldsymbol{\xi} = \{\xi_n\}$. The variational optimization for $Q_{\mathbf{w}}(\mathbf{w})$ yields a normal distribution of the

form

$$Q_{\mathbf{w}}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{S}) \tag{53}$$

$$\mathbf{S} = \left(\mathbf{A} + 2\sum_{n=1}^{N}\lambda(\xi_n)\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T\right)^{-1} \tag{54}$$

$$\mathbf{m} = \frac{1}{2}\mathbf{S}\left(\sum_{n=1}^{N}(2t_n-1)\boldsymbol{\phi}_n\right) \tag{55}$$

where $\mathbf{A} = \mathrm{diag}\langle\alpha_m\rangle$. Similarly, variational optimization of $Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$ yields a product of Gamma distributions of the form

$$Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) = \prod_{m=0}^{N}\Gamma(\alpha_m|\widetilde{a},\widetilde{b}_m) \tag{56}$$

$$\widetilde{a} = a + \frac{1}{2} \qquad \widetilde{b}_m = b + \frac{1}{2}\langle w_m^2\rangle. \tag{57}$$

Finally, maximizing (52) with respect to the variational parameters $\xi_n$ gives re-estimation equations of the form

$$\xi_n^2 = \boldsymbol{\phi}_n^T\langle\mathbf{w}\mathbf{w}^T\rangle\boldsymbol{\phi}_n. \tag{58}$$

We can also evaluate the lower bound given by the right hand side of (52)

$$\begin{aligned}\mathcal{L} = {} & \langle\ln F\rangle + \langle\ln P(\mathbf{w}|\boldsymbol{\alpha})\rangle + \langle\ln P(\boldsymbol{\alpha})\rangle \\ & -\langle\ln Q_{\mathbf{w}}(\mathbf{w})\rangle - \langle Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})\rangle\end{aligned} \tag{59}$$

where we have

$$\begin{aligned}\langle\ln F\rangle = \sum_{n=1}^{N}\Big\{ & \ln\sigma(\xi_n) + \frac{1}{2}(2t_n-1)\langle\mathbf{w}^T\rangle\boldsymbol{\phi}_n \\ & -\frac{1}{2}\,\xi_n - \lambda(\xi_n)\left(\boldsymbol{\phi}_n^T\langle\mathbf{w}\mathbf{w}^T\rangle\boldsymbol{\phi}_n - \xi_n^2\right)\Big\}\end{aligned} \tag{60}$$

$$\begin{aligned}\langle\ln P(\mathbf{w}|\boldsymbol{\alpha})\rangle = & -\frac{1}{2}\sum_{m=0}^{N}\langle\alpha_m\rangle\langle w_m^2\rangle \\ & +\frac{1}{2}\sum_{m=0}^{N}\langle\ln\alpha_m\rangle - \frac{(N+1)}{2}\ln(2\pi)\end{aligned} \tag{61}$$

$$\begin{aligned}\langle\ln P(\boldsymbol{\alpha})\rangle = {} & \sum_{m=0}^{N}\Big\{-b\widetilde{a}/\widetilde{b} + (a-1)\left(\psi(\widetilde{a}) - \ln\widetilde{b}\right) \\ & +a\ln b - \ln\Gamma(a)\Big\}\end{aligned} \tag{62}$$

$$-\langle \ln Q_{\mathbf{w}}(\mathbf{w}) \rangle = \frac{N+1}{2}\,(1+\ln 2\pi) + \frac{1}{2}\ln|\mathbf{S}| \tag{63}$$

$$\begin{aligned} -\langle \ln Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}) \rangle \quad = \quad & \sum_{m=0}^{N} \Big\{ -(\widetilde{a}_m - 1)\psi(\widetilde{a}_m) \\ & - \ln \widetilde{b}_m + \widetilde{a}_m + \ln \Gamma(\widetilde{a}_m) \Big\}. \end{aligned} \tag{64}$$

Predictions from the trained model for new inputs can be obtained by substituting the posterior mean weights into (8) to give the predictive distribution in the form

$$P(t|\mathbf{x}, \langle \mathbf{w} \rangle). \tag{65}$$

A more accurate estimate would take account of the weight uncertainty by marginalizing over the posterior distribution of the weights. Using the variational result $Q_{\mathbf{w}}(\mathbf{w})$ for the posterior distribution leads to convolution of a sigmoid with a Gaussian, which is intractable. From symmetry, however, such a marginalization does not change the location of the $p = 0.5$ decision surface. A useful approximation to the required integration has been given by MacKay (1992).

## 6 Experimental Results

### 6.1 Regression

We illustrate the operation of the variational relevance vector machine (VRVM) for regression using first of all a synthetic data set based on the function $sinc(x) = (\sin x)/x$ for $x \in (-10, 10)$, with added noise. Figure 3 shows the result from a Gaussian kernel relevance vector regression model, and Figure 4 illustrates the mean hyperparameter values and weights associated with the model of Figure 3. Results from averaging over 25 such randomly generated data sets are shown in Table 1.
As an example of a regression problem using real data, we show results in Table 2 for the popular Boston housing dataset.

### 6.2 Classification

We illustrate the operation of the VRVM for classification with some synthetic data in two dimensions taken from Ripley (1994). A randomly chosen subset of 100 training examples (of the original 250) was utilised to train an SVM, RVM and VRVM. Results from typical SVM and VRVM classifiers, using Gaussian kernels of width 0.5, are shown in Figures 5 and 6 respectively.

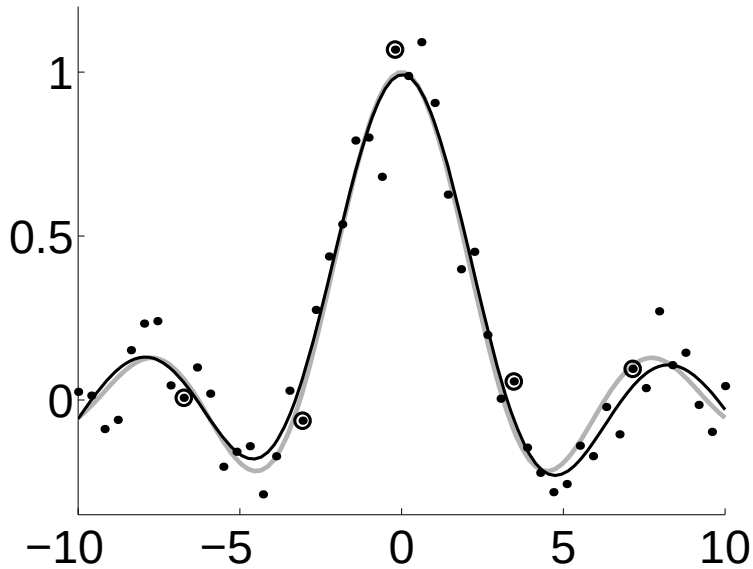



FIGURE 3. Example fit of a variational RVM to 50 data points generated from the 'sinc' function with added Gaussian noise of standard deviation 0.1. The sinc function and the mean interpolant are plotted in grey and black respectively, and the five relevance vectors (obtained by thresholding the mean weights at $10^{-3}$) are circled. The RMS deviation from the true function is 0.032, while a comparable SVM gave error of 0.038 using 36 support vectors. The VRVM also gives an estimate of the noise, which in this case had mean value 0.0945.

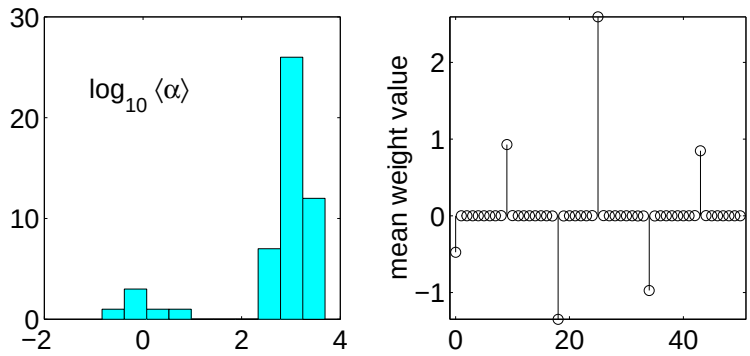



FIGURE 4. (Left) Histogram of the mean of the approximate $\alpha$ posterior. (Right) A plot of the 51 (unthresholded) mean weight values (the first weight is the bias, the next 50 correspond to the 50 data points, read left-to-right, in Figure 3). The dichotomy into 'relevant' and 'irrelevant' weights is clear.

| **Model** | **Error** | **# kernels** | **Noise estimate** |
|---|---|---|---|
| SVM | 0.0519 | 28.0 | – |
| RVM | 0.0494 | 6.9 | 0.0943 |
| VRVM | 0.0494 | 7.4 | 0.0950 |

TABLE 1. RMS test error, number of utilised kernels and, for the relevance models, noise estimates averaged over 25 generations of the noisy sinc dataset. For all models, Gaussian kernels were used with the width parameter selected from a range of values using 5-fold cross-validation. For the SVM, the parameters $C$ (the trade-off parameter) and $\epsilon$ (controlling the insensitive region of the loss function) were chosen via a further 5-fold cross-validation.

| **Model** | **Error** | **# kernels** | **Noise estimate** |
|---|---|---|---|
| SVM | 10.29 | 235.2 | – |
| RVM | 10.17 | 41.1 | 2.49 |
| VRVM | 10.36 | 40.9 | 2.49 |

TABLE 2. Squared test error, number of utilised kernels and noise estimates averaged over 10 random partitions of the Boston housing dataset into training/test sets of size 481 and 25 respectively. A third order polynomial kernel was used.

To assess the accuracy of the classifiers on this dataset, models with Gaussian kernels were used, with the width parameter of the Gaussian chosen by 5-fold cross-validation, and the SVM trade-off parameter $C$ was similarly estimated using a further 5-fold cross-validation. The results are given in Table 3.

| **Model** | **Error** | **# kernels** |
|---|---|---|
| SVM | 10.6% | 38 |
| RVM | 9.3% | 4 |
| VRVM | 9.2% | 4 |

TABLE 3. Percentage misclassification rate and number of kernels used for classifiers on the Ripley synthetic data. The Bayes error rate for this data set is 8%.

The 'Pima Indians' diabetes dataset is a popular classification benchmark. Table 4 summarises results on Ripley's split of this dataset into 200 training and 332 test examples.

## 7 Discussion

In this paper we have developed a practical variational framework for the Bayesian treatment of Relevance Vector Machines.

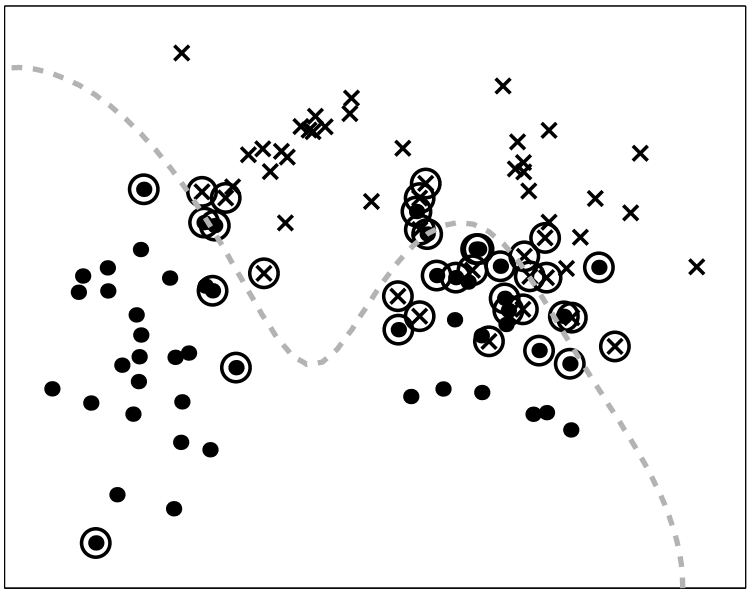


FIGURE 5. Support vector classifier of the Ripley dataset for which there are 38 kernel functions.

| **Model** | **Error** | **# kernels** |
|---|---|---|
| SVM | 69 | 110 |
| RVM | 65 | 4 |
| VRVM | 65 | 4 |

TABLE 4. Number of misclassifications and number of kernels used for classifiers on the Pima Indians data.

The variational solution for the Relevance Vector Machine is computationally more expensive than the type-II maximum likelihood approach. However, the advantages of a fully Bayesian approach are expected to be most pronounced in situations where the size of the data set is limited, in which case the computational cost of the training phase is likely to be insignificant.

## 8 References

Diggle, P.J., Liang, K-Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Berger J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition.

Bishop C.M. (1999). Bayesian PCA. In S. A. Solla S. A., Kearns M.S. and Cohn D.A., editors, *Advances in Neural Information Processing Systems*, volume **11**, 382-388. MIT Press.

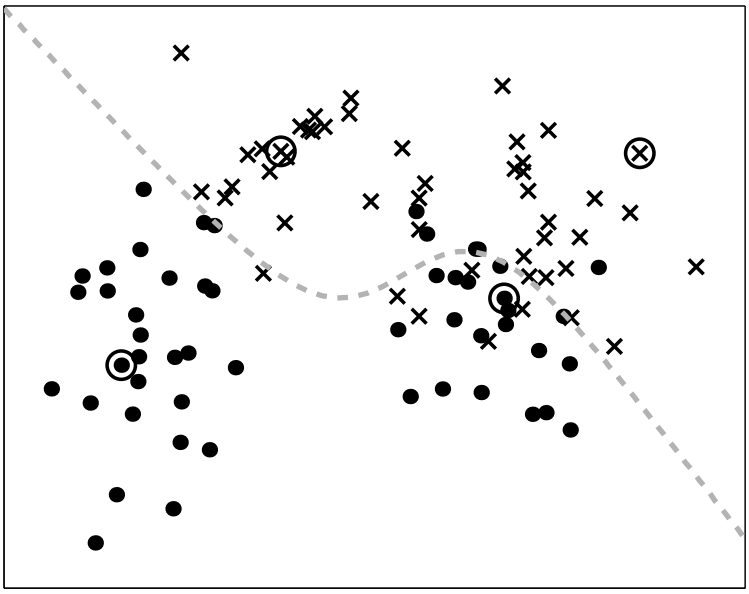


FIGURE 6. Variational relevance vector classifier of the Ripley dataset for which there are 4 kernel functions.


Jaakkola T. and M.I. Jordan (1998). Bayesian parameter estimation through variational methods, To appear in *Statistics and Computing*.

Jordan M.I., Z. Gharamani, T. S. Jaakkola, and L.K. Saul (1998). An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, 105-162. Kluwer.

MacKay D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation*, **4(5)**, 720-736, 1992.

Neal R.M. and G. E. Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer.

Ripley. B. D. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society, B*, **56(3)**, 409-456.

Tipping M. E. (2000). The relevance vector machine. In Sara A Solla, Todd K Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems*, **12**. Cambridge, Mass: MIT Press. To appear.

Vapnik V.N. (1998). *Statistical Learning Theory*. Wiley, New York, 1998.

Waterhouse S., D. MacKay, and T. Robinson (1996). Bayesian methods for mixtures of experts. In M. C. Mozer D. S. Touretzky and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, 351-357. MIT Press.

# Modelling Production with Undesirable Outputs

Carmen Fernández[1], Gary Koop[2], and Mark F.J. Steel[2]

[1] School of Mathematics and Statistics, University of St. Andrews, St. Andrews, KY16 9SS, U.K.
[2] Department of Economics, University of Edinburgh, Edinburgh EH8 9JY, U.K.

**Abstract:** Many production processes yield both good outputs and undesirable ones (*e.g.* pollutants). In this paper, we develop a generalization of a stochastic frontier model which is appropriate for such technologies. We discuss efficiency analysis and, in particular, define technical and environmental efficiency in the context of our model. Methods for carrying out Bayesian inference are described and applied to a longitudinal (or panel) data set of Dutch dairy farms.

**Keywords:** Bayesian methods; Dairy farms; Efficiency; Environment; Longitudinal data; Markov chain Monte Carlo; Stochastic frontiers.

## 1 Introduction

Stochastic frontier models are commonly used in the empirical study of production technology and the efficiency of economic agents, such as firms, individuals or countries. The seminal papers in the field are Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977), while a survey is provided in Bauer (1990). The ideas underlying this class of models can be applied to production models, but also to cost frontiers (by suitably redefining the quantities involved). The discussion here will focus on production frontiers, which aim to capture the maximum amount of output that can be obtained from a given level of inputs. Thus, they describe the best-practice technology for turning inputs into output. In practice, actual output of an individual production unit may fall below the maximum possible. The latter deviation from the frontier is a measure of inefficiency and is the focus of interest in many applications. The introduction of measurement or specification error is required by the fact that we do not know where the frontier is situated and have to estimate it from the available data. This makes the frontier stochastic, hence the term "stochastic frontier model".
The standard stochastic frontier model addresses the situation where only one output is produced, with a set of inputs. Inference with such standard stochastic frontier models can be done using classical or Bayesian approaches. In previous work, we have introduced and argued in favour of a Bayesian approach (see *e.g.* van den Broeck, Koop, Osiewalski and Steel,

1994). Some theoretical foundations for Bayesian analysis in stochastic frontier models are presented in Fernández, Osiewalski and Steel (1997) and an introductory survey of Bayesian methods in such models can be found in Koop and Steel (2000). Classical methods are discussed in *e.g.* Bauer (1990) or Horrace and Schmidt (1996). The present paper will take a Bayesian view.

An important extension of this framework is to allow for more than one type of output to be produced simultaneously. A Bayesian model for these multiple-output production processes is proposed in Fernández, Koop and Steel (2000a). The present paper further extends the above model in order to deal with situations where an individual unit produces undesirable outputs (such as pollution) as an inevitable by-product of the production of desirable outputs. In the application used in this paper, for example, Dutch dairy farms produce not only good outputs, such as milk, but also undesirable outputs, such as excessive nitrogen due to the application of manure and chemical fertilizers. It is thus important to understand the nature of the best-practice technology available to farmers for turning inputs into good and bad outputs. Furthermore, it is important to see how individual farmers measure up to this technology. In other words, evaluation of farm efficiency, both in producing as many good outputs and as few undesirable outputs as possible, is of interest. Here, we describe how extensions of stochastic frontier models can be used to shed light on these issues. We begin by explaining the generalization of the standard single-output stochastic frontier model to allow for several good outputs, following Fernández *et al.* (2000a). Next, we consider the more challenging case where some of these outputs can be undesirable. In the fourth Section, we shall briefly outline the prior used in the Bayesian model and the inference procedure used. Finally, we present some of the results for our empirical application involving Dutch dairy farms.

## 2 A Stochastic Frontier Model with Multiple Good Outputs

In Fernández *et al.* (2000a), we developed extensions of stochastic frontier models to allow for efficiency analysis in the presence of multiple outputs. Note that previous work with multiple outputs has often involved either having data on prices (*e.g.* in order to estimate a demand system) or on costs (*e.g.* in order to estimate a cost function). However, particularly in the case when some of the outputs are not sold in markets (*e.g.* pollution), such price or cost information is not available. Hence, it is important to develop methods which involve only output and input data.

The theoretical starting point in most analyses of multiple-output technology is a transformation function:

$$f(y, x) = 0,$$

where $y$ is a vector of $p$ good outputs and $x$ is a vector of inputs. If the transformation function is separable then we can write it as:

$$\theta(y) = h_g(x).$$

In the present paper, we assume a constant elasticity of transformation form for $\theta(y)$, but the basic ideas extend to any form.
To establish some terminology, note that $\theta(y) = constant$ maps out the output combinations that are equivalent. Hence, it is referred to as the production equivalence surface, which is $(p-1)$-dimensional. By analogy with the single output case, $h_g(x)$ defines the maximum output (as measured by $\theta(y)$) that can be produced with inputs $x$ and is referred to as the production frontier.
Since the empirical application used in the present paper involves (unbalanced) longitudinal or panel data, we assume that we have a set of $NT$ observations corresponding to outputs of $N$ different firms, where firm $i$ is observed for time periods $t = 1, \ldots, T_i$. The output of firm $i$ $(i = 1, \ldots, N)$ at time $t$ $(t = 1, \ldots, T_i)$ is $p$-dimensional and is given by the vector $y_{(i,t)} = (y_{(i,t,1)}, \ldots, y_{(i,t,p)})' \in \Re^p_+$. We use the following transformation of the $p$-dimensional output vector:

$$\theta_{(i,t)} = \left( \sum_{j=1}^{p} \alpha_j^q \, y_{(i,t,j)}^q \right)^{1/q}, \tag{1}$$

with $\alpha_j \in (0,1)$ for all $j = 1, \ldots, p$ and such that $\sum_{j=1}^{p} \alpha_j = 1$ and with $q > 1$. For fixed values of $\alpha = (\alpha_1, \ldots, \alpha_p)'$, $q$ and $\theta_{(i,t)}$, (1) defines a $(p-1)$-dimensional surface in $\Re^p_+$ corresponding to all the $p$-dimensional vectors of outputs $y_{(i,t)}$ that are technologically equivalent. In other words, (1) plots the production equivalence surface.
Given the transformation from the multivariate output vector $y_{(i,t)}$ to the univariate quantity $\theta_{(i,t)}$ (the parameters of which we estimate from the data), the basic problem of finding firm-specific efficiencies is essentially the same as in the single-output case. If we interpret the value $\theta_{(i,t)}$ as a kind of "aggregate output", and group these transformed outputs in an $NT$-dimensional vector

$$\log\theta = (\log\theta_{(1,1)}, .., \log\theta_{(1,T_1)}, \ldots, \log\theta_{(N,T_N)})', \tag{2}$$

we model $\log\theta$ through the following stochastic frontier model:

$$\log\theta = V\beta - Dz + \varepsilon_g. \tag{3}$$

In the latter equation, $V = (v(x_{(1,1)}), \ldots, v(x_{(N,T_N)}))'$ denotes an $NT \times k$ matrix of exogenous regressors, where $v(x_{(i,t)})$ is a $k$-dimensional function of the inputs $x_{(i,t)}$ corresponding to firm $i$ at time $t$. The particular choice

of $v(\cdot)$ defines the specification of the production frontier: *e.g.* $v(x_{(i,t)})$ is the vector of an intercept and all logged inputs for a Cobb-Douglas technology, whereas a translog frontier also involves squares and cross products of these logs. The corresponding vector of regression coefficients is denoted by $\beta \in \mathcal{B} \subseteq \Re^k$. Often, theoretical considerations will lead to regularity conditions on $\beta$, which will restrict the parameter space $\mathcal{B}$ to a subset of $\Re^k$, still $k$-dimensional and possibly depending on $x$. For instance, we typically want to ensure that the marginal products of inputs are positive.

Technical inefficiency is captured by the fact that firms may lie below the frontier, thus leading to a vector of inefficiencies $\gamma \equiv Dz \in \Re_+^{NT}$, where $D$ is an exogenous $NT \times M$ ($M \leq NT$) matrix and $z \in \mathcal{Z}$ with $\mathcal{Z} = \{z = (z_1, \ldots, z_M)' \in \Re^M : Dz \in \Re_+^{NT}\}$. Through different choices of $D$, we can accommodate various amounts of structure on the vector $\gamma$ of inefficiencies. For instance, taking $D = I_{NT}$, the $NT$-dimensional identity matrix, leads to an inefficiency term which is specific to each different firm and time period. For a balanced panel (*i.e.* $T_i = T, i = 1, \ldots, N$), $D = I_N \otimes \iota_T$, where $\iota_T$ is a $T$-dimensional vector of ones and $\otimes$ denotes the Kronecker product, implies inefficiency terms which are specific to each firm, but constant over time (*i.e.* "individual effects"). In our application we make the latter choice for $D$ (but with the obvious generalization to an unbalanced panel). Since we are working in terms of $\log\theta$, the log of the aggregate output, the *technical efficiency* corresponding to firm $i$ (at any period) will be defined as $\tau_{1i} = \exp(-z_i)$ where $z_i$ is the appropriate element of $z$. For more discussion and alternative definitions of efficiency measures, see Fernández, Koop and Steel (2000b). The term $\varepsilon_g$ in (3) is meant to capture all other influences, such as measurement or specification error, and is accordingly not restricted in its sign.

Stochastics will be introduced into the sampling model through distributions on $z$ (which could, equivalently, be considered part of the prior) and $\varepsilon_g$. Here, we use a choice of $D$ that makes $\mathcal{Z} = \Re_+^N$ and we assume independence across observations and between $z$ and $\varepsilon$. In order to fully specify a likelihood function for the $p$-dimensional outputs when $p > 1$, we also introduce a distribution on the weighted output shares, defined as

$$\eta_{(i,t,j)} = \frac{\alpha_j^q y_{(i,t,j)}^q}{\sum_{l=1}^p \alpha_l^q y_{(i,t,l)}^q}, \; j = 1, \ldots, p, \tag{4}$$

In particular, we group them into $\eta_{(i,t)} = (\eta_{(i,t,1)}, \ldots, \eta_{(i,t,p)})'$, and assume independent sampling from

$$p(\eta_{(i,t)}|s) = f_D^{p-1}(\eta_{(i,t)}|s), \tag{5}$$

where $s = (s_1, \ldots, s_p)' \in \Re_+^p$ and $f_D^{p-1}(\cdot|s)$ is the p.d.f. of a Dirichlet distribution with parameter $s$.

## 3 A Stochastic Frontier Model with Good and Bad Outputs

Important issues in environmental policy hinge on multiple output production technologies where some of the outputs are undesirable. For instance, we have data on farms which produce good outputs (*e.g.* dairy products) for the market and undesirable outputs (pollutants). We will refer to undesirable outputs as "bads". Efficiency analysis using stochastic frontier models can be used to shed light on practical policy questions, involving both the goods and the bads. For instance, if we find dairy farms to be environmentally efficient then pollution can only be reduced by reducing production at dairy farms. However, if many dairy farms are highly environmentally inefficient, then by adopting best-practice technology pollution can be reduced without harming production of milk.

The question now arises as to how to adapt the analysis of the previous section to allow for undesirable outputs and both technical and environmental inefficiency. Following Fernández *et al.* (2000a), we make one particular adaptation which we argue is reasonable. Others are clearly possible, and these are a topic of past and current research. For instance, Koop (1998) and Reinhard, Lovell and Thijssen (1999) assume that undesirable outputs can be treated as inputs. Fernández *et al.* (2000b) adapt the aggregator function in (1) to accommodate bad outputs. Here, we model the good outputs as in the previous section, but add a second frontier for the bad outputs. Environmental efficiency is then measured relative to this second frontier.

If we let $b$ indicate a vector of $m$ bad outputs, the most general description of best-practice technology is given by:

$$f(y, x, b) = 0.$$

We assume this transformation function can be broken down into:

$$\theta(y) = h_g(x),$$

and

$$\kappa(b) = h_b(y).$$

In other words, the general transformation function can be broken down into two equations involving a "goods production equivalence surface" $\theta(y)$, a "goods production frontier" $h_g(x)$, a "bads production equivalence surface" $\kappa(b)$, and a "bads production frontier" $h_b(y)$. The assumption that the amount of good outputs produced depends on the inputs, while production of bad outputs depends on the amount of good outputs is likely to be reasonable in many cases. If not, modifications of the present model can be implemented.

We begin with the model for the good outputs described in the previous section given by equations (1)-(5). We further let $b_{(i,t)} = (b_{(i,t,1)}, \ldots, b_{(i,t,m)})'$ be the vector of $m$ bad outputs for firm $i$ in period $t$. We define the environmental production equivalence surface through a similar constant elasticity of transformation form:

$$\kappa_{(i,t)} = \left( \sum_{j=1}^{m} \gamma_j^r \, b_{(i,t,j)}^r \right)^{1/r}, \tag{6}$$

with $\gamma_j \in (0,1)$ for all $j = 1, \ldots, m$ and such that $\sum_{j=1}^{m} \gamma_j = 1$ and with $0 < r < 1$.
Environmental inefficiency is measured using a stochastic frontier model with (6) as dependent variable. That is, we define $\log \kappa$ similarly to $\log \theta$ and set

$$\log \kappa = U\delta + Mv + \varepsilon_b \tag{7}$$

where $U = (u(y_{(1,1)}), \ldots, u(y_{(N,T_N)}))'$ is a function of the good outputs. $U$ plays a similar role to $V$ in equation (3) and, hence, the particular choice of $u(\cdot)$ defines the specification of the bads production frontier. Environmental inefficiencies are given by $Mv \in \Re_+^{NT}$. $M$ plays an analogous role to $D$ in the previous section and here we set $M = D$ which implies that the technical and environmental efficiency of each firm is constant over time. Thus, *environmental efficiency* of firm $i$ will be defined as $\tau_{2i} = \exp(-v_i)$.
To complete the sampling model, we shall introduce the following distributional assumptions. We link both frontiers by joint distributions on the inefficiency error terms and on the measurement error terms, while still retaining independence across observations. In particular, we assume a bivariate Normal distribution for $(\varepsilon_g, \varepsilon_b)$. That is, if we let $f_N^R(\varepsilon|a, A)$ denote the $R$-variate Normal p.d.f. with mean $a$ and covariance matrix $A$, evaluated at $\varepsilon$, we take:

$$p(\varepsilon_g, \varepsilon_b|\Sigma) = f_N^{2NT} \left( \begin{array}{c} \varepsilon_g \\ \varepsilon_b \end{array} \middle| 0, \Sigma \otimes I_{NT} \right) \tag{8}$$

where $\Sigma$ is a $2 \times 2$ P.D.S. matrix.
For the inefficiency error terms, we adopt a similar strategy, except that these have to be nonnegative. We assume independence between firms and for each $i = 1, \ldots, N$, we take a truncated Normal inefficiency distribution:

$$p(z_i, v_i|\mu, \Omega) = f_N^2((z_i, v_i)'|\mu, \Omega) f^{-1}(\mu, \Omega) I_{\Re_+^2}(z_i, v_i), \tag{9}$$

where $f(\mu, \Omega)$ is the integrating constant of the truncated Normal and $I_{\Re_+^2}(\cdot)$ is the indicator function for $\Re_+^2$.

Finally, we define a weighted vector of shares for the bads:

$$\zeta_{(i,t,j)} = \frac{\gamma_j^r b_{(i,t,j)}^r}{\sum_{l=1}^m \gamma_l^r b_{(i,t,l)}^r}, \; j = 1, \ldots, m, \tag{10}$$

stack them to form $\zeta_{(i,t)} = (\zeta_{(i,t,1)}, \ldots, \zeta_{(i,t,m)})'$, and assume independent sampling from

$$p(\zeta_{(i,t)}|h) = f_D^{m-1}(\zeta_{(i,t)}|h), \tag{11}$$

where $h = (h_1, \ldots, h_m)' \in \Re_+^m$

## 4 The Prior and Bayesian Inference

In the previous Sections, we have defined the sampling model, which depends on the parameters $(\beta, \delta, \Sigma, \mu, \Omega, \alpha, \gamma, q, r, s, h)$. We shall use the proper prior structure

$$p(\beta, \delta, \Sigma, \mu, \Omega, \alpha, \gamma, q, r, s, h) = p(\beta, \delta, \Sigma)p(\mu, \Omega)p(\alpha)p(\gamma)p(q)p(r)p(s)p(h)$$

The prior we assume on these parameters is chosen to be rather noninformative, except that we restrict $\beta$ and $\delta$ to their respective regularity regions, and we impose that $q > 1$ and $0 < r < 1$, again for economic theory considerations.
In particular, we take an Inverted Wishart prior on $\Omega$:

$$p(\Omega) = f_{IW}^2\left(\Omega|\Omega_0, \nu_0\right), \tag{12}$$

combined with

$$p(\mu|\Omega) = f_N^2(\mu|0, c\Omega), \tag{13}$$

Fernández, Koop and Steel (1999) report some simulation exercises to calibrate the prior of $(\mu, \Omega)$ (*i.e.* choose values for the hyperparameters in (12) and (13)) so as to induce a reasonable prior on the efficiencies $\tau_{1i}$ and $\tau_{2i}$. For the other parameters we assume

$$p(\beta, \delta, \Sigma) \propto f_N\left(\begin{array}{c}\beta \\ \delta\end{array} |b_0, H_0^{-1}\right) f_{IW}^2(\Sigma|\Sigma_0, \lambda_0) I_{\mathcal{RR}}(\beta, \delta, \Sigma), \tag{14}$$

where $\mathcal{RR}$ indicates the regularity region,

$$p(\alpha) = f_D^{p-1}(\alpha|a_0), \tag{15}$$

$$p(\gamma) = f_D^{m-1}(\gamma|g_0), \tag{16}$$

$$p(q) \propto f_G(q|1, q_0) I_{(1,\infty)}(q), \tag{17}$$

where $f_G(.|a, b)$ denotes a Gamma density function with shape parameter $a$ and mean $a/b$ (if $a = 1$, we have an Exponential),

$$p(r) \propto f_G(r|1, r_0) I_{(0,1)}(r), \tag{18}$$

$$p(s) = \prod_{j=1}^{p} f_G(s_j|1, k_j), \tag{19}$$

and, finally,

$$p(h) = \prod_{j=1}^{m} f_G(h_j|1, n_j). \tag{20}$$

We adopt noninformative choices for the hyperparameters in (14)-(20). The resulting posterior from combining the sampling model with the prior just described does not lend itself to immediate analytical analysis. Instead, we shall use a Markov chain Monte Carlo (MCMC) algorithm on the space of the parameters augmented with the inefficiencies $(z, v)$. The Markov chain will be constructed from Gibbs steps for $(z, v), (\beta, \delta), \Sigma$, where we can draw immediately from the conditionals, and Normal random walk Metropolis samplers for $\Omega, \mu, \alpha, \gamma, q, r, s, h$, since the conditionals for the latter do not have a well-known form. We fine-tune results from preliminary runs in order to select the variance for the increments in the random walk Metropolis samplers. The relevant conditional posterior distributions are described in detail in Fernández *et al.* (1999).

## 5 An Application to a Panel of Dutch Dairy Farms

We apply the model described in the previous Sections to a data set involving $N = 613$ Dutch dairy farms for the years 1991-94. It is an unbalanced panel with a total number of observations $NT = 1545$. For each farm, we have data on $p = 2$ good outputs, $m = 1$ bad output and 3 inputs:

- Good outputs: Milk (millions of kg) and Non-milk (millions of 1991 Guilders).
- Bad output: Nitrogen surplus (thousands of kg).
- Inputs: Family labor (thousands of hours), Capital (millions of 1991 Guilders) and Variable input (thousands of 1991 Guilders).

Variable input includes *inter alia* hired labor, concentrates, roughage and fertilizer. Non-milk output contains meat, livestock and roughage sold. The definition of capital includes land, buildings, equipment and livestock. Further detail on this data set is given in Reinhard *et al.* (1999).
Both the goods and bad production frontiers are here assumed to take Cobb-Douglas forms. A more detailed discussion of the empirical results can be found in Fernández *et al.* (1999).
Table 1 provides some characteristics of the posterior distribution. Note that the column labelled "Median" is the posterior median. The columns labelled "2.5%" and "97.5" are the 2.5% and 97.5% percentiles, respectively of the posterior distribution. "RTS" means returns to scale, which indicates the relative increase in aggregate output expressed as a fraction of a relative increase in all inputs (or good outputs for the bads frontier). We also summarize results for the technical and environmental efficiencies of a typical or average farm, $\tau_{1f}$ and $\tau_{2f}$. The latter results correspond to a predictive out-of-sample efficiency distribution, obtained by integrating out the distribution in (9) with the posterior of $(\mu, \Omega)$. Our model allows for technical and environmental efficiencies to be correlated with one another and that correlation is evaluated at 0.25, indicating that there is a slight tendency for technically inefficient farms to also be environmentally inefficient.

**Table 1: Posterior Results for Dutch Dairy Farm Data Set**

| | Median | 2.5% | 97.5% |
|---|---|---|---|
| $\beta_1$(Intercept) | -3.533 | -3.694 | -3.226 |
| $\beta_2$(Labour) | 0.120 | 0.090 | 0.150 |
| $\beta_3$(Capital) | 0.537 | 0.504 | 0.572 |
| $\beta_4$(Variable) | 0.487 | 0.463 | 0.509 |
| RTS (Goods) | 1.145 | 1.115 | 1.173 |
| $\delta_1$(Intercept) | 2.578 | 2.262 | 2.890 |
| $\delta_2$(Milk) | 0.889 | 0.858 | 0.921 |
| $\delta_3$(Non-milk) | 0.081 | 0.065 | 0.098 |
| RTS(Bads) | 0.971 | 0.940 | 1.001 |
| $q$ | 1.004 | 1.000 | 1.019 |
| $\alpha_1$ | 0.534 | 0.510 | 0.565 |
| $\tau_{1f}$ | 0.620 | 0.415 | 0.880 |
| $\tau_{2f}$ | 0.345 | 0.198 | 0.599 |

All results seem reasonably in accordance with economic intuition. Some of the more interesting results are:

- Firms tend to be more efficient technically than environmentally. In fact, the posterior median of the environmental efficiency for a typical

farm is only 0.345, indicating that the typical farm produces roughly three times as much nitrogen surplus as would be consistent with best practice! Using the same data, Reinhard *et al.* (1999) and Fernández *et al.* (1999), who both use a single-frontier model, find similar results.

- The small positive correlation between both types of efficiencies indicates that farms which tend to be less efficient technically also tend to be less efficient environmentally. In contrast, the single frontier analysis of Fernández *et al.* (1999) finds a moderately negative correlation.

- However, there is a large spread of efficiencies across farms, which manifests itself in large differences between the 2.5 and 97.5th percentiles of both technical and environmental efficiencies.

- Rather than conducting inference on the efficiency for a typical (unobserved) farm, we can also conduct inference on farm-specific efficiencies. Given that we have observed these farms, their efficiencies are less dispersed, and can lead to a ranking of firms, in the sense that *e.g.* the efficiencies of quartile firms (in the efficiency ranking) are quite well separated, and the posterior probability of these firms being reversed in the ranking is very low.

- Increasing returns to scale seem to exist for the production of good outputs, while slightly decreasing returns exists for bad output production.

- The elasticity of nitrogen production with respect to milk production ($\delta_2$) is much larger than the elasticity with respect to non-milk production ($\delta_3$). This finding indicates that it is the milk production side of dairy farming that is most associated with the production of nitrogen.

We hesitate to draw policy conclusions based solely on this one set of empirical results for one model specification. However, to illustrate the types of issues that our model can be used to address, we offer the following comments. The relatively large degree of environmental inefficiency indicates that pollution can be reduced in many farms at little cost in terms of foregone output. That is, if inefficient farms were to adopt best-practice technology and move towards their environmental production frontiers, production of pollutants could be reduced at no cost to milk or non-milk production. The positive correlation between the two types of efficiencies indicates that improving environmental efficiency could be associated with improvements in technical efficiency. Hence, policies aimed at improving efficiency (*e.g.* by educating farmers in best-practice technology) could have large payoffs. Furthermore, the pattern of returns to scale results indicate

that larger farms have advantages. Hence, policies which promote rationalization of farms (*e.g.* encouraging larger farms to purchase smaller farms) could result both in more production of milk and non-milk outputs (due to increasing returns to scale in the good production frontier) and less pollution (due to decreasing returns in the environmental production frontier).

## 6 Conclusions

In this paper, we have shown how the standard stochastic frontier model with a single output can be extended to multiple outputs where some of the outputs are undesirable. The model we develop can be used to model production technologies which produce *e.g.* pollutants. The empirical application to Dutch dairy farms shows the practicality of this approach and highlights some important policy issues which our model can address.

## References

Aigner, D.; Lovell, C.A.K. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, **6**, 21-37.

Bauer, P. (1990). Recent developments in the econometric estimation of frontiers. *Journal of Econometrics*, **46**, 39-56.

Osiewalski J. and Steel, M.F.J. (1994). Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics*, **61**, 273-303.

Fernández, C.; Koop, G. and Steel, M.F.J. (1999). Multiple output production with undesirable outputs: An application to nitrogen surplus in agriculture. manuscript.

Fernández, C.; Koop, G. and Steel, M.F.J. (2000a). A Bayesian analysis of multiple output production frontiers. *Journal of Econometrics*, forthcoming.

Fernández, C.; Koop, G. and Steel, M.F.J. (2000b). Alternative efficiency measures for multiple output production. manuscript.

Fernández, C.; Osiewalski, J. and Steel, M.F.J. (1997). On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics*, **79**, 169-193.

Horrace, W. and Schmidt, P. (1996). Confidence statements for efficiency estimates from stochastic frontiers. *Journal of Productivity Analysis*, **7**, 257-282.

Koop, G. (1998). Carbon dioxide emissions and economic growth: A structural approach. *Journal of Applied Statistics*, **25**, 489-515.

Koop, G.; and Steel, M.F.J. (2000). Bayesian analysis of stochastic frontier models. Forthcoming in B. Baltagi, ed., *Companion in Theoretical Econometrics*, Basil Blackwell, Oxford.

Meeusen, W. and van den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed errors. *International Economic Review*, **8**, 435-444.

Reinhard, S.; Lovell, C.A.K. and Thijssen, G. (1999). Econometric estimation of technical and environmental efficiency: An application to Dutch dairy farms, *American Journal of Agricultural Economics*, **81**, 129-153.

**Acknowledgements:** We acknowledge funding by the ESRC under grant number R34515 and we thank Stijn Reinhard and the Dutch Agricultural Economics Research Institute (LEI-DLO) for providing the data.

# Goodness of fit tests for regression models: A tutorial.

W. González Manteiga[1]

[1] Departamento de Estadística e I.O., Facultad de Matemáticas
Universidad de Santiago de Compostela, Santiago de Compostela
15706 Spain; wences@zmat.usc.es

**Abstract:** A brief account of the very recent work on goodness of fit for regression models is given. Two leading approaches based on smoothing and on empirical regression processes are provided. Some extensions to regression with binary responses and censored responses are also included.

**Keywords:** Goodness of fit; nonparametric regression; empirical regression process; bootstrap.

## 1 Introduction

Consider the regression model

$$Y_i = m\left(X_i\right) + \varepsilon_i = m\left(X_i\right) + \sigma\left(X_i\right)\eta_i \qquad i = 1, \ldots, n \tag{1}$$

where the random variable $Y$ is related to a p-dimensional vector $X$, and $\eta$ is a random error of zero mean and variance one, independent of $X$. In the called random design regression model, $X$ is random, and $\{(X_i, Y_i)\}_{i=1}^n$ is the initial random sample corresponding to the (p+1)-dimensional vector $(X, Y)$ with regression function $m(x) = \mathbb{E}(Y/X = x)$ and conditional variance $\sigma^2(x) = \mathrm{Var}(Y/X = x)$. In the case of fixed design, $X$ is constrained to fixed points $\{x_i\}_{i=1}^n$, following a design density function $f$. For example, in the one-dimensional case (p=1), the "regular fixed design", with finite support $[a, b]$ such that $\int_a^{x_i} f(x)\, dx = \frac{i-1/2}{n} \qquad i = 1, \ldots, n$.
One of the main tasks in regression analysis is to perform hypothesis tests of the form

$$H_0 : m \in \mathcal{M}_0 \qquad \text{against} \qquad H_1 : m \in \mathcal{M} \backslash \mathcal{M}_0, \quad \mathcal{M}_0 \subset \mathcal{M}$$

where $\mathcal{M}_0$ is a specified class of functions that, depending on the kind of problem in question, can correspond to linear models:

$$\mathcal{M}_0 = \left\{ m/m(x) = m_\theta(x) = A^t(x)\theta, \quad \theta \in \Theta \subset \mathbb{R}^q \right\},$$

where $A$ is a known function from $\mathbb{R}^p$ to $\mathbb{R}^q$ with $A^t(x)$ denoting the transpose of $A(x)$ (Seber (1977)), to generalized linear models:

$$\mathcal{M}_0 = \left\{ m/m(x) = g\left(A^t(x)\theta\right) \right\},$$

for a known function $g$ from $\mathbb{R}$ to $\mathbb{R}$ (Mc Cullagh and Nelder (1989)), to semiparametric models:

$$\mathcal{M}_0 = \left\{ m/m(x) = m\left(x_1^t, x_2^t\right) = A^t(x_1)\theta + m_2(x_2) \right\},$$

where $x_1 \in \mathbb{R}^{p_1}$, $x_2 \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$, $A$ is a function from $\mathbb{R}^{p_1}$ to $\mathbb{R}^q$ and $m_2$ is an unknown fuction from $\mathbb{R}^{p_2}$ to $\mathbb{R}$ (the called partially linear models, see for example Speckmann (1988)), etc. In this tutorial work our purpose will be only the goodness of fit test of a parametric regression model,

$$\mathcal{M}_0 = \{ m/m(x) = m_\theta(x), \theta \in \Theta \subset \mathbb{R}^q \},$$

under randon design.

A common feature of most work is to compare a parametric fit $m_{\theta_n}$ with a completely nonparametric estimator $m_n$ of $m$ and then to reject $H_0$ if $d_0\left(m_{\theta_n}, m_n\right)$ exceeds a critical value. Here $d_0$ is a proper distance on the space of all regression functions. The nonparametric estimation of $m$ requires selection of a smoothing parameter, and the power of the resulting tests is a function of this parameter. In the next section some examples are given.

Other possibility to test $H_0$ may also be based on an estimator of the integrated regression function

$$I(x) = \mathbb{E}\left[Y 1_{\{X \le x\}}\right] = \int_{-\infty}^{x} m \, dF_1, \qquad x \in \mathbb{R}^p,$$

where $F_1$ is the (unknown) distribution function of $X$ and the integral extends over the quadrant with right upper corner $x$. $I$ uniquely determines $m$ and a consistent estimator of $I$ is given by

$$I_n(x) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \le x\}} Y_i = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} y \, dF_n(u, v)$$

where $F_n(x, y) = \sum_{i=1}^{n} 1_{\{X_i \le x,\ Y_i \le y\}}$ is the empirical distribution function of $(X, Y)$.

The tests are now based on the distance $d_1\left(I_{\theta_n}, I_n\right)$, where

$$I_{\theta_n}(x) = \int_{-\infty}^{x} m_{\theta_n} \, dF_{1n} = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \le x\}} m_{\theta_n}\left(X_i\right)$$

is the parametric fit of $I$ under $H_0$, $F_{1n}$ is the empirical distribution function of the predictor variable $X$ and $d_1$ is a proper distance on the space

of all integrated regression functions. The nonparametric estimation of $I$ does not require bandwidth selection, but the asymptotic behaviour of $d_1$ is more complicated to obtain than the limit distribution of $d_0$ (in general asymptotically normal). Frequently $d_!$ can be seen as a continuous functional of an empirical regression process with parameter $\theta$ estimated (see Stute (1997) for more details). Section 3 is devoted to this approximation. In Section 4 a simulation example is given to show the behaviour of the last two goodness of fit test models. Section 5 is devoted to some extensions to regression with binary responses and censored responses. Finally, the tutorial work is ended with some comments of recent work on related topics, which could mean research for the the next years.

## 2 Some tests based on the estimation of the regression function

We will consider $\mathcal{M}_0$ the class of polynomials of degree less than or equal to (q-1) with $q$ fixed,

$$\mathcal{M}_0 = \left\{ m/m_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{q-1} x^{q-1},\ \theta \in \Theta \subset \mathbb{R}^q \right\}.$$

As $d_0$ we will use a weighted $L_2$-norm: $d_0(m_1, m_2) = \int (m_1 - m_2)^2 \omega$. As nonparametric estimator of the regression function a local polynomial regression is taken. The idea of local polynomial regression is to locally fit a low-order polynomial at a fixed point $x$, with observations receiving different weights. This leads to the following least-squares problem:

$$\min_\beta \sum_{i=1}^{n} \left( Y_i - \sum_{r=0}^{\bar{q}} \beta_r \left(X_i - x\right)^r \right)^2 K\left(\frac{X_i - x}{h}\right)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{\bar{q}})^t$, and the dependence of $\beta_r$ on $x$ is supressed. The bandwidth $h$ controls the size of the local neighbourhood and $K$ is a kernel function that assigns weights to each datum point.
For convenience, a matrix notation is introduced. Let $X$ be an $n \times (\bar{q}+1)$ design matrix, $X = \left((X_i - x)^j\right)_{1 \le i \le n,\ 0 \le j \le \bar{q}}$, and $W$ the $n \times n$ diagonal matrix with diagonal elements equal to $K\left(\frac{X_i - x}{h}\right)$, $1 \le i \le n$. The local $\bar{q}$-polynomial regression estimator is given by: $m_n(x) = \hat{\beta}_0(x) = \sum_{i=1}^{n} W_{n,\bar{q}}\left(\frac{X_i - x}{h}\right) Y_i$, where

$$W_{n,\bar{q}}(t) = u^t \left(X^t W X\right)^{-1} \left(1, ht, \ldots, h^{\bar{q}} t^{\bar{q}}\right)^t K(t)/h$$

with $u = (1, 0, \ldots, 0)^t$ a $(\bar{q}+1)$-dimensional vector, satisfying

$$\sum_{i=1}^{n} \left(X_i - x\right)^k W_{n,\bar{q}}\left(\frac{X_i - x}{h}\right) = \begin{cases} 1 & \text{if } k = 0 \\ 0 & \text{if } 1 \le k \le \bar{q} \end{cases}$$

>From these conditions it follows that the local polynomial fits polynomials of a degree not larger than $\bar{q}$, and then if $\bar{q} \geq q-1$,

$$\sum_{i=1}^{n} W_{n,\bar{q}}\left(\frac{X_i - x}{h}\right) m_\theta\left(X_i\right) = m_\theta(x) \qquad \forall \theta \in \Theta$$

In this way for the polynomial regression class, the use of local polynomials as the nonparametric estimators is especially interesting. When we compare a parametric polynomial fit against a local polynomial with the same or greater degree, we have two consistent unbiased estimators of the regression function under the null hypothesis.
Moreover, if the true $m(\cdot)$ has an orthogonal component that does not belong to $\mathcal{M}_0$ and we smooth by a local polynomial of a degree greater than $q-1$, for example $\bar{q}$, then the distance does not penalize the orthogonal terms up to degree $q-1$.
Let $m_{n,\bar{q}}(x)$ be the local poynomial smoother of order $\bar{q}$ and let $\theta_n$ be a $\sqrt{n}$-consistent estimator of $\theta_0$ under the null hypothesis. In Alcalá et al. (1999), under regularity conditions, the normal asymptotic distribution of $n\sqrt{h}d_0\left(m_{\theta_n}, m_{n,\bar{q}}\right)$ is obtained under the null hypothesis and also under local alternatives converging to the null hypothesis at a rate $\left(n\sqrt{h}\right)^{-1/2} = c_n$, this is to say, $m = m_{\theta_0} + c_n g$, with $g$ being a function belonging to the class of orthogonal functions of $\mathcal{M}_0$ with respect to the inner product

$$\langle m_1, m_2 \rangle = \int m_1 m_2 f \omega, \quad f \text{ being the density of } X.$$

However, the asymptotic distribution is inappropriate to calculate the critical values of the test for small or moderate sample sizes. A solution to this problem is the use of bootstrap procedures to obtain the critical value for finite samples: Bootstrapping the residuals if the error term is independent of the regressors ($\sigma^2(x) = \sigma^2$ in (1)) or using the wild bootstrap in more general situations. The two bootstrap versions are considered below. The first one consists of the following steps:

$a_1$) Consider the parametric residuals $e_i = Y_i - m_{\theta_n}\left(X_i\right), \qquad i = 1, \ldots, n.$

$b_1$) Recenter the previous residuals $\tilde{e}_i = e_i - \bar{e}, \quad i = 1, \ldots, n, \ \bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$.

$c_1$) Draw the bootstrap residuals $\{\varepsilon_i^*\}_{i=1}^n$ from the empirical distribution function of $\{\tilde{e}_i\}_{i=1}^n$.

$d_1$) Define the bootstrap sample by

$$Y_i^* = m_{\theta_n}\left(X_i\right) + \varepsilon_i^* \qquad i = 1, \ldots, n$$

and obtain $d_0\left(m_{\theta_n^*}, m_{n,\bar{q}}^*\right)$, where $m_{n,\bar{q}}^*$ and $\theta_n^*$ are constructed using the data $\{(X_i, Y_i^*)\}_{i=1}^n$.

$e_1$) Repeat the process B times and reject the hypothesis $H_0$ if $d_0\left(m_{\theta_n}, m_{n,\bar{q}}\right) > c^*_{1-\alpha}$, $c^*_{1-\alpha}$ being the $[B(1-\alpha)]$ th order statistic computed from the B replicates $\left\{d_0\left(m_{\theta^*_{ni}}, m^*_{n,\bar{q}i}\right)\right\} \qquad i = 1, \ldots, B$.

For the second one we have the next steps:

$a_2$) As in the previous $a_1$).

$b_2$) Draw bootstrap errors $\varepsilon^*_i$ such that $\mathbb{E}_*\left(\varepsilon^*_i\right) = 0$, $\mathbb{E}_*\left(\varepsilon^{*2}_i\right) = e^2_i$, $\mathbb{E}_*\left(\varepsilon^{*3}_i\right) = e^3_i \qquad i = 1, \ldots, n$, where $\mathbb{E}_*$ is the expectation operator in the probability space associated with the resampling.

For example $\varepsilon^*_i = e_i V^*_i$, $V^*_i$ being independent of the original sample with $\mathbb{E}^*\left(V^*_i\right) = 0$, $\mathbb{E}^*\left(V^{*2}_i\right) = 1$ and $\mathbb{E}^*\left(V^{*3}_i\right) = 1$

$c_2$) Define the bootstrap sample by

$$Y^*_i = m_{\theta_n}\left(X_i\right) + \varepsilon^*_i \qquad i = 1, \ldots, n$$

and proceed in a similar way as in $d_1$) and $e_1$).

## 3 Some tests based on the estimation of the integrated regression function

Considering the most simple case for the simple hypothesis $H_0 : m = m_{\theta_0}$ with $\theta_0 \in \Theta$ known, and taking into account the comments given in the introduction, we obtain:

$$\begin{aligned} I_n(x) - I_0(x) &= n^{-1}\sum_{i=1}^{n} 1_{\{X_i \le x\}} Y_i - \int_{-\infty}^{x} m_{\theta_0}(x)\, dF_{1n}(x) \\ &= n^{-1}\sum_{i=1}^{n} 1_{\{X_i \le x\}}\left(Y_i - m_{\theta_0}\left(x_i\right)\right), \end{aligned}$$

the difference between the estimations of the integrated regression function, under the general hypothesis and under the null hypothesis. Now, define $R_n(x) = n^{1/2}\left(I_n(x) - I_0(x)\right)$. $R_n$ constitutes the empirical process of the regressors marked by the errors: $\varepsilon_i = Y_i - m_{\theta_0}\left(X_i\right) \quad 1 \le i \le n$. For example, in the one-dimensional case (p=1), (see Stute (1997) for more details), $R_n \to R_\infty$ in distribution in the Skorokhod space $D[-\infty, \infty]$, where $R_\infty$ is a Brownian motion with respect to time

$$T(x) = \int_{-\infty}^{x} \mathrm{Var}(Y/X = u)\, dF(u) = \int_{-\infty}^{x} \sigma^2(u)\, dF(u)$$

The null hypothesis is rejected if $\sqrt{n} d_1\left(I_n, I_0\right) = \sqrt{n}\sup_x \left|I_n(x) - I_0(x)\right|$ (the Kolmogorov-Smirnov statistic) exceeds a critical value computed from

the corresponding functional of the Brownian motion. The same with other statistics as for example the Cramér-von Mises statistic.
For a composite hypothesis, $m_{\theta_0}$ must be replaced by $m_{\theta_n}$, where $\theta_n$ is a root-n estimator of $\theta_0$, and the goodness-of-fit statistics need to be based on the process $R_n^1(x) = n^{1/2}\left(I_n(x) - I_0(x)\right)$ whose limit distribution, a centered Gaussian process $R_\infty^1$, is much more complicated (see again Stute (1997) for more details). The null hypothesis is now rejected if $\sqrt{n}d_1\left(I_n, I_{\theta_n}\right) = \sqrt{n}\sup_x |I_n(x) - I_{\theta_n}(x)|$ exceeds a critical value computed from the corresponding functional of $R_\infty^1$. Again, as in the tests developed in the last section, the bootstrap approximation to this critical value, can be a good solution. For example, for the Kolmogorov-Smirnov statistic and using wild bootstrap we would proceed in the following way:

$a_3$) As in the previous $a_1$).

$b_3$) As in the previous $b_2$).

$c_3$) Define the bootstrap sample by

$$Y_i^* = m_{\theta_n}\left(X_i\right) + \varepsilon_i^* \qquad i = 1, \ldots, n$$

and obtain $d_1\left(I_n^*, I_{\theta_n^*}\right)$, where $I_n^*$ and $I_{\theta_n^*}$ are constructed using the data $\{(X_i, Y_i^*)\}_{i=1}^n$.

$d_3$) Repeat the process B times and reject the hypothesis $H_0$ if $d_1\left(I_n, I_{\theta_n}\right) > c_{1-\alpha}^*$, $c_{1-\alpha}^*$ being the $[B(1-\alpha)]$ th order statistic computed from the B replicates $\left\{d_1\left(I_{ni}^*, I_{\theta_n^* i}^*\right)\right\} \quad i = 1, \ldots, B$.

## 4 A simulation example

In this section we report simulation results to demonstrate the validity of the proposed tests. We have considered the next regression model:

$$Y_i = 5X_i + aX_i^2 + \varepsilon_i \qquad 1 \leq i \leq n$$

where the $X_i$'s have been generated from the uniform distribution on the unit interval and the $\varepsilon$'s were drawn from the normal distribution $N(0, \sigma^2)$, with different values for $\sigma^2$, and $\varepsilon_i$ independent of $X_i$.
First the statistic $d_0\left(m_{\theta_n}, m_{n,\bar{q}}\right)$ has been calculated using a local linear smoother ($\bar{q} = 1$) and a local quadratic smoother ($\bar{q} = 2$) with three different smoothing parameters in each case. We have taken three sample sizes $n = 50, 100, 200$ and the wild bootstrap resampling has been performed $B = 100$ times for each sample. For each combination of factors we have

replicated the experiment 500 times and recorded the proportion of rejections under the null hypothesis (a=0) and under alternatives (a=1, 5) (see Table 1).

Table 1. Percentage of times $H_0$ was rejected.

| | | | Local linear | | | Local quadratic | | |
|---|---|---|---|---|---|---|---|---|
| | | | $h$ | | | $h$ | | |
| $\sigma^2$ | $a$ | $n$ | 0.1 | 0.25 | 0.6 | 0.1 | 0.25 | 0.6 |
| 1 | 0 | 50 | 4.5 | 4.8 | 4.9 | 4.3 | 4.4 | 4.6 |
| | | 100 | 4.8 | 5.0 | 5.2 | 4.7 | 5.1 | 5.0 |
| | | 200 | 4.7 | 4.9 | 4.6 | 4.8 | 5.2 | 4.8 |
| 1 | 1 | 50 | 9.6 | 8.7 | 8.6 | 9.2 | 11.0 | 11.7 |
| | | 100 | 19.3 | 18.1 | 15.7 | 19.6 | 20.3 | 20.7 |
| | | 200 | 28.7 | 26.0 | 24.0 | 30.0 | 31.2 | 31.0 |
| 1 | 5 | 50 | 77.0 | 76.1 | 74.0 | 82.1 | 82.5 | 81.9 |
| | | 100 | 98.0 | 96.8 | 96.5 | 99.4 | 100.0 | 99.2 |
| | | 200 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| 2 | 0 | 50 | 5.1 | 4.5 | 4.3 | 4.2 | 4.1 | 3.8 |
| | | 100 | 5.3 | 5.5 | 4.7 | 4.7 | 4.9 | 4.8 |
| | | 200 | 5.5 | 5.8 | 4.9 | 4.7 | 5.2 | 4.6 |
| 2 | 1 | 50 | 6.8 | 7.2 | 6.3 | 8.8 | 7.6 | 8.0 |
| | | 100 | 11.9 | 12.5 | 12.0 | 13.8 | 14.6 | 15.1 |
| | | 200 | 18.0 | 17.0 | 17.3 | 23.0 | 24.1 | 22.7 |
| 2 | 5 | 50 | 54.1 | 53.6 | 50.2 | 74.3 | 75.2 | 76.7 |
| | | 100 | 84.0 | 84.8 | 83.7 | 92.0 | 91.6 | 91.4 |
| | | 200 | 97.6 | 96.7 | 97.2 | 100.0 | 100.0 | 100.0 |
| 3 | 0 | 50 | 4.9 | 4.5 | 4.2 | 4.2 | 4.4 | 4.0 |
| | | 100 | 5.1 | 4.8 | 4.6 | 4.9 | 5.3 | 5.0 |
| | | 200 | 5.0 | 5.1 | 4.9 | 5.1 | 4.8 | 4.7 |
| 3 | 1 | 50 | 7.0 | 6.9 | 6.5 | 7.2 | 8.0 | 7.5 |
| | | 100 | 8.7 | 8.4 | 8.6 | 10.1 | 9.6 | 10.5 |
| | | 200 | 13.0 | 13.5 | 12.6 | 15.0 | 14.6 | 14.8 |
| 3 | 5 | 50 | 34.1 | 35.2 | 36.0 | 42.0 | 41.2 | 40.5 |
| | | 100 | 66.6 | 68.1 | 65.2 | 84.1 | 83.6 | 83.4 |
| | | 200 | 84.8 | 85.1 | 81.0 | 96.1 | 94.2 | 97.3 |

Secondly, the Kolmogorov-Smirnov statistic $d_1\left(I_n, I_{\theta_n}\right)$ has been calculated in the same regression model in one independent experiment with the same sample sizes with B=500 and where the experiment is replicated 1000 times.

In Table 2 the proportion of rejections under a=0, 1, 5 is given.

Table 2. Percentage of times $H_0$ was rejected.

| $\sigma^2$ | $a$ | $n$ | $\hat{\alpha}$ | $\sigma^2$ | $a$ | $n$ | $\hat{\alpha}$ |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 50 | 6.4 | 3 | 0 | 50 | 5.4 |
| | | 100 | 6.1 | | | 100 | 6.7 |
| | | 200 | 6.4 | | | 200 | 5.4 |
| 1 | 1 | 50 | 10.1 | 3 | 1 | 50 | 6.4 |
| | | 100 | 15.9 | | | 100 | 9.8 |
| | | 200 | 29.5 | | | 200 | 24.4 |
| 1 | 5 | 50 | 89.1 | 3 | 5 | 50 | 46.1 |
| | | 100 | 99.4 | | | 100 | 78.7 |
| | | 200 | 100.0 | | | 200 | 97.4 |
| 2 | 0 | 50 | 6.1 | | | | |
| | | 100 | 5.0 | | | | |
| | | 200 | 5.9 | | | | |
| 2 | 1 | 50 | 9.5 | | | | |
| | | 100 | 9.9 | | | | |
| | | 200 | 16.6 | | | | |
| 2 | 5 | 50 | 63.7 | | | | |
| | | 100 | 90.4 | | | | |
| | | 200 | 99.6 | | | | |

It is clear for both methods how the power of the tests increases as sample size n and parameter a get large; that is, as the alternative moves away from $H_0$. On the other hand, power decreases as $\sigma^2$ gets large. This effect is, of course, not unexpected, because when noise becomes more relevant, any procedure will have difficulty to distinguish between deviations caused by noise and by a different model. The nominal level is reasonably approximated by each method. In the papers Alcalá et al. (1999) and Stute et al. (1998) more simulations and examples (including this one) are given.

## 5 Some extensions of interest

### 5.1 Testing the hypothesis of a generalized linear regression model with binary response

When the variable $Y$ is binary (i.e., takes the values 0 and 1), our interest is in estimating the function $p(x) = \mathbb{P}(Y = 1/X = x) = \mathbb{E}[Y/X = x] = m(x)$. The same kind of problem as mentioned above (in the previous sections) consists now of testing $H_0 : p = p_\theta, \quad \theta \in \Theta \subset \mathbb{R}^q$, a parametric model for the conditional probability that $Y$ equals 1. One of the most popular models in the literature of discrete response regression is the logistic model

which, in the context of a one-dimensional variable $X$, is given by

$$p_\theta(x) = \frac{\exp(\theta_0 + \theta_1 x)}{1 + \exp(\theta_0 + \theta_1 x)} = m_\theta(x)$$

In general we focus on testing $H_0 : m \in \{m_\theta(\cdot)\}_{\theta \in \Theta}$ versus the alternative $H_1 : m$ is a smooth function, where $m_\theta(x) = g(\theta^t x)$, $\theta \in \Theta \subset \mathbb{R}^p$, $x \in \mathbb{R}^p$ with $g$ a known link function (logit, probit, etc.).
In Rodríguez-Campos et al. (1998),

$$d_2(m_{\theta_n}, m_{n,\bar{q}}) = \frac{1}{n} \sum_{j=1}^{n} \left(m_{n,\bar{q}}(X_j) - m_{\theta_n}(X_j)\right)^2 \omega(X_j)$$

is used as test statistic, where $\theta_n$ is a root-n estimator for $\theta$ (for example the maximum likelihood estimator in the logistic model) and $\bar{q} = 0$ (the Nadaraya-Watson estimator as nonparametric regression estimator). As in the case of Section 2 the asymptotic normality is obtained for $d_2(m_{\theta_n}, m_{n,\bar{q}})$ and again the bootstrap is a good option to approximate the critical values in the test. For example, in our binary case the binary bootstrap must be used following the next steps:

$a_4$) Obtain the bootstrap version of the response variable sample:

$$Y_i^* = \begin{cases} 1 & \text{with probability} p_{\hat{\theta}}(X_i) \\ 0 & \text{with probability} 1 - p_{\hat{\theta}}(X_i) \end{cases} \qquad i = 1, \ldots, n$$

where $\hat{\theta}$ is the maximum likelihood estimator of the parameter $\theta$, obtained from the sample $\{(X_i, Y_i)\}_{i=1}^n$.

$b_4$) With the bootstrap resample $\{(X_i, Y_i^*)\}_{i=1}^n$ construct $m_{n,\bar{q}}^*$ and $\theta_n^*$ and define $d_2\left(m_{\theta_n^*}, m_{n,\bar{q}}^*\right)$.

$c_4$) Repeat the process B times and reject the hypothesis $H_0$ if $d_2(m_{\theta_n}, m_{n,\bar{q}}) > c_{1-\alpha}^*$, $c_{1-\alpha}^*$ being the $[B(1-\alpha)]$ th order statistic computed from the B replicates $\left\{d_2\left(m_{\theta_n^* i}, m_{n,\bar{q}i}^*\right)\right\}_{i=1}^B$.

In the paper by Rodríguez-Campos et al. (1998) the good behaviour of this test is also shown with different simulation examples.

### 5.2 Testing the hypothesis of parametric models with censored response

Sometimes the variable $Y$ may be viewed as a lifetime or a monotone transformation of it, while $X$ is a vector of covariables to be sampled at the entry into or in the course of a follow-up study. In the analysis of such survival data it is typical that, due to losses or other failures, $Y$ is not always available. Hence standard statistical methods which require knowledge of

all $Y$'s are not applicable. Under random right censorship, rather than $Y$, one observes $Z = \min\{Y, C\}$ where $C$ is a censoring variable. Hence the available data are $(X_i, Z_i, \delta_i)$ $1 \leq i \leq n$, where $\delta_i = 1_{\{Y_i \leq C_i\}}$ indicates the cause of failure. In this context $I(x)$ can be consistently estimated as

$$\hat{I}_n(x) = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} y \; d\hat{F}_n(u, v)$$

where $\hat{F}_n(x, y) = \sum_{i=1}^{n} W_{ni} 1_{\{X_{[i:n]} \leq x,\; Z_{i:n} \leq y\}}$ is an extension of $F_n$ to the censored data case, with $Z_{1:n} \leq \ldots \leq Z_{n:n}$ the $Z$-order statistics,

$$W_{ni} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left[ \frac{n-j}{n-j+1} \right]^{\delta_{[j:n]}}$$

the Kaplan-Meier weight attached to $Z_{i:n}$, and $\delta_{[i:n]}$, $X_{[i:n]}$ denoting the $\delta$ and $X$ concomitants associated with $Z_{i:n}$ (see Stute (1993, 1996, 1999) for more details).
To derive tests for $H_0 : m \in \{m_\theta\}_{\theta \in \Theta}$ we must now consider:

$$\hat{I}_n(x) - I_{\theta_n}(x) = n^{-1} \sum_{i=1}^{n} W_{ni} 1_{\{X_{[i:n]} \leq x\}} \left( Z_{i:n} - m_{\theta_n} \left( X_{[i:n]} \right) \right)$$

where $\sqrt{n} \left( \hat{I}_n - I_{\theta_n} \right)$ has a nondegenerate limit distribution (see Stute et al. (2000) for more details). In fact it is shown that uniformly in $x$,

$$R_n^2(x) = \sqrt{n} \left( \hat{I}_n(x) - I_{\theta_n}(x) \right) = n^{-1/2} \sum_{i=1}^{n} \xi_i(x) + o_{\mathbb{P}}(1) = R_n^3(x) + o_{\mathbb{P}}(1)$$

with $\xi_i(x)$ a random variable of mean zero based on $(X_i, Z_i, \delta_i)$ and on different parameters of the model. Under regularity conditions $R_n^2$ has the same limit distribution as $R_n^3$, a centered Gaussian process $R_\infty^2$. The null hypothesis is now rejected if

$$\sqrt{n} d_1 \left( \hat{I}_n, I_{\theta_n} \right) = \sqrt{n} \sup_x \left| \hat{I}_n(x) - I_{\theta_n}(x) \right| \approx \sqrt{n} \sup_x \left| n^{-1} \sum_{i=1}^{n} \xi_i(x) \right)$$

exceeds a critical value computed from the corresponding functional of $R_\infty^2$. The bootstrap approximation to this critical value is again a good solution. Now, rather than $R_n^2$, the wild bootstrap is taken over the leading process $R_n^3$:

$$R_n^{3*}(x) = n^{-1/2} \sum_{i=1}^{n} \hat{\xi}_i(x) V_i^*$$

where the $V_i^*$'s are i.i.d. random variables with expectation zero and variance one, being also independent of the original sample, and $\hat{\xi}_i$ an empirical estimation of $\xi_i$. Here $H_0$ is rejected if $d_1 \left( \hat{I}_n, I_{\theta_n} \right) > c_{1-\alpha}^*$, $c_{1-\alpha}^*$

being the $[B(1-\alpha)]$ th order statistic computed from the $B$ replicates $\left\{\sup_x \left|n^{-1}\sum_{j=1}^n \hat{\xi}_j(x)V_{ji}^*\right|\right\} \qquad i=1,\ldots,B.$

**Example. Stanford Heart Transplant Data.**
We now illustrate our method using data from the Stanford Heart Transplant program. Between October 1967 and February 1980, 184 of the 249 patients admitted to this program received a heart transplantation. We focus on only two variables, with the survival time as the response and age as the covariate. Patients alive beyond February 1980 were considered censored. We concentrate our analysis on the 157 patients out of 184 who had complete tissue typing. Among the 157 cases, 55 were censored.
The hypothesis of a linear model for $\log_{10}$ of survival time versus age was checked. The bootstrap approximations (with 10000 replicates) were used to obtain the $p$-values of the test statistics for this linear model. The $p$-value of $d_1\left(\hat{I}_n, I_{\theta_n}\right)$ was 0.8236.
Our $p$-values are typically larger than the ones obtained by other methods in the literature. This may be explained by the fact that our method incorporates Kaplan-Meier weights and thus attaches mass zero to the residuals related to censored data. Hence our approach is much more cautious about rejecting a model in particular when deviations are caused by censoring and not necessarily by a wrong assumption on the model.

## 6 Comments

The two methodologies (based on smoothing and on empirical regression processes, respectively) discussed in the different sections of this tutorial work are also applicable to test different restrictions on the nonparametric regression curve. Some relevant examples are the following:

1. "The significance testing".

$$H_0 : \mathbb{E}\left[Y/X\right] = \mathbb{E}\left[Y/X_1\right]$$

   where $X^t = (X_1^t, X_2^t)$ is the complete predictor vector of variables of dimension $p = p_1 + p_2$. It would be very useful to reduce the number of explanatory variables in the regression curve as much as possible.

2. "Testing of partially linear models".

$$H_0 : \mathbb{E}\left[Y/X\right] = X_1^t\theta_0 + \gamma\left(X_2\right)$$

   $\theta_0 \in \Theta \subset \mathbb{R}^{p_1}$ and $\gamma$ is an unknown function.

3. "Testing the additivity".

$$H_0 : \mathbb{E}\left[Y/X\right] = \sum_{j=1}^{p} f_j\left(X_j\right)$$

where $\{f_j\}_{j=1}^p$ are $p$ unknown functions.

The common factor in all these null hypotheses is the infinite dimensionality. Fan and Li (1996) using smoothing and Delgado and González-Manteiga (2000) with empirical regression processes are two examples of the very recent work in this field.

**Acknowledgements:** This work was partially supported by the Spanish Ministry of Education and Culture with the Grant PB98–0182–C02–02 and also by the Grant PGIDT99MA 20701 from "Xunta de Galicia" in Spain. The author is also very grateful to César Sánchez Sellero for his careful supervision of the paper.

## References

Alcalá, J.T., Cristóbal, J.A. and González-Manteiga, W. (1999). Goodness-of-fit Test for Linear Models Based on Local Polynomials *Statistics and Probability Letters*, **42**, 39–46.

Delgado, M.A. and González-Manteiga, W. (2000). Significance testing in nonparametric regression based on the bootstrap. Preprint.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865–890.

Mc Cullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* (Second Edition). Chapman and Hall, London.

Rodríguez-Campos, M.C., González-Manteiga, W. and Cao, R. (1998). Testing the Hypothesis of a Generalized Linear Regression Model Using Nonparametric Regression Estimation. *Journal of Statistical Planning and Inference*, **67**, 99–122.

Seber, G.A.F. (1977). *Linear Regression Analysis.* John Wiley and Sons, New York.

Speckman, P. (1988). Kernel Smoothing in Partially Linear Models. *Journal of the Royal Statistical Society, series B*, **50**, 413–446.

Stute, W. (1993). Consistent Estimation under Random Censorship when Covariables are Present. *Journal of Multivariate Analysis*, **45**, 89–103.

Stute, W. (1996). Distributional Convergence under Random Censorship when Covariables are Present. *Scandinavian Journal of Statistics*, **23**, 461–471.

Stute, W. (1997). Nonparametric Model Checks for Regression. *The Annals of Statistics*, **25**, 613–641.

Stute, W. (1999). Nonlinear Censored Regression. *Statistica Sinica*, **9**, 1089–1102.

Stute, W., González-Manteiga, W. and Presedo-Quindimil, M. (1998). Bootstrap Approximations in Model Checks for Regression. *Journal of the American Statistical Association*, **93**, 441, 141–149.

Stute, W., González-Manteiga, W. and Sánchez-Sellero, C. (2000). Nonparametric Model Checks in Censored Regression. *Communications in Statistics, Theory and Methods*, to appear.

# Semiparametric and Nonparametric Estimation in Econometrics

Joel L. Horowitz[1]

[1] Department of Economics, University of Iowa, Iowa City, IA 52242, U.S.A.

**Abstract:** Much empirical research in economics is concerned with estimating conditional mean functions. The most frequently used estimation methods assume that the conditional mean function is known up to a set of constant parameters that can be estimated from data. Such methods are called *parametric*. Their use greatly simplifies estimation and inference but is rarely justified by theoretical or other *a priori* considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading. Semiparametric methods reduce the strength of the assumptions required for estimation and inference, thereby reducing the opportunities for obtaining misleading results. In addition, semiparametric methods mitigate certain disadvantages of fully nonparametric methods that make no assumptions about the shape of the conditional mean function. This article describes three important semiparametric models for conditional mean functions. These are single index, partially additive, and additive models. They are compared with parametric and fully nonparametric models.

## 1 Introduction

Much empirical research in economics is concerned with estimating conditional mean functions. For example, labor economists are interested in estimating the mean wages of employed individuals conditional on characteristics such as years of education and work experience. The most frequently used estimation methods assume that the conditional mean function is known up to a set of constant parameters that can be estimated from data, possibly by ordinary least squares. Models in which the only unknown quantities are a finite set of constant parameters are called *parametric*. The use of a parametric model greatly simplifies estimation, statistical inference, and interpretation of the estimation results but is rarely justified by theoretical or other *a priori* considerations. Estimation and inference based on convenient but incorrect assumptions about the form of the conditional mean function can be highly misleading. Semiparametric statistical methods reduce the strength of the assumptions required for estimation and inference, thereby reducing the opportunities for obtaining misleading results. These methods are applicable to a wide variety of estimation problems in economics and other fields.

A conditional mean function gives the mean of a dependent variable $Y$ conditional on a vector of explanatory variables $X$. Denote the mean of $Y$ conditional on $X = x$ by $\mathbf{E}(Y|x)$. For example, suppose that $Y$ is a worker's weekly wage (or, more often in applied econometrics, the logarithm of the wage) and $X$ includes such variables as years of education, years of labor force experience, race, and sex. Then $\mathbf{E}(Y|x)$ is the mean wage (or logarithm of the wage) when education and the other explanatory variables have the values specified by $x$. In applications, $\mathbf{E}(Y|x)$ is unknown and must be estimated from data on the variables of interest. In the case of estimating a wage function, the data consist of observations of individuals' wages, years of education, and other characteristics. The most widely used method for estimating $\mathbf{E}(Y|x)$ assumes that this function is known up to finitely many constant parameters. This gives a *parametric model* for $\mathbf{E}(Y|x)$ . Often, $\mathbf{E}(Y|x)$ is assumed to be a linear function of $x$, in which case the parameters can be estimated by ordinary least squares (OLS), among other ways. OLS estimators are described many textbooks. See, for example, Goldberger (1998). However, the OLS estimator of $\mathbf{E}(Y|x)$ can be highly misleading if $\mathbf{E}(Y|x)$ is not linear in the components of $x$, that is if there is no $\beta$ such that $\mathbf{E}(Y|x) = \beta' x$ .

The opportunities for specification error increase if $Y$ is binary. For example, consider a model of the choice of travel mode for the trip to work. Suppose that the available modes are automobile and transit. Let $Y = 1$ if an individual chooses automobile and $Y = 0$ if the individual chooses transit. Let $X$ be a vector of explanatory variables such as the travel times and costs by automobile and transit. Then $\mathbf{E}(Y|x)$ is the probability that $Y = 1$ (the probability that the individual chooses automobile) conditional on $X = x$. This probability will be denoted $\mathbf{P}(Y = 1|x)$ . In applications of binary response models, it is often assumed that $\mathbf{P}(Y|x) = G(\beta' x)$ , where $\beta$ is a vector of constant coefficients and $G$ is a known probability distribution function. Often, $G$ is assumed to be the cumulative standard normal distribution function, which yields a *binary probit* model, or the cumulative logistic distribution function, which yields a *binary logit* model. The coefficients $\beta$ can then be estimated by the method of maximum likelihood (Amemiya 1985). However, there are now two potential sources of specification error. First, the dependence of $Y$ on $x$ may not be through the linear *index* $\beta' x$ . Second, even if the index $\beta' x$ is correct, the *response function* $G$ may not be the normal or logistic distribution function. See Horowitz (1993, 1998) for examples of specification errors in binary response models and their consequences.

Many investigators attempt to minimize the risk of specification error by carrying out a *specification search* in which several different models are estimated and conclusions are based on the one that appears to fit the data best. However, there is no guarantee that a specification search will include the correct model or a good approximation to it. If the search includes the correct model, there is no guarantee that it will be selected by the investiga-

tor's model selection criteria. Moreover, the search process invalidates the statistical theory on which inference is based. Thus, a specification search is not satisfactory way of dealing with the problem of specification error. The rest of this paper describes methods that deal with the problem of specification error by relaxing the assumptions about functional form that are made by parametric models. The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. These are described in Section 2. They assume that $\mathbf{E}(Y|x)$ is a smooth function but make no other assumptions about its shape or functional form. However, nonparametric methods have important disadvantages that seriously limit their usefulness in applications. Semiparametric methods, which are described in Section 3, offer a compromise. They make assumptions about functional form that are stronger than those of a nonparametric model but less restrictive than the assumptions of a parametric model, thereby reducing (though not eliminating) the possibility of specification error. In addition semiparametric methods avoid the most serious practical disadvantages of nonparametric methods.

## 2 Nonparametric Models

In nonparametric estimation $\mathbf{E}(Y|x)$ is assumed to satisfy smoothness conditions such as differentiability, but no assumptions are made about its shape or the form of its dependence on $x$. Härdle (1990) and Fan and Gijbels (1996) provide detailed discussions of nonparametric estimation methods. One easily understood and frequently used method is called *kernel estimation.* To describe the kernel method simply, assume that $X$ is a continuously distributed, scalar random variable. Let $\{Y_i, X_i : i = 1, \ldots, n\}$ be a random sample of n observations of $(Y, X)$ . Let $K$ be a probability density function that is bounded, continuous, and symmetrical about zero. For example, $K$ may be the standard normal density function. Let $\{h_n\}$ to be a sequence of positive numbers that converges to 0 as $n \to \infty$. For each $n = 1, 2, \ldots$ and $i = 1, \ldots, n$ define the function $w_{ni}(.)$ by

$$w_{ni} = \frac{k[(x - X_i)/h_n]}{\sum_{i=1}^{n} K[(x - X_i)/h_n]}.$$

Then the kernel nonparametric estimator of $\mathbf{E}(Y|x)$ is

$$H_n(x) = \sum_{i=1}^{n} w_{ni}(x) Y_i.$$

$H_n(x)$ is a weighted average of the observed values of $Y$. Observations $Y_i$ for which $X_i$ is close to $x$ get higher weight than do observations for which $X_i$ is far from $x$. It can be shown that if $h_n \to 0$ and $nh_n/(\log n) \to \infty$ as $n \to \infty$, then $H_n(x) \to \mathbf{E}(Y|x)$ with probability 1. Thus, if $n$ is large,

$H_n(x)$ is likely to be very close to $\mathbf{E}(Y|x)$ . Härdle (1990) provides a detailed discussion of the statistical properties of kernel nonparametric estimators. Nonparametric estimation minimizes the risk of specification error, but the price of this flexibility can be high. One important reason for this is that the precision of a nonparametric estimator decreases rapidly as the number of continuously distributed components of $X$ increases. This phenomenon is called the *curse of dimensionality*. As a result of it, impracticably large samples are usually needed to obtain acceptable estimation precision if $X$ is multidimensional, as it often is in econometric applications.

A further problem is that nonparametric estimates can be difficult to display, communicate, and interpret when $X$ is multidimensional. Nonparametric estimates do not have simple analytic forms. If $X$ is one- or two-dimensional, then the estimate of $\mathbf{E}(Y|x)$ can be displayed graphically, but only reduced-dimension projections can be displayed when $X$ has three or more components. Many such displays and much skill in interpreting them can be needed to fully convey and comprehend the shape of the estimate of $\mathbf{E}(Y|x)$.

Another problem with nonparametric estimation is that it does not permit extrapolation. That is, it does not provide predictions of $\mathbf{E}(Y|x)$ at points $x$ that are outside of the support (or range) of the random variable $X$ . This is a serious drawback in policy analysis and forecasting, where it is often important to predict what might happen under conditions that do not exist in the available data. Finally, it can be difficult to impose the restrictions of economic or other theory models in nonparametric estimation. Matzkin (1994) discusses this issue.

Semiparametric methods permit greater estimation precision than do nonparametric methods when $X$ is multidimensional. In addition, semiparametric estimates are easier to display and interpret than nonparametric ones and provide limited capabilities for extrapolation and imposing restrictions derived from economic or other theory models.

## 3 Semiparametric Models

The term *semiparametric* refers to models in which there is an unknown function in addition to an unknown finite dimensional parameter. For example, the binary response model $\mathbf{P}(Y = 1|x) = G(\beta' x)$ is semiparametric if the function $G$ and the vector of coefficients $\beta$ are both treated as unknown quantities. This section describes two semiparametric models of conditional mean functions that are important in applications. The section also describes a related class of models that has no unknown finite-dimensional parameters but, like semiparametric models, mitigates the disadvantages of fully nonparametric models.

In addition to the estimation of conditional mean functions, semiparametric methods can be used to estimate conditional quantile and hazard func-

tions, binary response models in which there is heteroskedasticity of unknown form, transformation models, and censored and truncated mean- and median- regression models, among others. Horowitz (1998) and Powell (1994) provide more comprehensive treatments in which these models are discussed.

### 3.1 Single Index Models

In a semiparametric single index model, the conditional mean function has the form

$$\mathbf{E}(Y|x) = G(\beta' x) \tag{1}$$

where $\beta$ is an unknown constant vector and $G$ is an *unknown* function. The quantity $\beta' x$ is called an *index* . The inferential problem is to estimate $G$ and $\beta$ from observations of $(Y, X)$ .

Model (1) contains many widely used parametric models as special cases. For example, if $G$ is the identity function, then (1) is a linear model. If $G$ is the cumulative normal or logistic distribution function, then (1) is a binary probit or logit model. When $G$ is unknown, (1) provides a specification that is more flexible than a parametric model but retains many of the desirable features of parametric models, as will now be explained.

One important property of single index models is that they avoid the curse of dimensionality. This is because the index $\beta' x$ aggregates the dimensions of $x$, thereby achieving *dimension reduction.* Consequently, the difference between the estimator of $G$ and the true function can be made to converge to zero at the same rate that would be achieved if $\beta' x$ were observable. Moreover, $\beta$ can be estimated with the same rate of convergence that is achieved in a parametric model. Thus, in terms of the rates of convergence of estimators, a single index model is as accurate as a parametric model for estimating $\beta$ and as accurate as a one-dimensional nonparametric model for estimating $G$. This dimension reduction feature of single index models gives them a considerable advantage over nonparametric methods in applications where $X$ is multidimensional and the single index structure is plausible.

A single-index model permits limited extrapolation. Specifically, it yields predictions of $\mathbf{E}(Y|x)$ at values of $x$ that are not in the support of $X$ but are in the support of $\beta' x$. Of course, there is a price that must be paid for the ability to extrapolate. A single index model makes assumptions that are stronger than those of a nonparametric model. These assumptions are testable on the support of $X$ but not outside of it. Thus, extrapolation (unavoidably) relies on untestable assumptions about the behavior of $\mathbf{E}(Y|x)$ beyond the support of $X$.

Before $\beta$ and $G$ can be estimated, restrictions must be imposed that insure their identification. That is, $\beta$ and $G$ must be uniquely determined by the population distribution of $(Y, X)$. Identification of single index models has

been investigated by Ichimura (1993) and, for the special case of binary response models, Manski (1988). It is clear that $\beta$ is not identified if $G$ is a constant function or there is an exact linear relation among the components of $X$ (perfect multicollinearity). In addition, (1) is observationally equivalent to the model $\mathbf{E}(Y|x) = G^*(\gamma + \delta\beta' x)$ , where $\gamma$ and $\delta$ are arbitrary and $G^*$ is defined by the relation $G^*(\gamma + \delta\nu) = G(\nu)$ for all $\nu$ in the support of $\beta' x$. Therefore, $\beta$ and $G$ are not identified unless restrictions are imposed that uniquely specify $\gamma$ and $\delta$. The restriction on $\gamma$ is called *location normalization* and can be imposed by requiring $X$ to contain no constant (intercept) component. The restriction on $\delta$ is called *scale normalization.* Scale normalization can be achieved by setting the $\beta$ coefficient of one component of $X$ equal to one. A further identification requirement is that $X$ must include at least one continuously distributed component whose $\beta$ coefficient is non-zero. Horowitz (1998) gives an example that illustrates the need for this requirement. Other more technical identification requirements are discussed by Ichimura (1993) and Manski (1988).

The main estimation challenge in single index models is estimating $\beta$ . Given an estimator $b_n$ of $\beta$, $G$ can be estimated by carrying out the nonparametric regression of $Y$ on $b_n' X$ (e.g, by using the kernel method described in Section 2). Several of estimators of $\beta$ are available. Ichimura (1993) describes a nonlinear least squares estimator. Klein and Spady (1993) describe a semiparametric maximum likelihood estimator for the case in which $Y$ is binary. These estimators are difficult to compute because they require solving complicated nonlinear optimization problems. Powell, et al. (1989) describe a *density-weighted average derivative estimator* (DWADE) that is non-iterative and easily computed. The DWADE applies when all components of $X$ are continuous random variables. It is based on the relation

$$\beta \propto \mathbf{E}[p(X)\partial G(\beta' X)/\partial X] = -2\mathbf{E}[Y\partial p(X)/\partial X], \tag{2}$$

where $p$ is the probability density function of $X$ and the second equality follows from integrating the first by parts. Thus, $\beta$ can be estimated up to scale by estimating the expression on the right-hand side of the second equality. Powell, et al. (1989) show that this can be done by replacing $p$ with a nonparametric estimator and replacing the population expectation $\mathbf{E}$ with a sample average. See Powell, et al. (1989) for details. Horowitz and Härdle (1996) extend this method to models in which some components of $X$ are discrete. Horowitz and Härdle (1996) also give an empirical example that illustrates the usefulness of single index models. Ichimura and Lee (1991) investigate a multiple index generalization of (1).

### 3.2 Partially Linear Models

In a partially linear model, $X$ is partitioned into two non-overlapping subvectors, $X_1$ and $X_2$. The model has the form

$$\mathbf{E}(Y|x_1, x_2) = \beta' x_1 + G(x_2), \tag{3}$$

where $\beta$ is an unknown constant vector and $G$ is an unknown function. This model is distinct from the class of single index models. A single index model is not partially linear unless $G$ is a linear function. Conversely, a partially linear model is a single index model only in this case. Stock (1989, 1991) and Engle et al. (1986) illustrate the use of (3) in applications. Identification of $\beta$ requires the *exclusion restriction* that none of the components of $X_1$ are perfectly predictable by components of $X_2$ . When $\beta$ is identified, it can be estimated with an $n^{-1/2}$ rate of convergence regardless of the dimensions of $X_1$ and $X_2$ . Thus, the curse of dimensionality is avoided in estimating $\beta$ .
An estimator of $\beta$ can be obtained by observing that (3) implies

$$Y - \mathbf{E}(Y|x_2) = \beta'[X_1 - \mathbf{E}(X_1|x_2)] + U, \tag{4}$$

where $U$ is an unobserved random variable satisfying $\mathbf{E}(U|x_1, x_2) = 0$ . Robinson (1988) shows that under regularity conditions, $\beta$ can be estimated by applying OLS to (4) after replacing $\mathbf{E}(Y|x_2)$ and $\mathbf{E}(X_1|x_2)$ with nonparametric estimators. The estimator of $\beta$, $b_n$, converges at rate $n^{-1/2}$ and is asymptotically normally distributed. $G$ can be estimated by carrying out the nonparametric regression of $Y - b_n' X_1$ on $X_2$ . Unlike $b_n$, the estimator of $G$ suffers from the curse of dimensionality; its rate of convergence decreases as the dimension of $X_2$ increases.

### 3.3 Nonparametric Additive Models

Let $X$ have $d$ continuously distributed components that are denoted $X_1, \ldots, X_d$. In a nonparametric additive model of the conditional mean function,

$$\mathbf{E}(Y|x) = \mu + f_1(x_1) + \ldots + f_d(x_d), \tag{5}$$

where $\mu$ is a constant and $f_1 + \ldots + f_d$ are unknown functions that satisfy a location normalization condition such as

$$\int f_k(\nu) d\nu = 0, \qquad k = 1, \ldots, d. \tag{6}$$

An additive model is distinct from a single index model unless $\mathbf{E}(Y|x)$ is a linear function of $x$. Additive and partially linear models are distinct unless $\mathbf{E}(Y|x)$ is partially linear and $G$ in (3) is additive.

An estimator of $f_k$ $(k = 1, \ldots, d)$ can be obtained by observing that (5) and (6) imply

$$f_k(x_k) = \int \mathbf{E}(Y|x) w_{-k}(x_{-k}) dx_{-k}, \tag{7}$$

where $x_{-k}$ is the vector consisting of all components of $x$ except the $k'$th and $w_{-k}$ is a weight function that satisfies $\int w_{-k}(x_{-k}) dx_{-k}$. The estimator of $f_k$ is obtained by replacing $\mathbf{E}(Y|x)$ on the right-hand side of (7) with a nonparametric estimator. Linton and Nielsen (1995) and Linton (1997) present the details of the procedure and extensions of it. Under suitable conditions, the estimator of $f_k$ converges to the true $f_k$ at rate $n^{-2/5}$ regardless of the dimension of $X$. Thus, the additive model provides dimension reduction. It also permits extrapolation of $\mathbf{E}(Y|x)$ within the rectangle formed by the supports of the individual components of $X$.
Linton and Härdle (1996) describe a generalized additive model whose form is

$$\mathbf{E}(Y|x) = G[\mu + f_1(x_1) + \ldots + f_K(x_d)], \tag{8}$$

where $f_1, \ldots, f_d$ are unknown functions and $G$ is a known, strictly increasing (or decreasing) function. Horowitz (2000) describes a version of (8) in which $G$ is unknown. Both forms of (8) achieve dimension reduction. When $G$ is unknown, (8) nests additive and single index models and, under certain conditions, partially linear models.

**Acknowledgements**: Research supported in part by NSF Grant SBR-96-17925.

## References

Amemiya, T (1985) *Advanced Econometrics.* Harvard University Press, Cambridge, MA.

Engle, R F, Granger C W J, Rice J, and Weiss A (1986) Semiparametric estimates of the relationship beteween weather and electricity sales. *Journal of the American Statistical Association* 81: 310-320.

Fan, J and Gijbels I (1996) *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London.

Goldberger, A S (1998) *Introductory Econometrics.* Harvard University Press, Cambridge, MA.

Härdle, W. (1990) *Applied Nonparametric Regression.* Cambridge University Press, Cambridge

Horowitz, J.L (1993) Semiparametric and nonparametric estimation of quantal response models. In: Maddala, G S, Rao, C R, and Vinod H D (eds.) *Handbook of Statistics, Vol. 11.* Elsevier, Amsterdam, pp. 45-72.

Horowitz, J.L (1998) *Semiparametric Methods in Econometrics.* Springer-Verlag, New York.

Horowitz, J L (2000) Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica,* forthcoming.

Horowitz J L and Härdle W (1996) Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association* 91: 1632-1640.

Ichimura, H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics* 58: 71-120.

Ichimura, H and Lee L-F (1991) Semiparametric least squares estimation of multiple index models: single equation estimation. In: Barnett W A, Powell J, and Tauchen G (eds) *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge University Press, Cambridge, pp. 3-49.

Klein R W and Spady R H (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61: 387-421.

Linton O B (1997) Efficient estimation of additive nonparametric regression models. *Biometrika* 84: 469-473.

Linton O B and Härdle, W (1996) Estimating additive regression models with known links. *Biometrika* 83: 529-540.

Linton O B and Nielsen J P (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* 82: 93-100.

Manski, C F (1998) Identification of binary response models. *Journal of the American Statistical Association* 83: 729-738.

Powell J L (1984) Estimation of semiparametric models. In: Engle R F and McFadden D L (eds) *Handbook of Econometrics,* Vol. 4. North-Holland, Amsterdam, pp. 2444-2521.

Powell J L, Stock J H, and Stoker T M (1989) Semiparametric estimation of index coefficients. *Econometrica* 51: 1403-1430.

Robinson, P M (1988) Root-N-consistent semiparametric regression. *Econometrica* 56: 931-954.

Stock J H (1989) Nonparametric policy analysis. *Journal of the American Statistical Association* 84: 567-575.

Stock J H (1991) Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits. In: Barnett W A, Powell J, and Tauchen G (eds) *Nonparametric and Semiparametric Methods in Econometrics and Statistics.* Cambridge University Press, Cambridge, pp. 77-98.

# Statistical Techniques for Quality Improvement: Improving the Manufacture of Viscose Fiber

Johannes Ledolter[1]

[1] Department of Statistics, Vienna University of Economics and Business Administration, Augasse 2-6, A-1090 Vienna, Austria Johannes.Ledolter@wu-wien.ac.at

**Abstract:** Continuous improvement of production and business processes is necessary if companies want to remain competitive . Statistics and statistical tools for stabilizing and improving processes are important components of improvement efforts. In this paper we illustrate how statistical techniques help improve the quality of a manufacturing process. Our investigation involves the production of a viscose fiber, referred to in this paper as fiber "M".

**Keywords:** Design of experiments; errors-in-variables model; fractional factorial design; regression.

## 1 Introduction

Fiber "M" is a wood-based product, made from beech trees. A vertical production process first transforms wood into pulp, and then pulp into the viscose-like fiber "M". The process from wood to pulp starts with wood chips being cooked in huge pressurized vats of a magnesium-bisulfite solution for approximately eight hours. The resulting pulp is bleached (using peroxide, sodium hydroxide, and ozone), drained, dried and pressed into cellulose fibers. The second stage of the process, the transformation of pulp into viscose fiber, involves several chemical and mechnical processes. Cellulose pulp is first treated with sodium hydroxide, a process that results in alkaline cellulose, a swollen crumbly mass. Excess sodium hydroxide is drained and the resulting mixture is pressed and shreded. A "ripening" process, where alkaline cellulose is stored for a certain period of time at 30 - 45 degrees C, in the alkaline cellulose is reacted with carbon disulfide and transformed into cellulose xanthate. Through the addition of water and sodium hydroxide the xanthate is then dissolved into a liquid form, called the viscose. After mechanical purifying steps the viscose is spun into viscose fibers. This is achieved by pressing viscose through very fine nozzles into a spinnbath solution containing sulfuric acid, and sodium and zinc sulfates, whereby the cellulose is regenerated. The resulting fibers are stretched, bundled, cut, washed and dried. The process from wood to viscose fiber takes approx. 34

hours. The special processes for fiber "M" differ from the standard viscose technology in that spinning conditions, additives and precipitating baths are changed somewhat resulting in better fiber performance. Fiber strength and fiber elongation are the most important quality characteristic, and the following report looks at these two quality measures.

## 2 A Designed Experiment on the Pilot Plant: Effects of Process Changes on Strength and Elongation

Fiber properties such as strength and elongation depend on the settings of numerous process factors. Relationships with some factors are well understood - either from theory or from years of practical experience - and great care is used to control these driving factors. However, dependencies with many other variables are neither known nor understood.

While statistical studies of historic process data are useful in uncovering relationships, such analyses have also shortcomings. They are often ineffective in uncovering dependencies as changes in controlled process factors either don't occur during the observed period or, if they do, are not large enough to trigger a response. Furthermore, important factors often change at the same time which makes it difficult to uncover interactions. Much more information can be gained from data if changes to the process factors are introduced deliberately according to a well-designed experiment.

We now discuss the design of such an experiment and the analysis of the resulting data. This particular experiment is conducted on the pilot plant, a small-scale production facility that simulates the production processes in the plant on a smaller scale. The goal of this experiment is to find and quantify the forces that influence strength and elongation of fiber "M".

### 2.1 The Design of the Experiment

We study the effects of ten process factors: three factors that have to do with the composition of the (liquid) viscose [that is, Cell (cellulose in percent), AV (ratio of sodium hydroxide to cellulose), and $CS_2$ (carbon disulfide, in percent)], temperature of the viscose during the "ripening" process, four factors that change the spinnbath [that is, the chemicals $H_2SO_4$, $Na_2SO_4$, $ZnSO_4$, and the temperature of the spinnbath], and two factors for making adjustments on the spinning machine [amount of stretch reduction and drawing speed]. These ten factors influence the second stage of the process from pulp to fiber.

A two-level fractional factorial design: $2^k$ factorial und $2^{k-p}$ fractional factorial designs are useful search designs as they can uncover important driving forces with relatively few experiments. An important feature of these

designs is that they don't just vary one factor at a time, but that they change factors together allowing the experimenter to uncover main effects as well as interactions. Excellent discussions of these designs can be found in Box and Draper (1969), Box, Hunter and Hunter (1978), Mason, Gunst and Hess (1989), Montgomery (1991), Hogg and Ledolter (1991), and Ledolter and Burrill (1999).

In a (full) $2^k$ faktorial experiment each of the k process factors is studied at a low and a high setting. The low and high settings are obtained by varying the currently used target value; these levels must be different enough so that one can estimate the effect of their change. However, these values must not be too different as performance problems must be kept low.

A disadvantage of a full factorial design is the large number of needed runs. Fractional factorial designs which consider only certain well-chosen fractions of the factorial plans lead to much more economical arrangements. In this paper we use a $2^{10-5}$ fractional factorial design with ten factors in 32 runs. A $2^5$ factorial design in the five factors [Cell, AV, $CS_2$, VisTemp, und $ZnSO_4$] is taken as the starting point of this design. The generators of the remaining five factors are taken as:

$$\begin{array}{ccc} H_2SO_4 & = & (\text{Cell})(\text{AV})(CS_2)(\text{VisTemp}) \\ Na_2SO_4 & = & (\text{Cell})(\text{AV})(CS_2)(ZnSO_4) \\ \text{SpTemp} & = & (CS_2)(\text{VisTemp})(ZnSO_4) \\ \text{Stretch} & = & (\text{AV})(\text{VisTemp})(ZnSO_4) \\ \text{Draw} & = & (\text{Cell})(\text{VisTemp})(ZnSO_4) \end{array}$$

The factors levels of the 32 runs, together with the results on fiber strength and fiber elongation, are listed in Table 1. The resolution of this design is IV. Ignoring interactions of order 3 or higher, this design allows us to estimate the main effects of all ten factors. Two-factor interactions, however, are confounded with certain other two-factor interactions; see Table 2.

<u>Block Structure:</u> Ideally, the temporal arrangement of the runs should be randomized. Randomization helps avoid biases that otherwise could get introduced due to unknown patterns of the process over time. A complete randomization is not possible as our experiment has to be carried out over several days. Table 1 shows that the experiment is blocked into eight days of four runs each. The liquid viscose depends on the three factors: Cell, AV, and $CS_2$. A batch of viscose is made up for each of the eight factor-level combinations, and on each day of the experiment one such batch is used for further study. The block structure is shown in Table 1.

Batch and day effects may be present. The blocking factors $B_1 = Cell, B_2 = AV, B_3 = CS_2$ introduce the additional confounding patterns. The blocking arrangement implies that all main effects and interactions of the three

factors Cell, AV, and $CS_2$ are confounded with block effects. For example, a large effect associated with AV may equally be due to a batch effect. Analysis of the confounding patterns also shows that the interactions (VisTemp)$(H_2SO_4) \approx (ZnSO_4)(Na_2SO_4)$ are confounded with the block effect. All other main and interaction effects remain unaffected by block effects.

Stretch factor: The amount of stretch that is applied by the spinning machine as fibers are pulled through the spinning bath is an important factor. The largest possible stretch is determined for each of the 32 runs. This is done by successively increasing the stretch until fiber problems become visible. A reduction of 30 precent from the largest possible stretch is taken as the low setting; a reduction of ten perecent is used as the high level. The effect of a 20 percent reduction is also studied, in addition to the 30 or 10 percent reductions that are part of the fractional factorial design.

**Table 1:** The $2^{10-5}$ fractional factorial design, and resulting measures of fiber strength and fiber elongation.

| | | Viscose | | | Spinnbath | | | | | Spinning Machine | | Results | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Day | Date | $CS_2$ | Cell. | AV | Temp Viscose | H2SO4 | Na2SO4 | ZnSO4 | Temp Spinnbath | Drawing Speed | Stretch | Strength | Elongation |
| | | % | % | | ° C | g/l | g/l | g/l | ° C | m / min. | Maximum -% | cN/tex | % |
| 1 | 27.10.99 | Low | Low | Low | High | Low | Low | High | Low | Low | -30% | 33,5 | 13,8 |
| | | | | | | | | | | | -20% | 33,9 | 13,2 |
| | | Low | Low | Low | Low | High | Low | High | High | High | -10% | 16,5 | 6,2 |
| | | | | | | | | | | | -20% | 15,3 | 7,3 |
| | | Low | Low | Low | High | Low | High | Low | High | High | -10% | 37,4 | 12,7 |
| | | | | | | | | | | | -20% | 34,4 | 13,3 |
| | | Low | Low | Low | Low | High | High | Low | Low | Low | -30% | 16,8 | 7,3 |
| | | | | | | | | | | | -20% | 17,0 | 7,3 |
| 2 | 30.10.99 | High | High | High | High | High | Low | Low | Low | Low | -30% | 33,0 | 12,1 |
| | | | | | | | | | | | -20% | 33,2 | 11,6 |
| | | High | High | High | High | High | High | High | High | High | -10% | 32,4 | 12,2 |
| | | | | | | | | | | | -20% | 32,0 | 14,8 |
| | | High | High | High | Low | Low | Low | Low | High | High | -10% | 30,9 | 12,8 |
| | | | | | | | | | | | -20% | 24,5 | 16,3 |
| | | High | High | High | Low | Low | High | High | Low | Low | -30% | 29,1 | 11,3 |
| | | | | | | | | | | | -20% | 32,6 | 11,7 |
| 3 | 02.11.99 | Low | High | Low | Low | Low | High | High | High | Low | -10% | 34,7 | 14,8 |
| | | | | | | | | | | | -20% | 32,9 | 15,1 |
| | | Low | High | Low | Low | Low | Low | Low | Low | High | -30% | 27,0 | 11,0 |
| | | | | | | | | | | | -20% | 28,6 | 11,2 |
| | | Low | High | Low | High | High | Low | Low | High | Low | -10% | 30,0 | 12,0 |
| | | | | | | | | | | | -20% | 28,2 | 11,4 |
| | | Low | High | Low | High | High | High | High | Low | High | -30% | 26,7 | 12,4 |
| | | | | | | | | | | | -20% | 28,8 | 12,3 |
| 4 | 04.11.99 | Low | High | High | Low | High | High | Low | Low | High | -10% | 32,5 | 12,5 |
| | | | | | | | | | | | -20% | 28,1 | 14,5 |
| | | Low | High | High | Low | High | Low | High | High | Low | -30% | 33,3 | 13,4 |
| | | | | | | | | | | | -20% | 35,7 | 13,8 |
| | | Low | High | High | High | Low | High | Low | High | Low | -30% | 34,2 | 16,5 |
| | | | | | | | | | | | -20% | 37,9 | 14,8 |
| | | Low | High | High | High | Low | Low | High | Low | High | -10% | 36,6 | 13,8 |
| | | | | | | | | | | | -20% | 33,8 | 13,9 |
| 5 | 05.11.99 | Med | Med | Med | Med | Med | Med | Med | Med | Med | -15% | 40,0 | 12,9 |
| | | | | | | | | | | | -20% | 39,2 | 13,1 |
| | | | | | | | | | | | -25% | 37,1 | 13,1 |
| 6 | 08.11.99 | High | Low | Low | High | High | Low | Low | Low | High | -10% | 19,7 | 6,7 |
| | | | | | | | | | | | -20% | 17,4 | 6,6 |
| | | High | Low | Low | High | High | High | High | High | Low | -30% | 34,5 | 15,7 |
| | | | | | | | | | | | -20% | 36,0 | 15,0 |
| | | High | Low | Low | Low | Low | Low | Low | High | Low | -30% | 22,1 | 7,9 |
| | | | | | | | | | | | -20% | 20,4 | 6,8 |
| | | High | Low | Low | Low | Low | High | High | Low | High | -10% | 37,8 | 11,9 |
| | | | | | | | | | | | -20% | 36,7 | 11,8 |
| 7 | 09.11.99 | Low | Low | High | Low | Low | High | High | High | High | -30% | 37,9 | 15,5 |
| | | | | | | | | | | | -20% | 41,8 | 15,6 |
| | | Low | Low | High | High | High | Low | Low | High | High | -30% | 36,5 | 14,7 |
| | | | | | | | | | | | -20% | 40,4 | 14,6 |
| | | Low | Low | High | Low | Low | Low | Low | Low | Low | -10% | 36,0 | 12,7 |
| | | | | | | | | | | | -20% | 35,6 | 13,5 |
| | | Low | Low | High | High | High | High | High | Low | Low | -10% | 38,5 | 13,2 |
| | | | | | | | | | | | -20% | 36,9 | 13,3 |
| 8 | 10.11.99 | High | Low | High | High | Low | Low | High | High | Low | -10% | 45,1 | 14,1 |
| | | | | | | | | | | | -20% | 42,2 | 14,4 |
| | | High | Low | High | High | Low | High | Low | Low | High | -30% | 37,2 | 13,1 |
| | | | | | | | | | | | -20% | 38,3 | 12,7 |
| | | High | Low | High | Low | High | Low | High | Low | High | -30% | 35,3 | 12,4 |
| | | | | | | | | | | | -20% | 36,6 | 12,1 |
| | | High | Low | High | Low | High | High | Low | High | Low | -10% | 42,2 | 15,0 |
| | | | | | | | | | | | -20% | 40,3 | 15,4 |
| 9 | 11.11.99 | High | High | Low | High | Low | Low | High | High | High | -30% | 28,3 | 15,4 |
| | | | | | | | | | | | -20% | 32,9 | 16,5 |
| | | High | High | Low | High | Low | High | Low | Low | Low | -10% | 34,5 | 13,6 |
| | | | | | | | | | | | -20% | 34,8 | 13,0 |
| | | High | High | Low | Low | High | High | Low | High | High | -30% | 28,0 | 14,8 |
| | | | | | | | | | | | -20% | 29,2 | 14,0 |
| | | High | High | Low | Low | High | Low | High | Low | Low | -10% | 29,8 | 9,2 |
| | | | | | | | | | | | -20% | 29,5 | 9,3 |
| 10 | 12.11.99 | Med | Med | Med | Med | Med | Med | Med | Med | Med | -15% | 40,4 | 13,1 |
| | | | | | | | | | | | -20% | 40,7 | 12,9 |
| | | | | | | | | | | | -25% | 40,0 | 13,3 |

**Table 2:** Confounding of the $2^{10-5}$ fractional factorial design in Table 1 with 10 factors in 32 experiments. The defining relation (ignoring words of length 5 or higher) is:

| | | | | |
|---|---|---|---|---|
| I | = | $(CS_2)$(VisTemp)$(ZnSO_4)$(SpTemp) | = | (AV)(VisTemp)$(ZnSO_4)$(Stretch) |
| | = | (Cell)(VisTemp)$(ZnSO_4)$(Draw) | = | (AV)$(CS_2)$(SpTemp)(Stretch) |
| | = | (Cell)$(CS_2)$(SpTemp)(Draw) | = | (Cell)(AV)(Stretch)(Draw) |
| | = | $(CS_2)(H_2SO_4)(Na_2SO_4)$(SpTemp) | = | (AV)$(H_2SO_4)(Na_2SO_4)$((Stretch) |
| | = | (Cell)$(H_2SO_4)(Na_2SO_4)$(Draw) | = | (VisTemp)$(ZnSO_4)(H_2SO_4)(Na_2SO_4)$ |

Ignoring interactions of order 3 or higher leads to unconfounded main effects. Two-factor interactions are confounded with other two-factor interactions, according to:

| | |
|---|---|
| (Cell)(AV) $\approx$ (Stretch)(Draw) | (Cell)$(CS_2)$ $\approx$ (SpTemp)(Draw) |
| (Cell)(VisTemp) $\approx$ $(ZnSO_4)$(Draw) | (Cell)$(ZnSO_4)$ $\approx$ (VisTemp)(Draw) |
| (Cell)$(H_2SO_4)$ $\approx$ $(Na_2SO_4)$(Draw) | (Cell)$(Na_2SO_4)$ $\approx$ $(H_2SO_4)$(Draw) |
| (Cell)(SpTemp) $\approx$ $(CS_2)$(Draw) | (Cell)(Stretch)$\approx$ (AV)(Draw) |
| $(CS_2)$ $\approx$ (SpTemp)(Stretch) | (AV)(VisTemp) $\approx$ $(ZnSO_4)$(Stretch) |
| (AV)$(ZnSO_4)$ $\approx$ (VisTemp)(Stretch) | (AV)$(H_2SO_4)$ $\approx$ $(Na_2SO_4)$(Stretch) |
| (AV)$(Na_2SO_4)$ $\approx$ $(H_2SO_4)$(Stretch) | (AV)(SpTemp) $\approx$ $(CS_2)$(Stretch ) |
| $(CS_2)$(VisTemp) $\approx$$(ZnSO_4)$(SpTemp) | $(CS_2)(ZnSO_4)$ $\approx$ (VisTemp)(SpTemp) |
| $(CS_2)(H_2SO_4)$ $\approx$ $(Na_2SO_4)$(SpTemp) | $(CS_2)(Na_2SO_4)$ $\approx$ $(H_2SO_4)$(SpTemp) |
| (VisTemp)$(H_2SO_4)$ $\approx$ $(ZnSO_4)(Na_2SO_4)$ | (VisTemp)$(Na_2SO_4)$ $\approx$ $(ZnSO_4)(H_2SO_4)$ |

(Cell)(Draw) $\approx$ (AV)(Stretch) $\approx$ $(CS_2)$(SpTemp) $\approx$ (VisTemp)$(ZnSO_4)$ $\approx$ $(H_2SO_4)(Na_2SO_4)$

**Table 3:** Confounding Patterns of the $2^{10-5}$ fractional factorial design in Table 1 in 8 Blocks with Blocking Variables $B_1$ = Cell, $B_2$ = AV and $B_3 = CS_2$.

Main effects of Cell, AV, und $CS_2$, and the interactions (Cell)(AV) $\approx$ (Stretch)(Draw), (Cell)$(CS_2)$ $\approx$ (SpTemp)(Draw), (AV)$(CS_2)$ $\approx$ (SpTemp)(Stretch), and (Cell)(AV)$(CS_2)$ $\approx$ (VisTemp)$(H_2SO_4)$ $\approx$ $(ZnSO_4)(Na_2SO_4)$ cannot be separated from block effects. All other two-factor interactions are confounded as in Table 2.

Centerpoints: Two centerpoints on days 5 and 10 are added to the 32 runs of the fractional factorial experiment. The factor levels of these runs are set midway between the low and high levels of the factorial experiment. In addition to a medium reduction (20 percent) from the largest possible stretch, we also reduce stretch by 15 und 25 percent. For this reason, three different levels of stretch reduction are available at the center points.

Measurements: A sample of fibers is taken from each production run. Each fiber is fastened to a holder, and is stretched by applying force. Fiber strength = Force/Thickness is calculated as the ratio of force (in cN) that is needed to tear the fiber and thickness of the fiber (tex = g/1000m). Elongation is expressed by the proportional amount a fiber can be stretched.

The average of results from n = 20 fibers make up the observations in Table 1.

## 2.2 Data Analysis

### 2.2.1 Effect of Stretch on Fiber Strength and Fiber Elongation

Strength and elongation are measured at each of the 32 runs of the $2^{10-5}$ fractional factorial experiment under two stretch reductions: the reduction specified by the fractional factorial design (30 or 10 perecent) and the added 20 percent reduction. At the centerpoints measurements are made at three different reductions.

As these two (three) measurements are taken on the very same product, these realizations cannot be treated as independent replications. In the following analysis we treat each of the 34 factor combinations as a separate block, define indicator variables for the blocks, and regress fiber strength (elongation) on these 34 indicators. The residuals from this regression are plotted against reduction in stretch. Figure 1 shows that strength decreases if we increase the amount of reduction from the largest feasible stretch. Elongation increases with increased reduction of stretch. The influence is linear over the studied range.

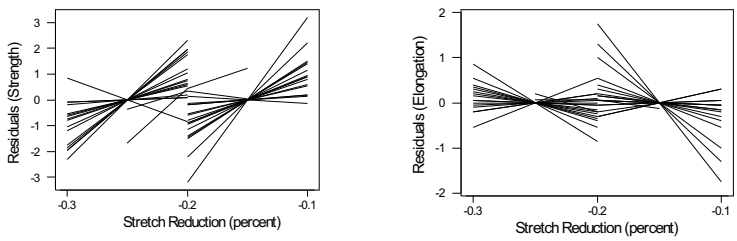



**Figure 1:** Residuals from the regression of fiber strength (fiber elongation) on 34 block indicators, plotted against different stretch reductions. Residuals from the same block are connected.

A regression of strength on stretch reduction

$$Strength_{Block,Stretch} = \mu_{Block} + \beta(Stretch) + Noise$$

leads to an estimate for $\beta$ of 1.92 cN/tex (standard error 0.28 cN/tex). Each additional ten percent reduction of stretch reduces strength by 1.92 cN/tex. A regression with fiber elongation leads to an estimate of $-0.44$ percent (standard error 0.15 percent); each additional ten percent reduction of stretch increases fiber elongation by 0.44 percent. We also consider a

model with a quadratic component; however the estimate of the quadratic effect is small and statistically insignificant.

### 2.2.2 Analysis of Strength and Elongation: Effects of Viscose Temperature and Spinnbath Factors

Main and interaction effects among Cell, AV, and $CS_2$ may be confounded with possible block effects. In order to be on the safe side, we define indicator variables for the ten days of our experiment in order to adjust our analysis for possible block effects. While losing the ability to estimate effects of Cell, AV, and $CS_2$, we obtain estimates of main and interaction effects of the other seven factors which are unconfounded with block effects.

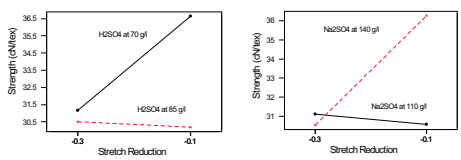
With the observations from the original $2^{10-5}$ fractional factorial experiment we estimate the regression model

$$\begin{aligned} Strength &= \mu_{Block} + \beta_1(VisTemp) + \beta_2(H_2SO_4) + \ldots + \beta_7(Draw) \\ &\quad +\beta_{12}(VisTemp)(H_2SO_4) + \beta_{13}(VisTemp)(Na_2SO_4) + \ldots \\ &\quad +Noise \end{aligned}$$

Not every two-factor interaction can be estimated as certain two-factor interactions are confounded with other two-factor interactions (see Table 2). The statistical significance of the estimated effects is determined through normal probability plots (see Hogg and Ledolter (1991, page 339)).

For strength the most important main effects are viscose temperature, $H_2SO_4$ and $Na_2SO_4$ in the spinnbath, and stretch reduction. Interactions between stretch reduction and $H_2SO_4$, and stretch reduction and $Na_2SO_4$ are found. For elongation the most important main effects are due to spinnbath temperature, $Na_2SO_4$ and $H_2SO_4$ in the spinnbath, and stretch reduction. Important interaction effects between stretch reduction and $Na_2SO_4$, and stretch reduction and $H_2SO_4$ are found. Important main and interaction effects for strength and elongation are displayed in Figure 2.

We also study strength and elongation at a fixed 20% stretch reduction. For strength we find significant main effects of viscose temperature, $H_2SO_4$ and $Na_2SO_4$. For elongation we find significant main effects of spinnbath temperature, $Na_2SO_4$, and $H_2SO_4$. These findings agree with the above results. Main effects of $H_2SO_4$ and $Na_2SO_4$ must now be interpreted as the effects at a



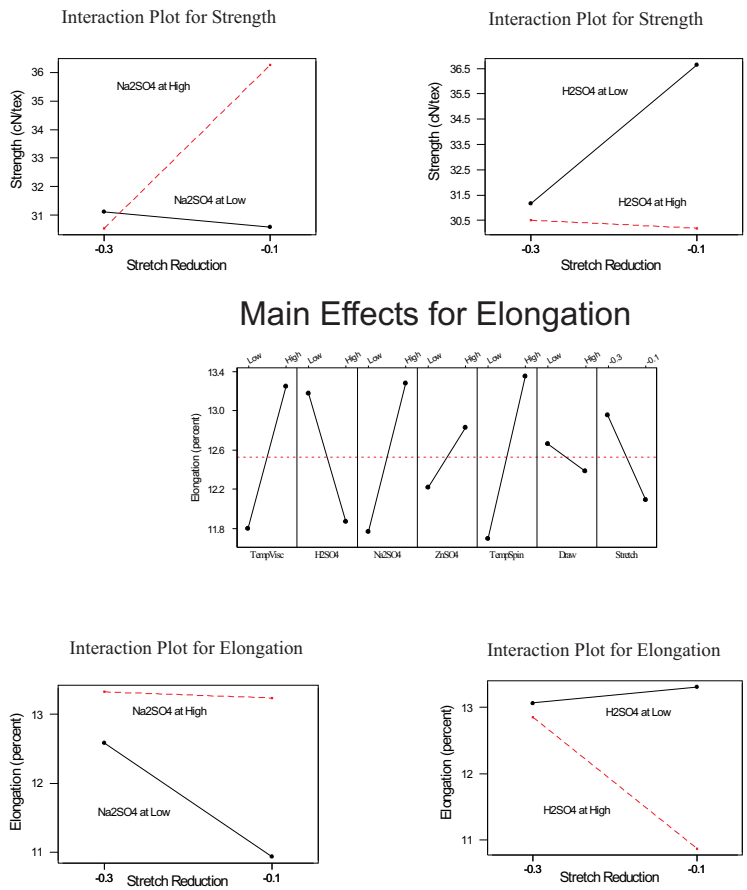



**Figure 2:** Analysis of viscose temperature and factors of the spinning process. Graphical representation of important main- and interaction effects.

### 2.2.3 Analysis of Strength and Elongation when Ignoring Possible Blockeffects

Ignoring the possibility of block effects allows us to estimate main effects of all ten factors including those of Cell, AV and $CS_2$, as well as (confounded) two-factor interactions. Stretch reduction turns out to be very important for both strength and elongation, confirming the results of Sections 2.2.1 and 2.2.2. Other important factors for strength are AV, the three chemicals of the spinnbath ($H_2SO_4$, $Na_2SO_4$, and $ZnSO_4$), and viscose temperature. Large interactions of AV with Cell, AV with $H_2SO_4$, and AV with $Na_2SO_4$ are also found. For elongation the effects appear smaller and distributed over more factors. In addition to the impact of stretch reduction, we find large main effects for AV, the chemicals in the spinnbath ($H_2SO_4$, $Na_2SO_4$), spinnbath temperature, and viscose temperature.

The standard deviation of the variation that is left unexplained by the model (mainly insignificant 2-factor interactions) is about 4.8 cN/tex for strength, and 1.8 percent for elongation. These are quite large compared to the standard deviations that are calculated from the replications at the centerpoints on days 5 and 10. Pooling the observations for the three levels of stretch reduction at the two centerpoints leads to a standard deviation of 1.34 cN/tex for strength, and 0.14 percent for elongation. These standard deviations appear to underestimate the natural variation as prior experiments with repeated measurements on fibers manufactured under identical process conditions have led to considerably higher variablity. Several other significant effects would have resulted if we had calculated the standard errors with the (smaller) centerpoint standard deviations. We are currently repeating the centerpoint experiments to obtain a more reliable estimate of the natural variability.

### 2.2.4 Additional Analyses

The scatter diagram of strength against elongation in Figure 3 shows connections between these two quality indicators. However, the correlation is not particularly strong, especially after ignoring a group of factor-level combinations with low strength and low elongation. Table 4 shows that low fiber strength and low elongation arises at runs with low levels for AV and Cell ($AV = 0.90, Cell = 5.3\%$), low viscose temperature ($15^oC$) and high levels of $H_2SO_4$ (85g/l). These combinations lead to unacceptable fiber properties and should be avoided.

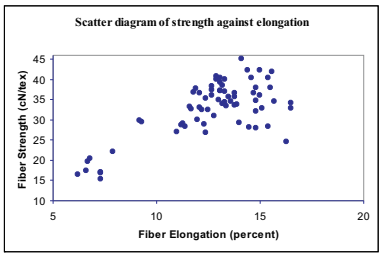



**Figure 3:** Scatter diagram of fiber strength against fiber elongation

**Table 4:** Factor combinations resulting in fibers with low strength and small elongation

| Day | $CS_2$ | Cell | AV | VisTemp | $H_2SO_4$ | $Na_2SO_4$ | $ZnSO_4$ | SpTemp | Draw |
|---|---|---|---|---|---|---|---|---|---|
| 1 | - | - | - | - | + | - | + | + | + |
| 1 | - | - | - | - | + | + | - | - | - |
| 6 | + | - | - | + | + | - | - | - | + |
| 6 | + | - | - | - | - | - | - | + | - |
| 9 | + | + | - | - | + | - | + | - | - |

### 2.3 Summary of Findings

Stretch has a large influence on fiber strength and fiber elongation. A ten percent reduction from the largest possible stretch reduces fiber strength by 1.92 cN/tex, and increases fiber elongation by 0.44 percent. Other influencing factors for strength are viscose temperature, $H_2SO_4$ and $Na_2SO_4$, where the influence of $H_2SO_4$ and $Na_2SO_4$ depends on the stretch reduction. Temperature of the spinnbath, $Na_2SO_4$ and $H_2SO_4$ are influencing factors for fiber elongation where, similar to strength, the influence of $Na_2SO_4$ and $H_2SO_4$ depends on stretch reduction. These findings are important for the company: (1) Changes in temperature (viscose as well as spinnbath temperatures) represent a fairly cheap way of improving fiber properties. (2) Interactions between stretch reductions and the chemicals in the spinnbath became visable through this experiment. These interactions will be explored in a next round of experiments.

## 3 Dependence of Fiber Strength on Fiber Thickness

Numerous studies have found a slight dependence of fiber strength on fiber thickness. Thickness of fiber is measured in "titers" (tex = grams/1000meters is proportional to the more commonly-used thickness measure denier = (0.9)titer). The slight dependence of fiber strength on thickness can be noticed in the figures on the left side of Figure 4 which display strength and thickness for samples from three fiber groups with different target

thicknesses. These graphs show that strength increases with decreasing fiber thickness.

Increased fiber strength is a desired quality characteristic. The above relationship has led to internal discussions whether one could use this dependence to gain a quality advantage. However, as the following discussion shows, this dependence is a mere artefact of the measurement process and cannot be used to gain a quality advantage.

Fiber strength is not measured directly; it is calculated by first measuring the force, F, that is necessary to tear the fiber and then relating this force to fiber thickness, titer T. That is, Strength = Force/Titer = F/T. A plausible model for force F and titer T is given by:

$$\begin{aligned} &(3.1a) \quad F &= a + b\mu_T + \varepsilon_F \\ &(3.1b) \quad T &= \mu_T + \varepsilon_T \end{aligned}$$

Equation (3.1b) describes the measurement error in determining thickness. Measured thickness is a combination of true thickness $\mu_T$ and a measurement error $\varepsilon_T$. Equation (3.1a) describes the model for force needed to tear a fiber with true thickness $\mu_T$. We assume that the expected force depends linearly on thickness, but do not suppose that it is necessarily proportional (a = 0). The measurement error for force is denoted by $\varepsilon_F$. The variances of the stochastic components $\varepsilon_K$ and $\varepsilon_T$ are denoted by $\sigma_{KK}$ and $\sigma_{TT}$. The model allows for covariance ($\sigma_{KT}$) between the measurement errors $\varepsilon_K$ and $\varepsilon_T$.

We want to explain the relationship that we see in the scatter diagrams of Figure 4 which relate strength to measured (but not necessarily true) thickness. For that we calculate the expected value of strength for a fiber with measured thickness t,

$$(3.2) \quad E[S = (F/T)|T = t] = (1/t)E[F|T = t].$$

Our model in equations (3.1), under the additional assumption of a bivariate normal distribution for $\varepsilon_K$ und $\varepsilon_T$, implies the conditional expectation

$$(3.3) \quad E[F|T = t] = a + b\mu_T + (\sigma_{KT}/\sigma_{TT})[t - \mu_T]$$

and

$$(3.4) \quad E[S = (F/T)|T = t] = (1/t)E[F|T = t] = \rho\lambda + (1/t)[a + (b - \rho\lambda)\mu_T]$$

where $\rho = Corr(\varepsilon_K, \varepsilon_T) = \sigma_{KT}/[\sigma_{KK}\sigma_{TT}]^{0.5}$ and $\lambda = [\sigma_{KK}/\sigma_{TT}]^{0.5}$.

For uncorrelated errors $\varepsilon_K$ and $\varepsilon_T$ (which is plausible as one can assume that measurement errors for force and thickness are unrelated), we obtain

$$(3.5) \quad E[S = (F/T)|T = t] = [a + b\mu_T]/t = [a/t] + b[\mu_T/t]$$

Equation (3.5) explains why there is a dependence of strength on thickness. First, it arises because of the possible violation of strict proportionality between force and thickness. A positive intercept in equation (3.1a) implies that the conditional expectation of strength increases with decreasing thickness. The fact that the intercept a tends to be positive is confirmed by the scatter diagrams of force against thickness on the right side of Figure 4.

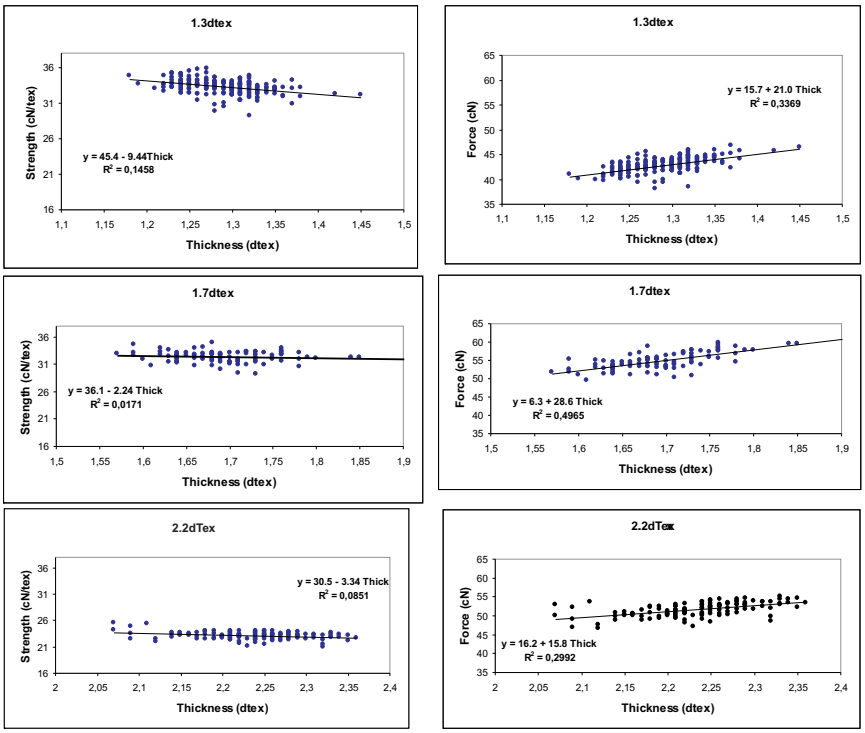



**Figure 4:** Scatter diagrams of strength against fiber thickness (left panel) and force against fiber thickness (right panel). Three fiber groups with different target thicknesses.

Second, the observed dependence arises because of the measurement error in determining thickness. The ratio $(\mu_T/t)$ in the second component of equation (3.5) relates measured and true thickness. The proportionality factor b is reduced if measured thickness exceeds the true value, and it is increased if measured thickness is smaller than the true value. In the general case with correlated measurement errors the dependence of strength

on thickness becomes larger for positive correlation $\rho = Corr(\varepsilon_K, \varepsilon_T)$, and smaller with negative correlation.

This derivation shows that the empirically observed relationship between force and thickness is due to a measurement problem. A plan to reduce the fiber thickness is not going to improve fiber strength. Improvements in strength must come from changes in the manufacturing process.

## 4 Concluding Remarks

Many other analyses and experiments were conducted over the last two years, and there are plans for many more. (1) A detailed statistical analysis was conducted with historic process data. Relationships between several quality indicators (such as fiber strength, elongation, coloring properties, an indicator thought responsible for downstream spinning problems) and more than sixty process variables were explored. The investigation involved more than one million records. It identified the factors which were varied as part of the experiment discussed in Section 2. (2) A study currently underway looks at links between the quality of wood (age, color), the quantity of chemicals needed in the bleaching process to attain desirable white fibers, and fiber characteristics. Increased bleeching is thought to affect fiber properties. (3) Improved sampling methods for measuring fiber characteristics are being investigated. (4) The results of the designed experiment in Section 2 need to be confirmed with additional experiments. Furthermore, it must be investigated that "off-line" results on the pilot plant (which involve the production of small quantities under highly-controlled conditions) also apply "on-line" when several tons of material are produced each day.

## References

Box,G.E.P., Draper, N.R. (1969). *Evolutionary Operation.* New York: Wiley.

Box,G.E.P., Hunter, W.G., and Hunter, S. (1978). *Statistics for Experimenters.* New York: Wiley.

Burill, C.W., and Ledolter, J. (1999). *Achieving Quality through Continual Improvement.* New York: Wiley.

Hogg, R.V., and Ledolter, J. (1991). *Statistics for Engineers and Physical Scientists (second edition).* New York: Prentice Hall.

Mason, R.L., Gunst R.F., and Hess, J.L. (1989). *Statistical Design and Analysis of Experiments.* New York: Wiley.

Montgomery, D. (1991). *Design and Anaysis of Experiments.* New York: Wiley.

Ledolter, J., and Burril, C.W. (1999). *Statistical Quality Control: Strategies and Tools for Continual Improvement.* New York: Wiley.

# Iterative and recursive estimation in structural non-adaptive models

V. Patilea [1], E. Renault [2]

[1] LEO, Université d'Orléans.
[2] CREST-INSEE and Université Paris IX.

## 1 Introduction

Structural econometric models often lead to an explicit formulation for the conditional distribution of the endogenous variables given the exogenous variables and the lagged endogenous values. However, such explicit and apppealing formulations may imply a partition of the set of the endogenous variables in two subsets $\{Y_t\}$ and $\{Y_t^*\}$, that is the observed and the latent endogenous variables, respectively. The variables $\{Y_t^*\}$ are unobserved by the econometrician but appear in the structural relationships, for instance because they are observed by the economic agents. Due to the presence of these latent variables, it may be very difficult to efficiently estimate the parameters of interest because of the intractability of the likelihood function corresponding to the observable variables.
Consider a general setting for nonlinear state space models:

$$Y_t^* = \varphi\left(u_t, X_t; \theta\right), \tag{1}$$

$$Y_t = r\left(u_t, X_t; \theta\right), \quad \theta \in \Theta \subset \mathbf{R}^p. \tag{2}$$

The first equation (transition equation) characterizes the dynamics of the latent variables $Y_t^*$ through a known function $\varphi\left(\cdot\right)$ depending on the unknown parameters $\theta$, the current value $X_t$ of the exogenous variables and a latent stochastic process $u_t$. Typically, $u_t$ may gather lagged values of $Y_t^*$ as well as the current value of a white noise process with a distribution independent of $\theta$, the parameters of interest. The second equation (measurement equation) relates the observed variables $Y_t$ to the latent and the exogenous ones and the unknown parameters through a known function function $r(\cdot)$. In other words, we are faced with an incomplete data problem.
A popular and useful iterative procedure for finding an efficient estimator in incomplete data models is the EM algorithm (Dempster *et al.* (1977)).

However, the standard asymptotic properties of the EM procedures seems no longer guaranteed when the support of the conditional probability distribution of the latent variables given the observable ones depends on the unknown parameters. More recently, Gouriéroux *et al.* (1993) proposed a simulation based procedure well-suited for estimating nonlinear state space models like (1) - (2). Unfortunately, this so-called indirect inference method, an extension of the simulated method of moments (Duffie and Singleton (1993)), does not provide a guideline for a clever choice of the moments to match which is needed for reaching efficiency[1]. The basic idea of this paper can be seen as an extension of the EM methodology. Each iteration of the algorithms we propose consists of two simple steps. Like in the M-step of a EM algorithm, we obtain our estimating sequence based on a so-called latent criterion (extremum estimator or moment estimator) built from the latent model (1). The motivation for this is the belief that the latent model would provide, if complete data were available, the best estimator both in terms of precision and computational costs. Clearly, in our general setting where the relationship between latent and observable variables may depend upon the unknown parameters, it is not automatically true that the latent variables convey more information than the observed ones (see section 2 for a more detailed discussion on the information bounds in the latent and observable world, respectively). Nevertheless, for the sake of both economic interpretation and computational complexity, we will develop a methodology always based on the preference for the latent model. Of course, this M-step has to be coupled with something like a E-step where the occurence of the latent variables in the latent criterion of interest is replaced by a 'guess' on these variables (or at least a 'guess' of the latent criterion itself) associated to the observation of $(X, Y)$ and the 'guess' on the value of the unknown parameters. Like in standard EM procedure, this later guess is the reason why the procedure has to be iterative.
It is often the case that the conditional expectation (or, more generally, the guess of the latent criterion) required in the E-step involves computer intensive calculations and/or simulations, like in the simulated EM algorithms (*e.g.*, Ruud (1991)). This is the reason why we also propose some recursive procedures. Actually, Young (1985) distinguishes two types of data processing for inference purposes in complex dynamic models. On one hand, iterative data processing used to maximize a given criterion as, for example, in "iterative least squares or some other relaxation procedures". Such kind of procedures involve, at each stage, the batch processing of the full block of data, but the highly nonlinear computation of the required estimator is decomposed in a sequence of simpler computations of intermediate estimators. On the other hand, the recursive approach which "introduces an extra dimension to estimation : in addition to the en bloc estimates based on complete data set (...), the analyst is also able to obtain estimates of the parameters for (...) subsets of the data, in a computationally elegant and efficient manner". This approach "also opens up the way for the estimation

of (...) states in stochastic-dynamic systems".
In this paper we show how the two aspects, iterative/recursive, are both useful for extremum estimation in a large class of nonlinear dynamic systems. The basic framework we have in mind is a case where a natural sample based criterion, say, $Q_T[\theta, \lambda]$ for extremum (argmax) estimation of a structural parameter $\theta$ is 'contaminated' by other occurrences of $\theta$ in a 'nuisance function' $\lambda = \lambda(\theta)$. From the consistency and/or simplicity point of view, maximizing $Q_T[\theta, \lambda(\theta)]$ with respect to $\theta$ does not produce a satisfactory procedure. However, maximizing the criterion $Q_T\left[\theta, \lambda(\theta^0)\right]$ with respect to $\theta$ should provide a convenient (consistent, easy to compute, precise,...) estimator of $\theta^0$, the true unknown value of the parameter, provided that we are able to compute this criterion. The problem is, of course, that we are not able to compute it. Below we will exhibit a class of structural econometric models where the inference issue reveals, in a natural way, a pattern as described above. In the cases we have in mind, while standard efficient inference on $\theta$ is infeasible, or at least unpalatable, some structural features of the model or some properties of the estimation method suggest a simple, but uncomputable, criterion $Q_T\left[\cdot, \lambda(\theta^0)\right]$ which, roughly speaking, would allow inference without a significant loss of efficiency. Typically, $Q_T\left[\cdot, \lambda(\theta^0)\right]$ corresponds to a latent criterion which depends upon $\lambda(\theta^0)$ due to the fact that the latent variables $Y_t^*$ have to be recovered from the observations thanks to a structural relationship which involves $\theta^0$.
This motivates us to look for continuously updated proxies of this criterion, either by 'nearly efficient' iterative procedures, or by less computationally demanding recursive approaches. Most of the recursive estimators we propose are defined through the first order conditions associated to suitable proxies of the simple criterion in hand, rather than as maximizers of these proxy criteria. Moreover, we argue that in many econometric models leading to a sample based criterion $Q_T[\theta, \lambda(\theta)]$ there exists a duality between the 'nuisance function' $\lambda(\theta)$ and unobserved state variables.
The paper is organized as follows. In section 2 we describe the general framework and analyze the identification issue. In section 3 we define a class of iterative estimators by extending a procedure due to Renault and Touzi (1996). In section 4 we argue that in the framework considered in this paper, the recursive approach may be much more user friendly than the iterative one. This might be particularly the case in the presence of an unobserved state variable process that enters the structural model nonlinearly. We propose several recursive (mainly of Robbins-Monro type) methods for which we derive the asymptotic properties. In section 5 we apply the general procedures developed in the previous sections to a class of structural econometric models stemming from the nonlinear rational expectation literature. More precisely, we focus on a class of option pricing models involving unobservable state variables (see also Renault (1997)).

## 2 The general framework and the identification problem

Consider that we observe a sequence of random vectors $(Y_t)_{t\geq 1}$ defined on some abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and taking values in a subset of $\mathbf{R}^J$, $J \geq 1$. Say that the structural econometric model we consider, or the estimation method we consider, impose a sample based criterion (objective function)

$$Q_T\left[\,\theta,\, \lambda(\theta)\right] = Q_T\left[\,\theta,\, \lambda(\theta), (Y_t)_{1\leq t\leq T}\right]$$

for the estimation of $\theta$. The parameter of interest $\theta$ lies in $\Theta$, a subset of $\mathbf{R}^p$. The function $\lambda(\cdot)$ is known and it is defined on $\Theta$ and it takes values in $\Gamma$, a subset of $\mathbf{R}^q$. Typically, $\lambda(\Theta) = \Gamma$. Let us remark that for convergence issues the sets $\Theta$ and $\Gamma$ could be quite general (see Proposition 3.1 below). However, most of the time we consider them subsets of Euclidean spaces.
As announced in the introduction, we have in mind the case where we cannot, or we should not, maximize $Q_T$ with respect to all the occurrences of the parameter of interest. However, $Q_T\left[\,\theta,\, \lambda(\theta^0)\right]$ represents a simple, but impossible to compute, criterion for estimating $\theta$. The natural idea is to use a two stage estimation strategy, that is first to replace the unknown value $\lambda(\theta^0)$ by some 'good proxy' and afterwards to maximize the simple criterion obtained in this way. The 'good proxies' of $\lambda(\theta^0)$ will be obtained from the previous steps of the estimation procedure. In other words, once we have, say, $\theta^1$ an estimate of $\theta^0$, $\lambda(\theta^1)$ will be a proxy for $\lambda(\theta^0)$. Maximizing $Q_T\left[\,\theta,\, \lambda(\theta^1)\right]$ with respect to $\theta$ we get a new (updated) estimate of $\theta^0$ which can be used in the next step. The steps could be considered for a fixed sample (iterative, or cross-sectional procedures), or, each time that a new data arrive, a new step is performed (recursive procedures). Moreover, the maximizers of the proxy criteria could be also defined through the first order conditions.
These facts lead us to the definition of a two parameter criterion

$$Q_T\left[\,\theta,\, \lambda\right] = Q_T\left[\,\theta,\, \lambda, (Y_t)_{1\leq t\leq T}\right], \quad (\theta, \lambda) \in \Theta \times \Gamma,$$

deduced from the original one. We consider a first set of assumptions similar to those usually imposed for $M-$estimation in the presence of nuisance parameters (see, e.g., section 4 of Wooldridge (1994)).

**Assumption 2.1** *i) For any $T \geq 1$, $Q_T(\cdot\,,\,\cdot)$ satisfies the standard measurability and continuity conditions, i.e., it is measurable as function of observations and it is continuous as a function of parameters $(\theta, \lambda)$. ii) There exists a limit criterion $Q_\infty(\cdot\,,\,\cdot) : \Theta \times \Gamma \to \mathbf{R}$ such that*

$$\lim_{T\to\infty} E\left[Q_T(\theta,\lambda)\right] = Q_\infty(\theta,\lambda), \quad \forall(\theta,\lambda)\in\Theta\times\Gamma,$$

*where the expectation is computed with respect to the true distribution of the observations which depends on* $\theta^0$.

A problem to be addressed is, of course, the identifiability of $\theta^0$. A usual identification condition for argmax estimation based on the full criterion $Q_T\left[\theta,\lambda(\theta)\right]$ would state that $\theta^0$ is the unique maximizer of $Q_\infty\left[\theta,\lambda(\theta)\right]$. However, the starting point for the estimation procedures we consider herein is the simpler criterion $Q_T\left[\theta,\lambda(\theta^0)\right]$. Therefore, a basic identification condition to be imposed in our framework is that $\theta^0$ is the unique maximizer of the simplified limit criterion[2] $Q_\infty\left[\theta,\lambda(\theta^0)\right]$. This condition is quite similar with the usual identification condition that would have been imposed if $\lambda=\lambda(\theta)$ was a nuisance parameter (see, *e.g.* Wooldridge (1994), Theorem 4.3). The basic identification condition should be strengthened in order to take into account our two stage estimation strategy.
In order to get an idea about how we should reinforce the basic identifiability condition, assume that, for $\lambda\in\Gamma$, the limit criterion $Q_\infty(\theta,\lambda)$, viewed as a function of $\theta$, admits a unique maximizer denoted $\overline{\theta}(\theta^0,\lambda)$, that is

$$\overline{\theta}(\theta^0,\lambda) = \arg\max_{\theta\in\Theta} Q_\infty(\theta,\lambda).$$

Since the limit criterion depends on $\theta^0$, this unique maximizer also depends on $\theta^0$. Note that the basic identification condition can be written under the form

$$\overline{\theta}(\theta^0,\lambda(\theta^0)) = \theta^0. \tag{3}$$

Intuitively, given the estimation approach we follow, the strengthened identification condition should impose that the only fixed point of the function $\overline{\theta}(\theta^0,\lambda(\cdot))$ is $\theta^0$. Otherwise, it may happen that the statistician will not be able to reject a 'bad guess' $\theta^1\neq\theta^0$ used to build the simplified criterion $Q_T\left[\theta,\lambda(\theta^1)\right]$. These remark lead us to the following identification assumption.

**Assumption 2.2** *i) For any* $\theta^1\in\Theta$, *the function* $\theta\to Q_\infty\left[\theta,\lambda(\theta^1)\right]$ *admits a unique maximizer denoted by* $\overline{\theta}(\theta^0,\lambda(\theta^1))$., *ii) If* $\theta^0$ *is the value of the parameter corresponding to the distribution governing the observations, then* $\theta^0$ *is the unique fixed point of the map* $\overline{\theta}(\theta^0,\lambda(\cdot))$.

A justification of the fact that the unique fixed point condition make the parameter $\theta$ identifiable from the observations goes as follows. Let $\theta^1$ and $\theta^0$

be two different parameters corresponding to the same distribution of the observations. Maximizing the corresponding limit criteria (see Assumption 2.1 *ii)*) we obtain the same $\overline{\theta}$ functions, that is

$$\overline{\theta}(\theta^1, \lambda(\cdot)) = \overline{\theta}(\theta^0, \lambda(\cdot)).$$

Finally, the unique fixed point condition considered for $\overline{\theta}(\theta^1, \lambda(\cdot))$ and $\overline{\theta}(\theta^0, \lambda(\cdot))$ implies $\theta^1 = \theta^0$.
Imagine now that we want to estimate the parameter of interest using the first order conditions associated to the given maximization problem instead of maximizing the criterion in hand. Assuming the necessary regularity conditions define

$$M(\theta, \theta^1) = \frac{\partial Q_\infty}{\partial \theta}\left[\theta, \lambda(\theta^1)\right] \tag{4}$$

and assume that, for any $\theta^1 \in \Theta$, $\overline{\theta}(\theta^0, \lambda(\theta^1))$ is the unique solution of the equation $M(\theta, \theta^1) = 0$. Then, the unique fixed point property of $\overline{\theta}(\theta^0, \lambda(\cdot))$ is equivalent with the condition that $M(\theta^1, \theta^1) = 0$ implies $\theta^1 = \theta^0$. In other words, Assumption 2.2 could be replaced in this case by the following one.

**Assumption 2.3** *Let $\theta^0$ be the true unknown value of the parameter and assume that, for any $\theta^1 \in \Theta$, $\overline{\theta}(\theta^0, \lambda(\theta^1))$ defined in Assumption 2.2 i) is the unique solution of the equation $M(\theta, \theta^1) = 0$. Moreover,*

$$M(\theta^1, \theta^1) = 0 \Longrightarrow \theta^1 = \theta^0.$$

Assumption 2.2 will be used for iterative estimation procedures, while Assumption 2.3 will be invoked for recursive (Robbins-Monro type) estimation procedures.
In order to relate these procedures to the standard issue of estimation in the presence of nuisance parameter, let us deduce from the definition of $\overline{\theta}(\theta^0, \lambda(\cdot))$ the following identity :

$$M(\overline{\theta}(\theta^0, \lambda(\theta^1)), \theta^1) = 0, \quad \theta^1 \in \Theta.$$

If second derivatives exist, we may deduce that, for any $\theta^1$,

$$\frac{\partial M}{\partial \theta'}(\overline{\theta}(\theta^0, \lambda(\theta^1)), \theta^1)\frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \lambda(\theta^1)) + \frac{\partial M}{\partial \theta^{1\prime}}(\overline{\theta}(\theta^0, \lambda(\theta^1)), \theta^1) = 0,$$

where

$$\frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \lambda(\theta^1)) = \frac{\partial \overline{\theta}}{\partial \lambda'}(\theta^0, \lambda(\theta^1))\frac{\partial \lambda}{\partial \theta^{1\prime}}(\theta^1).$$

In particular, for $\theta^1 = \theta^0$, we get

$$\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta^0, \lambda\left(\theta^0\right)\right]\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left(\theta^0, \lambda\left(\theta^0\right)\right) + \frac{\partial^2 Q_\infty}{\partial\theta\partial\lambda'}\left[\theta^0, \lambda\left(\theta^0\right)\right]\frac{\partial\lambda}{\partial\theta^{1\prime}}\left(\theta^0\right) = 0. \tag{5}$$

It is well known (see *e.g.* Wooldridge (1994)) that the case where the cross-derivatives matrix $\partial^2 Q_\infty/\partial\theta\partial\lambda'\left[\theta^0, \lambda\left(\theta^0\right)\right]$ vanishes is precisely the case where an extremum estimator

$$\widetilde{\theta}_T = \arg\max_\theta Q_T\left(\theta, \widetilde{\lambda}_T\right)$$

has an asymptotic distribution which does not depend upon the choice of a consistent estimator $\widetilde{\lambda}_T$ of $\lambda^0$, the true unknown value of the nuisance parameter. As announced above, our focus of interest is the more difficult situation where we do not have at our disposal a consistent estimator of $\lambda^0$, but we must recover it through an iterative or recursive procedure providing a sequence of 'guesses' of $\lambda^0$. However, one may hope that in the framework we consider herein, the nullity of the cross-derivatives will be precisely the property that ensures that our procedures are as accurate as the infeasible extremum estimation method of $\theta$ based on the incomputable criterion $Q_T\left[\theta,\ \lambda\left(\theta^0\right)\right]$. Indeed, this is the case because, first, we will see below that there is no loss of accuracy if and only if $\partial\overline{\theta}/\partial\theta^{1\prime}\left(\theta^0, \lambda\left(\theta^0\right)\right) = 0$. On the other hand, from (5) we deduce

$$\frac{\partial^2 Q_\infty}{\partial\theta\partial\lambda'}\left[\theta^0, \lambda\left(\theta^0\right)\right] = 0 \quad\Longrightarrow\quad \frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}\left(\theta^0, \lambda\left(\theta^0\right)\right) = 0,$$

provided that the matrix $\partial Q_\infty/\partial\theta\partial\theta'\left(\theta^0, \lambda\left(\theta^0\right)\right)$ is nonsingular, which represents a standard assumption.

## 3 Iterative extremum estimation

### 3.1 The general principle and consistency

In order to develop the estimation strategy described in the previous section, let us define [3], for any $\lambda \in \Gamma$,

$$\overline{\theta}_T(\lambda) = \arg\max_{\theta\in\Theta} Q_T(\theta, \lambda). \tag{6}$$

Note that $\overline{\theta}_T(\lambda(\theta^0))$ is nothing else than the extremum (argmax) estimator defined through the simple criterion $Q_T(\theta, \lambda(\theta^0))$. Moreover, remark that $\overline{\theta}_T(\lambda)$ is the sample based counterpart of $\overline{\theta}(\theta^0, \lambda)$ defined in section 2. We need the following uniform convergence assumption on $Q_T(\theta, \lambda)$ in order to obtain the uniform convergence of $\overline{\theta}_T(\lambda)$ towards $\overline{\theta}(\theta^0, \lambda)$ (see also the

usual uniform assumptions imposed for $M-$estimation in the presence of nuisance parameters; e.g., Wooldridge (1994)).

**Assumption 3.1** *Assume that $Q_T(\cdot,\cdot)$, $T \geq 1$, satisfies the measurability condition of Assumption 2.1 i). Moreover,*

$$\sup_{(\theta,\lambda)\in\Theta\times\Gamma} |Q_T(\theta,\lambda) - Q_\infty(\theta,\lambda)| \xrightarrow{p} 0.$$

*with $Q_\infty$ defined as in Assumption 2.1 ii).*

The convergence results of this section rely on the following proposition which we state for a quite general parameter space.

**Proposition 3.1** *Assume that $\Theta$ and $\Gamma$ are compact subsets of some metric spaces $(\tilde{\Theta}, d_1)$ and $(\tilde{\Gamma}, d_2)$, respectively. Moreover, the map $\overline{\theta}(\theta^0,\cdot) : \Gamma \to \Theta$ is continuous. If Assumption 2.1 and Assumption 3.1 hold and $\overline{\theta}(\theta^0,\lambda)$ is the unique maximizer of $\theta \to Q_\infty(\theta,\lambda)$, then*

$$\sup_{\lambda\in\Gamma} d_1\left(\overline{\theta}_T(\lambda), \overline{\theta}(\theta^0,\lambda)\right) \xrightarrow{p} 0.$$

Following the estimation strategy announced above, let us now consider the following extremum estimation procedure, an extension of those presented by Renault and Touzi (1996), Patilea *et al.* (1995) and Renault (1996) : consider an arbitrary sequence $p(T)$, $T \geq 1$, of positive integers such that $\lim_{T\to\infty} p(T) = \infty$ and define

$$\theta_T = \theta_T^{(p(T)+1)}, \;\; T \geq 1, \tag{7}$$

where, for any $p \geq 1$,

$$\theta_T^{(p+1)} = \overline{\theta}_T\left(\lambda(\theta_T^{(p)})\right) \tag{8}$$

and $\theta_T^{(1)} \in \Theta$ is some starting value. This estimator is an extension and a practical version of the Renault and Touzi's estimator defined for a log-likelihood type criterion and $p(T) \equiv \infty$. Patilea *et al.* (1995) and Renault (1996) impose a contracting assumption on the $\overline{\theta}$ function, a reinforcement of the unique fixed point condition states in Assumption 2.2 *ii)*. We restate this assumption in our framework.

**Assumption 3.2** *Let $(\tilde{\Theta}, d_1)$ be a metric spaces containing the set $\Theta$. If $\theta^0$ is the true unknown value of the parameters, the mapping*

$$\overline{\theta}(\theta^0, \lambda(\cdot)) : \Theta \longrightarrow \Theta$$

*is strongly contracting, that is there exists* $k = k(\theta^0) \in (0,1)$ *such that,* $\forall\, \theta_1,\, \theta_2 \in \Theta$,

$$d_1\left(\overline{\theta}(\theta^0, \lambda(\theta^1))\,,\, \overline{\theta}(\theta^0, \lambda(\theta^2))\right) \leq k\, d_1\left(\theta^1\,,\, \theta^2\right).$$

We are able now to state the weak consistency result for the estimator defined in (7). Again, $\Theta$ and $\Gamma$ could be quite general parameter sets.

**Proposition 3.2** *If* $\Theta$ *is as in Proposition 3.1 and the Assumptions 2.1, 2.2 i), 3.1* [4] *and 3.2 hold, then* $\theta_T$ *defined in (7) is weakly consistent.*

### 3.2 Asymptotic distribution

Below in this section we deduce the limit distribution of the estimator defined in (7). Before stating our results we need some hypotheses (see, e.g., Wooldridge (1994), section 4). The sets $\Theta$ and $\Gamma$ are assumed to be subsets of Euclidean spaces, in particular $\Theta \subset \mathbf{R}^p$, for some $p \geq 1$.

**Assumption 3.3** *If* $\theta^0$ *is the true unknown value of the parameters, then*

$$\sqrt{T}\ \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0} \xrightarrow{d} N_p\left(0,\, B(\theta^0)\right)$$

*with*

$$B(\theta^0) = \lim_{T\to\infty} \mathrm{Var}\left(\sqrt{T}\ \frac{\partial Q_T}{\partial \theta}\left[\theta, \lambda(\theta^0)\right]\Big|_{\theta=\theta^0}\right)$$

*which is supposed to be positive definite.*

The asymptotic normality of the score of the simplified criterion $Q_T[\cdot, \lambda(\theta^0)]$ is as usual assumptions for extremum estimators. Another usual assumption is the uniform convergence, in probability, of the Hessian matrix of the criterion to be maximized. In our framework this corresponds to the assumption below. Let $\|\cdot\|_\lambda$ denote the spectral for the $p-$dimensional squared matrices, i.e., $\|A\|_\lambda$ denotes the square root of the largest eigenvalue of $A'A$.

**Assumption 3.4** *The matrix*

$$\Sigma(\theta,\theta^1) = -\frac{\partial^2 Q_\infty}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right],$$

*exists for any* $\theta,\theta^1 \in \Theta$ *and its components are continuous functions of* $(\theta,\theta^1)$. $\Sigma(\theta^0,\theta^0)$ *is assumed positive definite. Moreover, when* $T\to\infty$,

$$\sup_{\theta,\,\theta^1\in\Theta}\left\|\frac{\partial^2 Q_T}{\partial\theta\partial\theta'}\left[\theta,\lambda(\theta^1)\right]-\Sigma(\theta,\theta^1)\right\|_\lambda \longrightarrow 0, \tag{9}$$

*in probability.*

Note that $-\Sigma(\theta^0,\theta^0)$ is nothing else that the Hessian matrix, considered for $\theta=\theta^0$, of the simple limit criterion $Q_\infty[\cdot,\lambda(\theta^0)]$. This matrix is usually assumed to be negative definite. The next assumption is more specific to our framework.

**Assumption 3.5** *For any* $\theta^1\in\Theta$, *define*

$$H(\theta^1) = \frac{\partial^2 Q_\infty}{\partial\theta\partial\lambda'}\left[\theta,\lambda(\theta^1)\right]\Big|_{\theta=\theta^1}\frac{\partial\lambda}{\partial\theta^{1\prime}}(\theta^1),$$

$$H_T(\theta^1) = \frac{\partial^2 Q_T}{\partial\theta\partial\lambda'}\left[\theta,\lambda(\theta^1)\right]\Big|_{\theta=\theta^1}\frac{\partial\lambda}{\partial\theta^{1\prime}}(\theta^1),$$

*and assume that* $H(\cdot)$ *is continuous. Moreover,*

$$\sup_{\theta^1\in\Theta}\left\|H_T(\theta^1)-H(\theta^1)\right\|_\lambda \longrightarrow 0,$$

*in probability.*

A classic Taylor expansion argument that we will use for proving the asymptotic normality of our iterative estimator demands the matrix $H(\theta^0)+\Sigma(\theta^0,\theta^0)$ to be invertible. From (5) we may deduce

$$H(\theta^0) = -\Sigma(\theta^0,\theta^0)\,\frac{\partial\overline{\theta}}{\partial\theta^{1\prime}}(\theta^0,\lambda(\theta^0))\Big|_{\theta^1=\theta^0}.$$

Since $\|I_p-A\|_\lambda<1$ implies $A$ invertible ($I_p$ stands for the identity matrix), the following assumption ensures the desired property for $H(\theta^0)+\Sigma(\theta^0,\theta^0)$.

**Assumption 3.6** *Let $\theta^0$ be the true unknown value of the parameters and assume that the function $\overline{\theta}(\theta^0, \lambda(\cdot))$ defined in Assumption 2.2 i) is continuously differentiable and*

$$\left\| \frac{\partial \overline{\theta}}{\partial \theta^{1\,\prime}}(\theta^0, \lambda(\theta^1)) \Big|_{\theta^1=\theta^0} \right\|_\lambda < 1.$$

Note that Assumption 3.6 implies Assumption 3.2 and thus it ensures the local identifiability of $\theta^0$ from the observed variables. Moreover, we will see in section 4 that Assumption 3.6 is a natural condition for ensuring convergence of the Robbins-Monro type recursive estimators we propose. Note that

$$\frac{\partial \overline{\theta}}{\partial \theta^{1\,\prime}}(\theta^0, \lambda(\theta^0))\Big|_{\theta^1=\theta^0} = -\Sigma(\theta^0, \theta^0)^{-1} \frac{\partial^2 Q_\infty}{\partial \theta \partial \lambda'} \left[\theta^0, \lambda(\theta^1)\right]\Big|_{\theta^1=\theta^0} \frac{\partial \lambda}{\partial \theta^{1\,\prime}}(\theta^0)$$

which shows that, in some sense, we extend the standard nuisance parameter framework where, basically, $\partial\overline{\theta}/\partial\theta^{1'}$ is assumed to be zero. Renault and Touzi (1996) checked Assumption 3.6, by simulations, in the case of a stochastic volatility option pricing model.

**Proposition 3.3** *Assume that $\theta^0$, the true unknown value of the parameters, is an interior point of $\Theta \subset \mathbf{R}^p$. Consider that Assumptions 2.1, 2.2 i), 3.1, 3.2 and 3.3 to 3.6 hold and that $Q_\infty[\cdot, \lambda(\cdot)]$ is continuous on $\Theta \times \Theta$. If, in addition, the sequence $p(T)$, $T \geq 1$, considered in (7) is such that*

$$\sqrt{T}\left(\theta_T - \theta_T^{(p(T))}\right) = \sqrt{T}\left(\theta_T^{(p(T)+1)} - \theta_T^{p(T)}\right) \longrightarrow 0, \tag{10}$$

*in probability, then the $\theta_T$ is asymptotically normal with asymptotic variance matrix*

$$V(\theta^0) = A(\theta^0)^{-1}\Sigma(\theta^0, \theta^0)^{-1}B(\theta^0)\Sigma(\theta^0, \theta^0)^{-1}A(\theta^0)^{\prime\,:-1}.$$

*where*

$$A(\theta^0) = I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\,\prime}}(\theta^0, \theta^1)\Big|_{\theta^1=\theta^0}.$$

Proposition 3.3 extends to a general extremum estimators framework the Theorem 5.2 of Renault and Touzi (1996) on maximum likelihood type

estimators. Indeed, the asymptotic variance $V(\theta^0)$ is closely related to the standard asymptotic variance

$$W(\theta^0) = \Sigma(\theta^0, \theta^0)^{-1} B(\theta^0) \Sigma(\theta^0, \theta^0)^{-1},$$

of the hypothetical extremum estimator associated to the criterion $Q_T(\cdot, \lambda(\theta^0))$. This remark raises a natural question : is $V(\theta^0)$ greater or smaller (as a positive definite matrx) than $W(\theta^0)$? Using an example, we show in section 5 that there is no generally valid order relationship between those two matrices. Formally, we can easily imagine this if we remark that, for instance, for an unidimensional parameter the derivative of $\overline{\theta}(\theta^0, \lambda(\cdot))$ could be positive or negative. Intuitively, the full criterion $Q_T[\cdot, \lambda(\cdot)]$ could be 'more, or less informative' than the simplified criterion $Q_T[\cdot, \lambda(\theta^0)]$ and this may still be the case even when we adopt a two stage estimation strategy as in this paper.

As far as we are concerned by a comparison with the Renault and Touzi (1996) result on the asymptotic distribution of their estimator, two generalizations have to be stressed. First, since they considered a maximum likelihood framework, in their case $\Sigma(\theta^0, \theta^0)$ coincides with $B(\theta^0)$ and it is a 'latent' Fisher information matrix (see section 5 for the detais). Secondly, Renault and Touzi's asymptotic result rests upon a slightly stronger assumption than (10). More precisely, in order to ensure that, for a fixed $T$, $\theta_T^{(p)}$ defined in (8) converges when $p$ grows to infinity, they assume that, for $T$ sufficiently large, $\overline{\theta}(\lambda(\cdot))$ becomes strongly contracting and, consequently, it has a unique fixed point.

## 4 Recursive $M-$ estimation

Given the computational burden that might be implied by the iterative estimator presented above, we propose in this section several recursive procedures. The asymptotic results on the corresponding recursive estimators are quite similar and comparable with the ones obtained for the iterative estimation (see Patilea and Renault (1997) for the detailed results and proofs).

## 5 Application to a class of structural econometric models

### 5.1 Latent and observable statistical models

Many structural econometric models (nonlinear rational expectations, option pricing, auction models, ...) characterize observable variables as highly

nonlinear functions of some latent variables. These functions are one-to-one, but they depend on the unknown distribution of the latent variables through the equilibrium of the game or the learning process. Therefore numerical complexity of the equilibrium definition generates substantial obstacles for the direct implementation of maximum likelihood inference. Motivated by the fact that the law of motion of the latent variables is often defined in a fairly simpler way, simulation-based strategies (e.g., indirect inference) have been developed recently. The methodology we proposed above allows for alternative estimation procedures based on learning on the latent variables in order to perform approximated MLE directly inside the more tractable latent model. We briefly present these alternative estimation procedures below.
Let $Y_t^*$, $t = 0, 1, 2, ...$, be a sequence of homogeneous Markovian, of order one, $J \times 1$ random vectors defined on some abstract probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and taking values in some $\mathcal{Y}^* \subset \mathbf{R}^J$. We specify a $p-$dimensional parametric model for the conditional probability distribution of the process $Y^*$, given the initial value $Y_0^*$, through the family of transition densities

$$\mathcal{M}^* := \{f^*(\cdot \mid \cdot\,; \theta),\ \ \theta \in \Theta \subset \mathbf{R}^p\} \tag{11}$$

defined with respect to the Lebesgue measure on $\mathbf{R}^J$. The model $\mathcal{M}^*$ is such that for any $\theta \in \Theta$ the transition $f^*(\cdot \mid \cdot\,; \theta)$ allows for a unique stationary initial distribution. Moreover, we assume that the model we consider is correctly specified, that is for some $\theta^0 \in \Theta$, called the true unknown value of the parameters,

$$\prod_{t=1}^{T} f^*(y_t^* \mid y_{t-1}^*\ ; \theta^0)$$

is a correct description (in terms of densities) of the DGP providing the latent data ${Y^*}_1^T = (Y_t^*)_{1 \le t \le T}$, given that $Y_0^* = y_0^*$.
The maximum likelihood estimator in the latent (unobservable) world, denoted by $\hat{\theta}_T^*$, is defined as

$$\hat{\theta}_T^* = \arg\max_{\theta \in \Theta} \sum_{t=1}^{T} l^*(Y_t^* \mid Y_{t-1}^*\ ; \theta) \tag{12}$$

where, for any $\theta$,

$$l^*(\cdot \mid \cdot\,; \theta) := \log f^*(\cdot \mid \cdot\,; \theta).$$

This estimator is not computable since we do not observe the process $Y^*$. As it has been stressed above, the specificity of the framework we consider lies in the observation scheme. The observed data $Y_1^T = (Y_t)_{1 \le t \le T}$ and the

observed initial value $Y_0$ belong to some $\mathcal{Y} \subset \mathbf{R}^J$ and are given by a known one-to-one transformation $g(\cdot, \theta^0)$ (depending on the true unknown parameter [5] ) of the latent data ${Y^*}_1^T$ and the latent initial value, respectively. That is, for any $t = 0, 1, 2, ...,$

$$Y_t = g(Y_t^*, \theta^0) \Longleftrightarrow Y_t^* = g^{-1}(Y_t, \theta^0). \tag{13}$$

In this framework, a maximum likelihood procedure in the observable world would maximize the criterion

$$\begin{aligned} \tilde{Q}_T[\theta, \lambda(\theta)] &= \frac{1}{T}\sum_{t=1}^{T} l^* \left(g^{-1}(Y_t, \theta) \,|\, g^{-1}(Y_{t-1}, \theta) \;; \theta\right) \\ &+ \frac{1}{T}\sum_{t=1}^{T} \log \left|J_y g^{-1}(Y_t, \theta)\right|. \end{aligned} \tag{14}$$

We denote by $|J_y g^{-1}(\cdot, \theta)|$ the absolute value of the Jacobian with respect to $y$ of the one-to-one mapping $y \to g^{-1}(y, \theta)$. We have in mind the case where this criterion is very complicated. This happens, for instance, in the option pricing model considered by Christensen (1992). In such a framework, at any time $t$, $Y_t^*$ represents a vector of unobservable state variables while $Y_t$ is a vector of observed option prices. The relationship $g(\cdot, \theta^0)$ is provided by arbitrage-based derivative asset pricing à la Harrison and Kreps (1979). The observable model behind (15) was called by Christensen 'the empirical martingale model'. Renault and Touzi (1996) focused on the at-the-money (ATM) option case. They considered $Y^*$ as being the (latent) volatility process of the underlying asset and $Y_1^T$ a time series of prices of ATM European options (with fixed maturity period) written on this underlying asset. Moreover, they used the Hull and White (1987) option pricing formula for the passage $g(\cdot, \theta)$ between the two worlds, latent and observable. Typically, the dynamics of the latent process $Y^*$ considered by Renault and Touzi can be described by a simple diffusion process (for instance the exponential of an Ornstein-Uhlenbeck process) in such a way that the latent log-likelihood is "nice". Meanwhile, their observable log-likelihood is "not nice", that is cumbersome to maximize, since it involves highly nonlinear functions of the unknown parameters. Facing such kind of problems, Renault and Touzi (1996) (see also Pastorello *et al.* (2000)) have proposed a two stage estimation procedure which represented the starting point of this paper.

In the framework of this section, the criteria appearing in the previous sections become

$$Q_T[\theta, \lambda(\theta^1)] = Q_T[\theta, \theta^1] = \frac{1}{T}\sum_{t=1}^{T} l^* \left(g^{-1}(Y_t, \theta^1) \,|\, g^{-1}(Y_{t-1}, \theta^1) \;; \theta\right)$$

and

$$Q_\infty[\theta, \lambda(\theta^1)] = Q_\infty[\theta, \theta^1] = E_{\theta^0}\left[l^* \left(g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \;; \theta\right)\right]$$

(in this section $E_{\theta^0}$ denotes the expectation considered w.r.t. the stationary distribution of $(Y_1, Y_0)$, characterized by $\theta^0$).
Let us notice that, despite what happens in usual partial observation schemes in econometrics (as those corresponding to Probit, Tobit,..., models), in our framework there is no automatic loss of information when passing from the latent to observable world. In order to justify this statement we recall that $g(\cdot, \theta^0)$ is a one-to-one transformation and we remark that the observable and latent models may swap their roles. Thus our estimation procedures could not be immediately be justified by a necessary richer information about $\theta^0$ contained in the latent variables. However, we have in mind some structural models written in a simple and informative way in what concerns the structural parameters $\theta$. Meanwhile, the observable likelihood is clearly more complicated, even intractable, and *intuitively* less informative. The example below reveals some information and identifiability aspects specific for our framework.

**Example 1** (*About the loss of information and identifiability*)
Let us assume that the variables $Y_t^*$, $t = 1, ..., T$, are i.i.d., normal with mean $\theta$ and variance one. Assume that $Y_t = Y_t^* - \theta/2$ and remark that the Cramer-Rao (CR) bound in the observable world equals four while in the latent world it is equal to one. In other words, there is an information loss in estimating $\theta$ when passing from the latent world to the observable one[6].
Let us remark that in our framework we may even loose the whole information on (a part of) the parameters, that is (a part of) the parameters may become unidentifiable. To see this let us consider $Y_t^*$, $t = 1, ..., T$, as above and take $Y_t = Y_t^* - \theta$ which makes $\theta$ unidentifiable from the observables. We can also exemplify the identification problems in a dynamic model by extending the specification above to a AR(1) Gaussian process:

$$Y_t^* = \rho Y_{t-1}^* + \varepsilon_t, \quad t \in \mathbb{Z},$$

with $|\rho| < 1$ and $\varepsilon_t$, $t \in \mathbb{Z}$, i.i.d. $(0, \sigma^2)$ r.v.'s. In this case we remark that a transformation like $Y_t = Y_t^*/\sigma$ does not allow to identify $\sigma$, even though $\rho$ is still identifiable. $\diamondsuit$
The previous example warns us that in our framework, due to (apparently innocent) one-to-one transformation, the structural parameters may easily become partially or totally unidentifiable in the observable world, even if they are identifiable in the latent world[7]. When the parameters are not identifiable in the observable world, it is clear that the two stage estimation strategy we adopt will fail to consistently estimate the true parameters.
Following the facts presented in section 2, let us denote

$$\begin{aligned} \overline{\theta}(\theta^0, \theta^1) &:= \arg\max_{\theta \in \Theta} E_{\theta^0} \left[ l^* \left( g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \; ; \theta \right) \right], \qquad (15) \\ &= \arg\max_{\theta \in \Theta} E_{\theta^0} \left[ l^* \left( g^{-1}(Y_1, \theta^1) \,|\, g^{-1}(Y_0, \theta^1) \; ; \theta \right) + \log \left| J_y g^{-1}(Y_1, \theta^1) \right| \right]. \end{aligned}$$

Recall that the basic identification condition in the latent model $\mathcal{M}^*$ can be rewritten as $\overline{\theta}(\theta^0, \theta^0) = \theta^0$. Moreover, the strengthened identification condition for our framework where we consider a two stage estimation strategy was that the unique fixed point of $\overline{\theta}(\theta^0, \cdot)$ is $\theta^0$ (see Assumption 2.2). Clearly, the reinforcement of the basic identification condition, is stronger than the identification in the observable world. The remark above on the possible loss of indentifiability when passing from one world to the other gives an idea about how restrictive our reinforced identification condition could be.
The intuitions we sketched above and in section 2, leading us to state Assumption 2.2, may raise the question if the unique fixed point property for the function $\overline{\theta}(\theta^0, \cdot)$ is necessary for ensuring the identification in the observable model

$$\mathcal{M} = \left\{ f^* \left( g^{-1}(\cdot, \theta) \,|\, g^{-1}(\cdot, \theta) \; ; \theta \right) \left| J_y g^{-1}(\cdot, \theta) \right|, \;\; \theta \in \Theta \right\}.$$

The counterexample below shows that the answer is negative.

**Example 2** (*Counterexample*)
Let us consider a sequence of i.i.d. r.v.'s $Y^* = (Y_t^*)_{t \geq 1}$ and a family of exponential probability density functions

$$f^*(y_t^*; \theta) = \theta \exp(-\theta y_t^*) \, 1\!\!1_{(0,\infty)}(y_t^*),$$

with $\theta \in \Theta = [1/2, \, 3/2]$. We assume that the distribution of $Y_1^*$ is given by $f^*(\cdot; \theta^0)$, for some $\theta^0 \in \Theta$. The observable variables are obtained via the transformation

$$Y_t = g(Y_t^*, \theta^0) = Y_t^* + \theta^0.$$

In this case

$$E_{\theta^0} \left[ l^* \left( g^{-1}(Y_1, \theta^1); \theta \right) \right] = \log \theta - \theta \left( \frac{1}{\theta^0} + \theta^0 - \theta^1 \right),$$

and

$$\overline{\theta}(\theta^0, \theta^1) = \left( \frac{1}{\theta^0} + \theta^0 - \theta^1 \right)^{-1}.$$

As $\theta^0$ and $1/\theta^0$ are both fixed points for $\overline{\theta}(\theta^0, \cdot)$, we obtain that Assumption 2.2 is violated, provided that $\theta^0 \neq 1$. However, $\theta^0$ is clearly identifiable in the observable world since the support of the variables $(Y_t)_{t \geq 1}$ is $[\theta^0, \infty)$.
$\diamondsuit$
The previous example shows that it may happen that the two stage strategy we adopt in this paper cannot be applied, while the ML estimation based on the full criterion (15) could, *theoretically*, provide consistent estimation. However, we argue that there exists an important class of structural

econometric models where the ML estimator based on the full criterion is not computable. Meanwhile, the two stage strategy we considered herein applies. For example, Renault and Touzi (1996) reported that the unique fixed point property of $\overline{\theta}(\theta^0, \cdot)$ seems to be verified in the Hull and White (1987) option pricing model.

### 5.2 Iterative estimators

In the previous subsection we stressed our preference for the latent world and, guided by our choice, we gave additional motivations for the general identification condition written in section 2. Now remark that in this framework, definition (6) becomes

$$\overline{\theta}_T(\theta^1) = \arg\max_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^{T} l^* \left( g^{-1}(Y_t, \theta^1) \,|\, g^{-1}(Y_{t-1}, \theta^1) \; ; \theta \right), \quad \theta^1 \in \Theta. \tag{16}$$

The motivation for this definition is that, once we have in hand a proxy value of $\theta^0$, we will always look for bringing the observations in the latent world where the structural model is defined. The estimation strategy of transforming data via an approximating value of $\theta^0$ and performing ML estimation in the latent model $\mathcal{M}^*$ has some common features with the strategy of the EM algorithms introduced by Dempster *et al.* (1977). Thus we can consider that the argmax estimators obtained using $\overline{\theta}_T(\cdot)$ are EM type estimators. However, as it has been remarked by Renault and Touzi (1996), the general theory of EM algorithms can not be used in this framework since, for any $t$, the distribution of $Y_t^*$ given $Y_t$ is a degenerate one the support of which depends on the unknown value of the parameters.
We may remark that $\overline{\theta}_T(\theta^0)$ coincides with the MLE in the latent model! Using $\overline{\theta}_T(\cdot)$ written in (16) we may rewrite the two estimators of section 3. Note that in the framework considered in this section the matrix $B(\theta^0)$ defined in Assumption 3.3 represents the inverse of the Fisher information in the latent model and coincides with the matrix $\Sigma(\theta^0, \theta^0)$ of Assumption 3.4. Proposition 3.3 tells that, under some conditions, the Renault and Touzi (1996) estimator and the iterative estimator presented in section 3 are asymptotically normal with the same asymptotic variance

$$V(\theta^0) = \left( I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \theta^1) \Big|_{\theta^1 = \theta^0} \right)^{-1} \Sigma(\theta^0, \theta^0)^{-1} \left( I_p - \frac{\partial \overline{\theta}'}{\partial \theta^1}(\theta^0, \theta^1) \Big|_{\theta^1 = \theta^0} \right)^{-1}.$$

In the limit case where, for any $\theta^1 \in \Theta$, $g^{-1}(\cdot, \theta^1)$ is the identity function, that is when one observes directly the latent variables, as expected, $V(\theta^0)$ becomes equal to $B(\theta^0)^{-1}$, the CR bound in the latent model. Let us recall that in subsection 5.1 we have remarked that in our framework there is no

general rule stating that the latent world should reveal more information than the observable one does.
On the other hand, it is clear that $V(\theta^0)$ should be greater than the asymptotic CR bound corresponding to the observable model $\mathcal{M}$. In other words, it is obvious that the iterative estimator given in (7) should be asymptotically less concentrated than the intractable MLE obtained from the observable log-likelihood (15). It remains a natural question : is $V(\theta^0)$ greater or smaller than $B(\theta^0)^{-1}$? In the example below we show that there is no generally valid order relationship between these two matrices.

**Example 3**
For the sake of simplicity, let us consider, as in Example 2, that the variables $Y_t^*$, $t = 1, ..., T$, are i.i.d. as normals of mean $\theta$ and variance 1. Moreover, $Y_t = Y_t^* + \theta/2$, thus $g^{-1}(y, \theta) = y - \theta/2$. In this case $\overline{\theta}(\theta^0, \theta^1) = 3\theta^0/2 - \theta^1/2$ and $(d\overline{\theta}/d\theta^1)(\theta^0, \theta^1) = -1/2$, for any $\theta^1 \in \Theta$. As a consequence $V(\theta^0) = 4/9$ and we remark that $V(\theta^0)$ coincides with the CR bound in the observable world, that is the two argmax estimators mentioned above are asymptotically equivalent with the MLE obtained from the observable log-likelihood. Moreover, $V(\theta^0)$ is smaller than $B(\theta^0)^{-1}$ which equals one. Swapping the roles of the variables (as we have done in Example 2) we obtain a reverse conclusion: $B(\theta^0)^{-1}$ is smaller than $V(\theta^0)$. It is clear that there exists a close link between the form of the matrix

$$I_p - \frac{\partial \overline{\theta}}{\partial \theta^{1\prime}}(\theta^0, \theta^1)\Big|_{\theta^1 = \theta^0}$$

and the relationship between the information bounds in $\mathcal{M}^*$ and $\mathcal{M}$. This relationship is to be examined in future work. $\diamondsuit$

**Footnotes**
[1] The so-called efficient method of moments (EMM; see Gallant and Tauchen (1995)) recovers efficiency through the choice of an infinite set of moments to match stemming from an Hermite type expansion of the conditional probability distribution. However, there is no general asymptotic theory on the choice of the 'tunning parameter', that is the number of moments to match for a given sample size.
[2] It is clear that, in general, the identification condition based on the simplified limit criterion neither implies, nor is implied by the identification condition based on the full limit criterion.
[3] Remark that $\overline{\theta}_T(\lambda)$ may not be uniquely defined. The asymptotic results hold for any version of $\overline{\theta}_T(\lambda)$, $T \geq 1$.
[4] Note that we only need the uniform convergence of $Q_T[\cdot, \lambda(\cdot)]$, in probability. By abuse, from now on, any time we invoke Assumption 3.1 we think at the convergence of $Q_T[\theta, \lambda(\theta^1)]$, in probability, uniformly with respect to $\theta$, $\theta^1 \in \Theta$.

[5] In this framework, the function $\lambda(\cdot)$ appearing in the previous sections is the identity function. A more general case can be also considered.
[6] Of course, if we swap the roles of the variables, that is if we consider $Y_t^*$, $t = 1, ..., T$, normally distributed with mean $\theta/2$ and variance one and $Y_t = Y_t^* + \theta/2$, we find a reverse result: the latent variables are less informative then the observable ones.
[7] If we swap the roles of $Y$ and $Y^*$ we can also deduce from Example 1 above that identification could hold in the observable model without necessarily being ensured in the latent one. Due to the fact that herein we favor the latent world where we build our structural model, we will always look for identification and consistency assumptions for the observable model starting from assumptions on the latent one.

**Acknowledgements:** V. Patilea gratefully acknowledges research support from the contract "Projet d'Actions de Recherche Concertées" (PARC No.93/98-164) of the Belgium Government. We thank Halbert White for drawing our attention on the recursive estimation procedures relevant for the framework considered in this paper. At a preliminary stage of this paper we benefited from the discussions with Marie-Pierre Ravoteur and Nizar Touzi.

## 6 References

Christensen, B.J. (1992). Asset prices and the empirical martingale model, *Working Paper, New-York University.*

Clark, D. (1984). Necessary and sufficient conditions for the Robbins-Monro method, *Stochastic Processes and their Applications*, **17**, 359-367.

Davidson, J. (1994). *Stochastic Limit Theory*, Advanced Texts in Econometrics, Oxford Univ. Press.

Gouriéroux C., A. Monfort and E. Renault (1993). Indirect inference, *Journal of Applied Econometrics*, vol. **8,** , S85-S118.

Harrison, J. and D. Kreps (1979). Martingales and arbitrage in multiperiod security markets, *Journal of Economic Theory*, **20**, 381-408.

Horn, R. and C. Johnson (1985). *Matrix Analysis*, Cambridge University Press.

Horn, R. and C. Johnson (1991). *Topics in Matrix Analysis*, Cambridge University Press.

Hull, J. and White A. (1987). The pricing of options on assets with stochastic volatilities, *Journal of Finance*, **42**, 281-300.

Kuan, C.M. and H. White (1994a). Artificial neural networks: an econometric perspective, *Econometric Reviews*, **13(1)**, 1-91.

Kuan, C.M. and H. White (1994b). Adaptive learning with nonlinear dynamics driven by dependent processes, *Econometrica*, **62**, 1087-1114.

Kushner, H.J. and D. Clark (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New-York.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing, *Handbook of Econometrics, Vol IV*, R.F. Engle and D. McFadden eds., 2111 - 2245.

Pastorello S., E. Renault and N. Touzi (2000). Statistical inference for random variance option pricing, forthcoming *JBES*.

Patilea, V and E. Renault (1997). Continuously updated estimators,*CORE DP 9776*, Louvain-la-Neuve.

Patilea, V., M.P. Ravoteur and E. Renault (1995). Multivariate time series analysis of option prices, *Mimeo, GREMAQ*, Toulouse.

Renault, E. (1997). Econometric models of option pricing errors, *Advances in economics and econometrics: theory and applications, Seventh World Congress*,vol. **III**, chapter **8**, D. Kreps and K. Wallis (eds.), Cambridge Univ. Press.

Renault, E. and N. Touzi (1996). Option hedging and implied volatilities in a stochastic volatility model, *Mathematical Finance*, Vol. **6**, No. **3** , 279-302.

Robbins, H. and S. Monro (1951). A stochastic approximation method, *Annals of Mathematical Statistics*, **22**, 400-407.

Rouche, N. and J. Mawhin (1980). *Ordinary Differential Equations*, Pitman Advanced Publishing Program.

White, H. (1989). Some asymptotic results for learning in single Hidden-Layer feedforward network models, *Journal of the American Statistical Association*, Vol. **84**, No. **408**, 1003-1013.

Wooldridge, J. (1994). Estimation and inference for dependent processes, *Handbook of Econometrics*, Vol **IV**, R.F. Engle and D. McFadden eds., 2641-2739.

Young, P. (1985) Recursive Identification, Estimation and Control, *Handbook of Statistics*, Vol. **5**, Elsevier Science Publishers.

# Jump Diffusion Processes with Shot Noise Effects and their Applications to Finance

Winfried Stute[1]

[1] Mathematical Institute, University of Giessen, Arndtstr. 2, D-35392 Giessen, Germany

**Abstract:** In this paper we consider an extension of a jump diffusion process to one incorporating shot noise effects. The main purpose is to provide a model for stock prices which consists of two components: a smooth one driven by a geometric Brownian Motion and another one consisting of a compound Poisson Process. In our extension a decay function is introduced designed to model a situation, in which the jump effects may fade away on the long run. The main result of the paper provides a risk-neutral measure under which the observed (discounted) process is a martingale. This measure is crucial for pricing derivatives of the underlying. A small simulation study is also included.

**Keywords:** Jump diffusion; Shot noise effects; Martingale measure.

## 1 Introduction

If one tries to summarize what "Mathematics in Finance" is all about, one may come up with three goals, to mention only the most important:

1. Provide Methods designed to price Financial Derivatives (like Options or Futures)
2. Provide Methods for Hedging Portfolios
3. Study Methods to assess Financial and Credit Risks

At the same time the methodology is expected to also have a stabilizing effect on the market, as (on the long run) pricing should prevent riskless profits which may cause undesirable market instabilities. Clearly, the absence of "Arbitrage" does not exclude unexpected (random) movements resulting in nonpredictable returns and risks. Mathematically, "No Arbitrage" is achieved if pricing takes place in a so-called risk-neutral world. For example, if $(S_t)_t$ is a stochastic process representing the prices of an underlying, then ideally the price of a corresponding contingent claim at time $t$ with maturity $T$ should be

$$C_t = e^{-r(T-t)}\mathbb{E}_Q\left[C_T|\mathcal{F}_t\right],$$

where $Q$ is a measure such that the (discounted) process $(S_t)_t$ is a martingale w.r.t. $Q$, $C_T$ is the payoff of the derivative at time $T$ and $(\mathcal{F}_t)_t$ is a proper filtration. Here $r$ denotes the interest rate of a riskless asset. Of course, the measure $Q$ will depend on the structure of the underlying model for $(S_t)_t$. Informally speaking, if it exists, it will be unique only if the model for $(S_t)_t$ is simple. For example, in the Black-Scholes model, $(S_t)_t$ satisfies the stochastic differential equation

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t, \tag{1}$$

where $\mu$ is a constant drift, $\sigma$ is a constant volatility and $(W_t)_t$ is a Brownian Motion. Equation (1) admits the explicit solution

$$S_t = S_0 \exp\left\{\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right\}, \tag{2}$$

a geometric Brownian Motion with drift.

Typically, derivatives on $(S_t)_t$ have a maturity $T$ so that it suffices to restrict the process to the time interval $[0, T]$, say. Since the martingale measure $Q$ will be assumed to be absolutely continuous w.r.t. the true measure, it is only required to compute the Radon-Nikodym derivative $\xi = \xi_T$. In the case of (1) respectively (2), we have

$$\xi_T = e^{-\frac{(r-\mu)^2 T}{2\sigma^2} + \frac{r-\mu}{\sigma} W_T}. \tag{3}$$

Now, in practice, it is known that model (1) is unable to produce several phenomena which can be observed on markets. One of the most serious drawbacks is the inability of (1) to explain volatility clusters in the returns. Another point of criticism is concerned with the smoothness of the sample paths, ignoring the possibility that in reality jumps may occur. In the literature there have been several attempts to extend or modify model (1) in order to incorporate such effects:

(i) Replacing the constant $\sigma$ with an external stochastic process $\sigma_t$ leading to varying (conditional) volatilities.

(ii) Replacing the Brownian Motion with a more general stable process, to the effect that log-returns need not be normally distributed and jumps may occur.

(iii) Adhering to the Brownian Motion, but incorporating also a (compound) Poisson Process being responsible for random jumps.

In practice, these abrupt changes may be caused by political or corporate news, as the resignment of Boris Yeltsin on December 31, 1999, which

launched a rapid increase of stock prices at the Moscow Stock Exchange. On the other hand, it is often seen that after some time these effects fade away, at least partly. There are simple explanations for that. An upward jump may be followed by profit taking, while a jump downward may encourage new investors to buy the asset. In terms of the stochastic process literature, these impulses create Shot Noise effects. We now present a model which captures all effects described above and which may be expected to offer a better market model:

(i) A continuous part, which models the behavior of stock prices between two successive jumps.

(ii) A "jump"-part, which takes care of the possibility that abrupt changes may occur.

(iii) A decay function $h$, which controls the effect of noise depending on the time elapsed since a jump occurred.

**Model:**

$$S_t = S_0 e^{(\mu - \frac{\sigma^2}{2})t + \sigma W_t} \prod_{j=1}^{N_t} [1 + U_j h(t - \tau_j)] \tag{4}$$

Here

- $(N_t)_{t \geq 0}$ is a Poisson Process with intensity $\lambda$
- $(U_j)_j$ is a sequence of independent identically distributed "jump sizes"
- $h$ is an appropriate decay function
- $\tau_j$ is the time of the $j$-th jump since $t = 0$

To define the joint stochastic behavior, we assume that

$$(W_t)_t, (N_t)_t \text{ and } (U_j)_j \text{ are independent.}$$

The function $h$ is that part of the model which may be chosen so as to create special effects for the price process of the underlying. For $h = 1_{\{t \geq 0\}}$, we obtain the usual jump diffusion process. See Merton (1976) and Lamberton and Lapeyre (1997). It is easy to see that in this case $(S_t)_t$ is Markov. For a general $h$, this need no longer be true. The aforementioned shot noise effects may be obtained, e.g., if for $h$ we take

$$h(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ \exp(-at) & \text{for} \quad t \geq 0 \end{cases} .$$

The nonnegative parameter $a$ takes care of the speed at which a jump may fade away. In this example, it completely vanishes on the long run. If we

add a positive constant to $h$ on $t \geq 0$, we obtain abrupt changes whose influence survives at least partly.
Interestingly enough, Model (4) is flexible enough to also generate completely different effects. In the so-called "rumor-model", an impulse (the "rumor") at $\tau_j$ may create some continuous increase (or decrease) of the price until a maximum effect is reached with a certain delay. A possible candidate for such a function $h$ is given by

$$h(t) = \begin{cases} 0 & \text{for} \quad t \leq 0 \\ \exp\left[-\frac{(\ln t - a)^2}{b}\right] & \text{for} \quad t > 0 \end{cases},$$

where $a \in \mathbb{R}$ and $b > 0$.
Finally, we would like to mention that Model (4) can be enlarged by incorporating another factor $\prod_{j=1}^{\tilde{N}_t}[1+\tilde{U}_j\tilde{h}(t-\tilde{\tau}_j)]$. In this way we could come up with a model featuring, e.g., both "(real) shot noise" and "rumor" effects.
In Section 2 we briefly describe a general method how to construct a martingale measure $Q$ for discrete time processes and apply this to Model (4). In Section 3 we go to the limit and present the Likelihood Process in continuous time. Finally, Section 4 presents some simulation results. Detailed proofs will appear elsewhere.

## 2 Computation of Martingale Measures in Discrete Time

Generally speaking, the martingale measure $Q$ will always depend on the model for the underlying process. Hence risk neutrality does not mean that pricing takes place in an abstract world, but is oriented towards our understanding of what may happen in reality. A wrong model may therefore lead to a wrong specification of prices. For example, smile effects are just an outgrowth of applying a model which does not fit well in practice. These comments were included only to make clear, that a good pricing policy is not just to apply some formulae from probability theory, but is also a statistical issue in that one should aim at finding models with a satisfactory goodness-of-fit. On the other hand, for complicated models the measure $Q$ may not be unique. In other words, there may be no general agreement on how to price an asset. So, in this sense, there is a tradeoff between the goodness-of-fit of a model and the wish for a unique pricing policy.
The last remarks in particular apply to (4). In what follows we shall therefore concentrate on deriving a special $Q$. The choice of an appropriate martingale measure in so-called incomplete models has been the subject of much research. See, e.g., Delbaen and Schachermayer (1990) and Schweizer (1996). Typically, $Q$ is chosen so that its density with respect to $\mathbb{P}$ minimizes a given objective function. Often only existence can be proved, and an explicit computation of $Q$ in complex models may be complicated. In

this paper we follow Dothan's (1990) approach which is based on a decomposition of the returns into an innovation and a predictable part. We shall come up with explicit formulae which allow for an implementation and therefore for a numerical assessment of the method. For the sake of completeness we briefly discuss the major steps. Put, for any sequence $(\tilde{S}_n)_n$ adapted to a filtration $(\mathcal{F}_n)_n$,

$$r_n = \frac{\tilde{S}_n - \tilde{S}_{n-1}}{\tilde{S}_{n-1}},$$

the associated process of returns. Set

$$\mu_n = \mathbb{E}_{\mathbb{P}}(r_n|\mathcal{F}_{n-1}) \text{ – the expected return –}$$

and define

$$M_n := M_{n-1} + \frac{\tilde{S}_n}{\tilde{S}_{n-1}} - \mathbb{E}_{\mathbb{P}}\left(\frac{\tilde{S}_n}{\tilde{S}_{n-1}}|\mathcal{F}_{n-1}\right).$$

Hence $(M_n)_n$ is a martingale with respect to $\mathbb{P}$, the true measure driving $\tilde{S}_n$. Putting $\Delta M_n = M_n - M_{n-1}$, we then obtain

$$\frac{\tilde{S}_n}{\tilde{S}_{n-1}} = 1 + \mu_n + \Delta M_n.$$

This is the representation of the returns in terms of the innovations $\Delta M_n$ and the predictable $\mu_n$. Conclude that

$$\tilde{S}_n = \tilde{S}_0 \prod_{i=1}^{n} [1 + \mu_i + \Delta M_i].$$

This is the exponential representation of $\tilde{S}_n$ in terms of $\mu$ and $M$. As mentioned earlier the desired $Q$ we are looking for will be abolutely continuous w.r.t. $\mathbb{P}$ and is therefore uniquely determined through its Radon-Nikodym derivative

$$\xi = \frac{dQ}{d\mathbb{P}}.$$

Put

$$L_n = \mathbb{E}_{\mathbb{P}}[\xi|\mathcal{F}_n] \text{ – the likelihood process –}$$

From Girsanov's Theorem we then obtain

$$\mathbb{E}_{\mathbb{P}}[\Delta M_n L_n|\mathcal{F}_{n-1}] = -(\mu_n - r)L_{n-1}.$$

One solution is given by

$$L_n = \prod_{i=1}^{n} \left[1 - \frac{(\mu_i - r)\Delta M_i}{\mathbb{E}_{\mathbb{P}}[\Delta^2 M_i|\mathcal{F}_{i-1}]}\right] \equiv \prod_{i=1}^{n} [1 - l_i], \tag{5}$$

provided the factors are nonnegative.
In the following we consider arbitrary $t < T$. As before, time $T$ may be interpreted as the maturity of a contingent claim.
Coming back to (4), we put for a given $N$:

$$\tilde{S}_i = S_{t+\frac{i}{N}(T-t)}.$$

Hence

$$\tilde{S}_N = S_T \text{ for each } N.$$

For $\mathcal{F}_i$ we take the $\sigma$-field containing the information prior to $t_i := t + \frac{i}{N}(T-t)$. Exploiting the very structure of (4), we get the following result, which presents an explicit formula for $l_i$. Note that the interest rate needs to be adjusted to the new time scale. $t$ will be set $t = 0$.

**Theorem 2.1** *For model (4), we have for each* $1 \le i \le N$*:*

$$l_i = \frac{e^{(\mu-r)\frac{T}{N}+\lambda\mathbb{E}_{\mathbb{P}}(U_1)\int_0^{T/N} h(x)dx} - P_{i-1}^{-1} e^{(r-\mu)\frac{T}{N}-\lambda\mathbb{E}_{\mathbb{P}}(U_1)\int_0^{T/N} h(x)dx}}{e^{\sigma^2\frac{T}{N}+\lambda\mathbb{E}_{\mathbb{P}}(U_1^2)\int_0^{T/N} h^2(x)dx} - 1} \times$$

$$\left( e^{-\frac{\sigma^2 T}{2N}+\sigma\Delta W_{t_i}-\lambda\mathbb{E}_{\mathbb{P}}(U_1)\int_0^{T/N} h(x)dx} \left\{ \prod_{k=N_{t_i-1}+1}^{N_{t_i}} \left[ 1 + U_k h\left(\frac{iT}{N} - \tau_k\right)\right]\right\} - 1 \right).$$

*Here*

$$P_i = \prod_{k=1}^{N_{t_i}} \frac{1 + U_k h\left(\frac{(i+1)T}{N} - \tau_k\right)}{1 + U_k h\left(\frac{iT}{N} - \tau_k\right)}.$$

*Note that* $P_i = 1$ *for* $h = 1_{\{\cdot>0\}}$*, i.e., when there are jumps but no shot noise effects.*

## 3 The Likelihood Process in Continuous Time

>From Theorem 1 we see that the likelihood process is not simple at all. Actually, we preferred to give the formulas for $l_i$ rather than $L_n$ which would be even more complicated. But, as we let the mesh of the grid tend to zero, there is some hope that due to the continuity of $(W_t)$ and the jump character of $(N_t)_t$, things may simplify (to a certain extent). The following Theorem presents the limit of (5), when $T$ is kept fixed but the mesh of the grid goes to zero. Assume that $h$ is continuously differentiable on $[0, T]$, with derivative $h'$.

**Theorem 3.1** *On* $[0,T]$, $\xi_T = dQ/d\mathbb{P}$ *equals*

$$\begin{aligned}\xi_T &= \prod_{k=1}^{N_T}\left[1 - U_k h(0)\frac{\sum_{l=1}^{k-1}\frac{U_l h'(\tau_k-\tau_l)}{1+U_l h(\tau_k-\tau_l)} - C_1}{C_2}\right] \\ &\times \exp\left[\frac{C_1\sigma}{C_2}W_T + \left(\frac{C_1\sigma^2}{2C_2} + \frac{C_1C_3}{C_2} - \frac{C_1^2\sigma^2}{2C_2^2}\right)T - \frac{\sigma}{C_2}C_4(T)\right. \\ &\quad \left. + \left(\frac{C_1\sigma^2}{C_2^2} - \frac{\sigma^2}{2C_2} - \frac{C_3}{C_2}\right)C_5(T) - \frac{\sigma^2}{2C_2^2}C_6(T)\right].\end{aligned}$$

*Here,*

$$\begin{aligned}C_1 &= r - \mu - \lambda\mathbb{E}_{\mathbb{P}}(U_1)h(0) \\ C_2 &= \sigma^2 + \lambda\mathbb{E}_{\mathbb{P}}(U_1^2)h^2(0) \\ C_3 &= -\frac{\sigma^2}{2} - \lambda\mathbb{E}_{\mathbb{P}}(U_1)h(0) \\ C_4(T) &= \int_0^T \sum_{k=1}^{N_x}\frac{U_k h'(x-\tau_k)}{1+U_k h(x-\tau_k)}dW_x \\ C_5(T) &= \int_0^T \sum_{k=1}^{N_x}\frac{U_k h'(x-\tau_k)}{1+U_k h(x-\tau_k)}dx \\ C_6(T) &= \int_0^T \left(\sum_{k=1}^{N_x}\frac{U_k h'(x-\tau_k)}{1+U_k h(x-\tau_k)}\right)^2 dx.\end{aligned}$$

Apart from $\mu, \sigma$ and $r$, which already appeared in the simplest case, namely (3), also $\lambda$ and $\mathbb{E}_{\mathbb{P}}(U_1)$ enter the formula now, not to mention $h$. This is not unexpected, of course, since these quantities represent the jump part and describe, how many jumps per time unit and in what direction they preferrably occur.

The formula dramatically simplifies if no shot noise effects are present, because in this case $h \equiv 1$ on the positive real line and therefore $h' \equiv 0$. Conclude $C_4(T) = C_5(T) = C_6(T) = 0$. Further simplifications are obtained if

$$r - \mu - \lambda\mathbb{E}_{\mathbb{P}}(U_1) = 0, \tag{6}$$

a situation which is discussed in detail in Lamberton and Lapeyre (1997). Now $C_1 = 0$ and therefore $\xi_T \equiv 1$. This of course is self-evident, since under (6) the discounted process $(S_t)_t$ is a martingale under $\mathbb{P}$ already and no change in the measure is necessary.

Finally, as another simplification, if there are no jumps at all but $r$ and $\mu$ are arbitrary, then

$$C_1 = r - \mu \qquad C_2 = \sigma^2 \qquad C_3 = -\frac{\sigma^2}{2}$$

and $\xi_T$ reduces to (3).
Even though, for a general $h$, the formula for $\xi_T$ seems complicated, the method may be easily implemented. In the following section we present a simulation study which demonstrates the impact of Shot Noise effects on pricing a European Call option.

## 4 Simulations

In our simulation study we shall compare prices for a European Call option with maturity $T = 150$ days. The underlying price process was generated according to (4), with

$$h(t) = \begin{cases} 0 & \text{for} \quad t < 0 \\ \exp[-0.02t] & \text{for} \quad t \geq 0 \end{cases} .$$

The interest rate per time unit (one day) was set $r = 0.035/365$. For $\mu, \sigma$ and $\lambda$ we took

$$\mu = 0.06/365, \sigma = 0.2/\sqrt{365} \text{ and } \lambda = 1/40,$$

i.e., on the average we expect a jump every 40 days. The jump sizes were drawn from a centered lognormal distribution:

$$U = \exp\left[-\frac{\beta^2}{2} + \beta\xi\right] - 1, \text{ with } \xi \sim \mathcal{N}(0,1).$$

We have

$$\mathrm{Var}U = e^{\beta^2} - 1,$$

so that $\beta$ determines the variance of $U$. In our simulation study we put $\beta = 0.06$. Finally, for the strike price we considered $K = 116, 131$ and $146$. We computed 4 prices. The first gives the classical Black-Scholes price with the exact $\sigma$:

$$C_t = S_t\Phi(d_1) - e^{-r(T-t)}K\Phi(d_2)$$

with

$$\begin{aligned} d_1 &= \frac{\ln\frac{S_t}{K} + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \\ d_2 &= d_1 - \sigma\sqrt{T-t}. \end{aligned}$$

Hence we apply the B-S formula ignoring the fact that jumps and Shot Noise effects may occur in the future. In other words, we apply a pricing formula in a situation where the model is misspecified. For the second price we again apply the B-S formula, but with $\sigma$ estimated from a historical chart (of length 150). This is a more realistic situation, since in reality

we never know the true $\sigma$. The third price is based on 6500 Monte Carlo simulations for possible future paths of the misspecified B-S model, again with estimated $\sigma$. Finally, we computed the prices of the Call based on 6500 paths simulated for the true model (4), upon using the representation of $\xi_T$ in Theorem 2. The estimated $\sigma$ was

$$\hat{\sigma} = 0.23690/\sqrt{365},$$

which was slightly above the true $\sigma$.

| Strike $K$ | 116 | 131 | 146 |
|---|---|---|---|
| B-S Price ($\sigma$) | 18.32 | 7.95 | 2.52 |
| B-S Price ($\hat{\sigma}$) | 19.01 | 9.17 | 3.57 |
| Monte Carlo Price (B-S; $\hat{\sigma}$) | 19.23 | 9.27 | 3.60 |
| Monte Carlo Price (Shot Noise) | 20.28 | 10.43 | 4.58 |

The B-S price with estimated $\sigma$ exceeds the theoretical B-S price based on $\sigma$, since $\hat{\sigma}$ overestimates $\sigma$. The Monte Carlo approximation is very close to the corresponding B-S price. Finally, the price taking into account future Shot-Noise effects is the largest. This is not unexpected, since possible jumps increase the volatility of the underlying price process even if their effect may diminish after a while.

## References

Delbaen, F. and Schachermayer, W. (1996). The variance-optimal martingale measure for continuous processes. *Bernoulli* **2**, 81-105.

Dothan, M. U. (1990). *Prices in Financial Markets.* Oxford University Press, New York.

Lamberton, D. and Lapeyre, B. (1997). *Introduction to Stochastic Calculus applied to Finance.* Chapman-Hall, London.

Merton, R. C. (1970). Option pricing when underlying stock returns are discontinuous. *J. of Financial Economics* **3**, 125-144.

Schweizer, M. (1996). Approximation pricing and the variance – optimal martingale measure. *Ann. Probability* **24**, 206-236.

# Combining Statistical and Computer Models for Health Risk Assessment (Exposure Analysis)

James V Zidek [1], Jean Meloche[1], Nhu D Le[2] and Li Sun[1].

[1] Department of Statistics, University of British Columbia, Vancouver, CANADA, V6T 1Z2i

[2] Cancer Control Research Program BC Cancer Agency, 600 W 10th Ave, Vancouver, CANADA, V5Z 4E6

**Abstract:** This paper gives a general model for estimating population exposure to one or more pollutants through one or more media. Although the approach must be adapted for use in particular contexts, it does provide the blueprint. Moreover general inferential procedures usable in particular contexts are given. The model has a modular structure. Its modules can to some extent be developed independently of one another. A major module represents the random behavior of individuals. For that module, we describe a flexible, internet based implementation developed to allow 24-hour recall time-activity data to be used.

**Keywords:** pNEM;$PM_{10}$;air pollution;heath risk;time-activity; hierarchical models;spatial prediction; health impact analysis; environmental risk.

## 1 Introduction

This paper features a current trend in statistical modeling. That trend is expressed here by a model that:

- is intrinsically computational rather than mathematically formulated;
- generates its stochastic outputs quickly enough to permit multiple runs and thereby yield, effectively, a predictive distribution;
- lives on the internet, can be remotely accessed by any browser on any computing platform and downloads its output onto the remote user's spreadsheet or datafile;
- admits remote user redesign and uploaded data-input.

However, like its more mathematical "cousins", it can serve as a hierarchical modeling component.
We call the model as currently implemented "pCNEM " for "a probabilistic, Canadian version of NEM". "NEM" stands for "NAAQS Exposure Model"

and in turn NAAQS stands for "North American Air Quality Standards". (*c.f.* "Probabilistic Version of the NAAQS Exposure Model'; Johnston 1984, 1987; Johnson et al 1992) However "pC" also ambiguously refers to "PC", the machine on which it is currently implemented (using Linux).
We should call the model "pCNEM/$PM_{10}$" since its current "template" is for predicting exposures to $PM_{10}$, a species of particulate air pollution of great current interest from a health and regulatory perspective. (pCNEM/$PM_{10}$ can be used to forecast the exposure to $PM_{10}$ of a random individual from any designated population cohort.) However, pCNEM's general platform and can readily be modified to suit a variety of other agents. In this paper, we restrict discussion to the current $PM_{10}$ template. Alternatives to the pCNEM approach are found in BEADS (MacIntosh et al 1995), and SHAPE (1988). However neither of these alternatives have been implemented as adaptive, web based exposure models and we confine ourselves to pCNEM in this paper
In Section 3, we will describe briefly how pCNEM/$PM_{10}$ fits into a hierarchical model for assessing the adverse health effects of $PM_{10}$, an assessment that is currently underway.

## 2 Estimating Population Exposures

The specific version of any given exposure model depends on the nature of the environmental hazard under consideration. pNEM was originally written for ozone ($O_3$)and carbon monoxide (CO). However, the importance recently attached to airborne particulates led to our subsequent development of pCNEM/$PM_{10}$. That UBC program runs quickly on a PC. It can be operated interactively by remote web-users who can upload their data-inputs and download the program's outputs onto a spreadsheet in their local computer for further analysis. Moreover, these users can change the model on-line for use in their own contexts for predicting exposures to environmental. The template we now describe if for $PM_{10}$ in Vancouver. Developing an alternative can be facilitated by loading and editing that template while at the same time adopting appropriate new input data files. Users need first to fix the study area (SA) under consideration, for example Vancouver. They then need to subdivide the SA into exposure districts (ED's) for which ambient levels of $PM_{10}$ can be predicted or measured. In an application of this model we have made, hourly values of $PM_{10}$ was predicted at 299 census tracts. However for reasons given below, these were averaged over 9 census subdivisions that formed the eventual ED's.
The population of the SA is broken down into demographic groups (D's), 14 in the case of the template for pCNEM/$PM_{10}$. Each D represents an age range, a gender and a working status (Yes or No). Each D must in turn be subdivided into cohorts of individuals sharing similar exposure histories. Thus, each D is sub-divided by ED's of residence (WD) and work (WD

where appropriate) and stove type (gas or not). The latter is important since gas stoves can be an indoor source of exposure. In this way, the population of the SA will be subdivided into a large number of homogeneous exposure subgroups.

An individual's exposure is influenced by his or her behavior. On hot days when ambient $PM_{10}$ levels are high, individuals may tend to remain in air-conditioned residential or non-residential micro-environments (ME's) including car-cabins.

pCNEM reflects the diversity of human behavior conditioned on the outdoor environment through the use of behavior "pools". In pCNEM/$PM_{10}$ each pool includes daylong records from a preliminary time-activity survey of actual surveys. Those pools are formed from a 24 hour time-activity recall data set called NHAPS.

Currently only the NHAPS data for Canada is used since the template is for Vancouver. However, the corresponding dataset for the US is also available on the site. Applications to urban areas in the US could use the latter.

Each pool corresponds to a single D. Records for individuals in that D are broken down into eight pools one for each combination of "Season" (winter or summer), "Daytype" (weekday or weekend), and "Temperature" (hot or cold). Thus for any given SA and time-period of analysis, temperature data would need to be uploaded to replace the file currently installed with the template.

Each day-long record of time and activity gives an "exposure event sequence" (EES) for that day. The elements of that sequence represent periods of actual behavior from one minute's to one hour's duration. [Each hour begins anew.] During any given period each individual's activity recorded in the NHAPS survey questionnaire data is classified by the ME in which that activity took place along with the breathing rate for the activity involved ("low", "medium" and "high"). In some applications as in the case of CO the latter will affect to dosage and physiological effects.

The ME's adopted will in general vary according to the pollutant species involved. For example being at an automobile refueling station would be an important ME for assessing exposures to benzene, a pub or restaurant for particulates.

Exposure to a pollutant such as $PM_{10}$, depends on both outdoor as well as indoor concentrations. Hourly ambient levels are estimated by the ambient monitors modulated by a spatial predictor. The values we have used are averages over the census tracts comprising the ED.

Indoor sources can be specified by the user but in the template, include the gas stove (for those cohorts where this is relevant) and presence or absence of a smoker. It would also depend on whether windows were open or closed. Other indoor sources like vacuum cleaners will be added as data become available.

After the pooling of activity patterns has been completed and the cohort

specified, a pNEM run begins at day and hour specified by the user, say hour 1 of Jan 1 for definiteness. The day is classified with the help of meteorological data by "Season", "Daytype" and "Temperature." The cohort determines D. A record can be selected at random for that day from the associated "Pool" and an EES obtains as a result.
pNEM/$PM_{10}$ works its way through the periods defined by that EES. Each period has a ME associated with it, the latter's geographical location estimated by the parameters defining the cohort. Ambient $PM_{10}$ levels estimate outside concentration, an air exchange model determines the net air flows into or out of indoor ME's through the windows. Inside such ME's, windows open, smokers "light-up" and gas stoves turn on, all at random according to empirically estimated distributions specified by the user.
Indoor concentrations vary randomly depending on such things as the random ME volume for indoor ME's, pollution decay rates, and random emission rates for stoves and cigarettes. Breathing rates together with demographic information will determine physiological reactions as a result of exposure to the random concentration of the pollutant in the ME during that period. [In the case of CO that reaction is the build-up of carbyoxihemoglobin, the rate being different for pre- and post-menopausal woman.]
Accumulating exposures over successive period will yield hourly exposure estimates along with any aggregate physiological reactions that are tracked. For the first day (Jan 1) in our hypothetical run sequence, 24 successive random hourly outputs are generated for the single randomly selected time-activity record from that associated pool. On Jan 2, a second record is randomly drawn from the associated pool and the process begins anew. The process continues through all the days of the year (or other time period selected by the user) for the selected cohort. A single record of random exposures over all the days in the period of interest for a composite individual representing that cohort will have thus been produced. [Results for that one individual can then be scaled up to the cohort according to its estimated size to generate the population distribution.] For instance if the number of days when the maximum hourly level of exposure concentrations of $PM_10$ exceeded 50 $\mu m^{-3}$ were 4 for the individual, the level for his or her cohort with say 100 individuals for example, would be estimated as 400 person-days. An estimate of the expected number of days during the period under study where the maximum daily exposures of an randomly drawn individual from the population exceeded 50 $\mu m^{-3}$ can readily be estimated by combining these cohort estimates.
A particularly important feature of pCNEM is its "roll-back" option for assessing the benefit of reducing ambient levels by any proposed new abatement policy. The model takes the fixed ambient field for the period of interest say 1996 for example and adjusts it downward according to a proposed abatement scenario. pCNEM/$PM_{10}$ can be rerun with the hypothetical new field to determine the highly nonlinear effect of the abatement program across cohorts. Moreover, differential effects for sensitive groups such

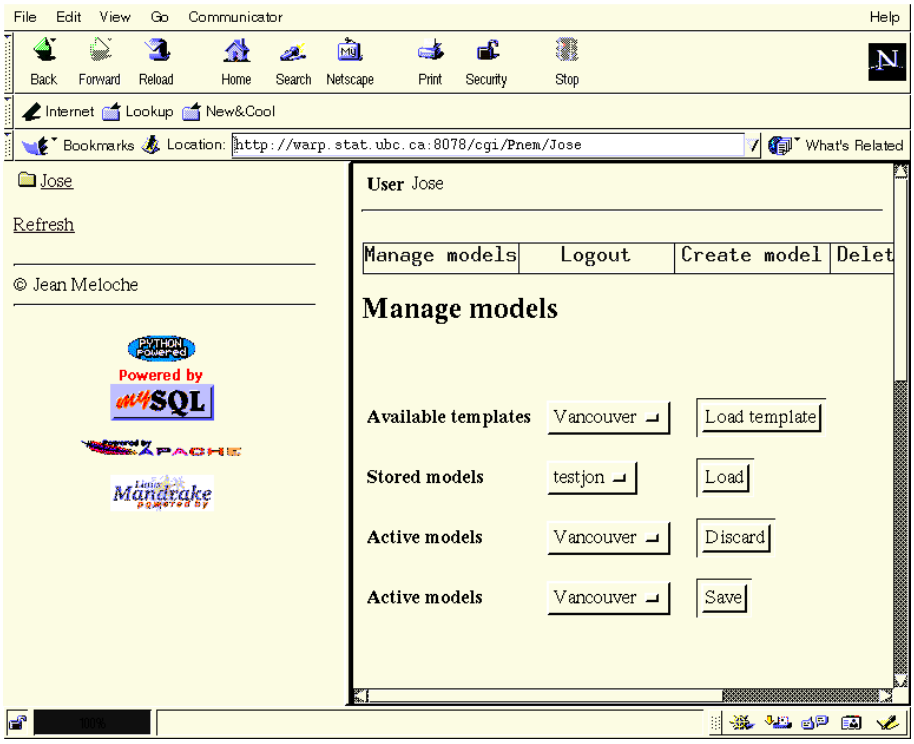



FIGURE 1. Graphical Interface for pCNEM.

as the elderly or small children can be estimated.
In Figure 1 we depict the interface for pCNEM, user "Jose" sees when he logs on to his model-site. He sees what looks like a standard windows screen and he "points-and-clicks" on it in the same was as he would if he were in fact running Windows. However, it is entirely Linux based.
At this stage, Jose is offered the choice of loading the Vancouver template or constructing his own model. If he does load it, and "clicks" on the "Vancouver folder" that ppears inside his folder he would see Figure 2. Notice in that figure the commands like those under the "Manage model" available to him. By clicking on the various folders, Jose can open up a whole area of programming options that for brevity will not be described further here.

## 3 A Model for Environmental Health Risk Analysis.

In this concluding section we briefly sketch how pCNEM/$PM_{10}$ described in the last section is to be used in a hierarchical model for assessing the health impacts of $PM_{10}$. The hierarchical model has three components:

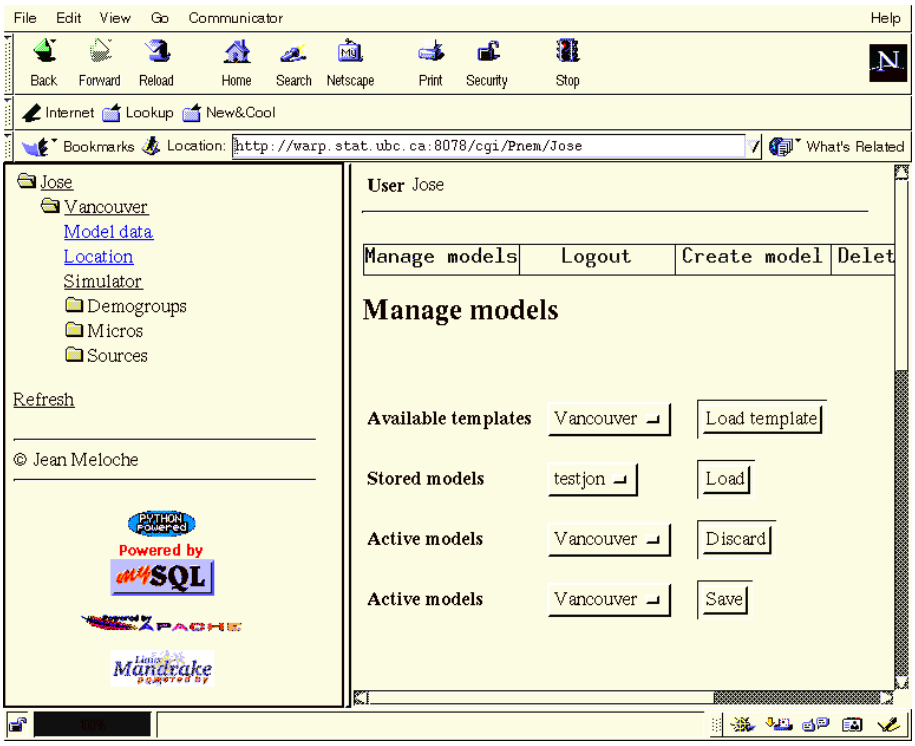



FIGURE 2. Graphical Interface After Loading the Vancouver Template.

- a spatial predictive model of ambient levels of $PM_{10}$ based on It's measured hourly levels at a number of monitoring sites;
- pCNEM/$PM_{10}$, the conditional predictive distribution of the exposure to $PM_{10}$ of a randomly selected member of a population cohort of specific interest conditional on the ambient input;
- the health effects model conditional on the unmeasured, predicted exposures.

The model in 2 (described in the previous section) incorporates the influence of human behavior and takes account of human-environment interaction. For example, in summer people may tend to stay indoors in warm weather, while in winter the reverse is true. That interaction being too complex to model mathematically is represented through time-activity data as described above. As for the model in 3, that which we briefly describe below has the associated testing theory needed to explore the association between environmental hazards and health effects.
Vancouver, where we have implemented pCNEM, has 10 (TEOM) ambient monitors for measuring levels hourly concentrations of $PM_{10}$. At the same

time pCNEM/$PM_{10}$ has 8 ED's covering a very large non-homogeneous geographical region. (That number could be changed but happens to be convenient, since the ED's can then be taken as census subdivisions for which demographic and other census data are available.) The ambient hourly level of $PM_{10}$ for any given ED could be estimated as the average of the values at the 1 or 2 ambient monitors located within its boundaries.

However because of the large geographical size of ED's and the heterogeneity of the $PM_{10}$ field, substantial measurement error would be incurred. Such error can have potentially significant, very deleterious and unpredictable consequences in the context of interest(*c.f.* Zidek et al 1997; Fung and Krewski 1999). To mitigate these potentially negative effects of measurement error the use of what Carroll et al (1995) call "regression calibration" has been advocated (*c.f.* Pierce et al 1992). In effect this process entails the prediction of the true from the measured values.

Our approach uses an adaptation of a spatial predictor ((Le and Zidek 1992; Brown, Le and Zidek 1994; Sun 1994; Le, Sun and Zidek 1997; Sun, Le, Zidek and Burnett 1998). That predictor has been shown to be well-calibrated for agents other than $PM_{10}$ (Sun 1998) and for $PM_{10}$ in Vancouver as well (Sun et al 2000). Thus for example, a 95% prediction interval covers approximately 95being predicted according to cross-validation studies. Therefore, we have some confidence in using it.

Moreover for future applications where a suite of hazardous agents are studied, the method offers great flexibility. It can handle multivariate responses. Spatial-temporal data can be accommodated and non-stationary random ambient fields are admitted.

For our current application we spatially predict $PM_{10}$ concentrations over a grid of 299 sites spread over Vancouver and environs. The predicted values over an ED are averaged to obtain a prediction for the ED as whole. This is believed to be substantially more accurate than relying on just 1 or 2 distant monitors since we would expect measurement obtained at the latter would mis-estimate the true values in the field close to the ED boundaries. We illustrate the theory for daily levels of $PM_{10}$ in Figure 1.

The figure shows the interpolated $PM_{10}$ surface on four selected summer days in Vancouver. The peaks obtain close to existing monitors. Between them, the predictor tends toward the spatial mean as one would expect due to the "regression effect." Clearly the surface is not flat. That means that a mobile individual moving will undergo variable exposures. By using the spatial predictor in conjunction with pCNEM/$PM_{10}$ we would hope to get improved estimates of actual exposures.

Key to the development of this predictor is a method for handling the non-stationarity of the spatial field. That method is due to Sampson and Guttorp (1992, hereafter SG; see also Le, Sun and Zidek 1999). That method fits to estimates of the hypercovariances between stations i and j in a hierarchical Bayesian model, $\{\Psi_{ij}\}$, a monotone decreasing function say $\zeta$ of $||\mathbf{d}_i - \mathbf{d}_j||$. Here $\mathbf{d}_i = (d_{i1}, d_{i2})$ and $d_{ik},\ k = 1, 2$ is a smooth (thin-plate

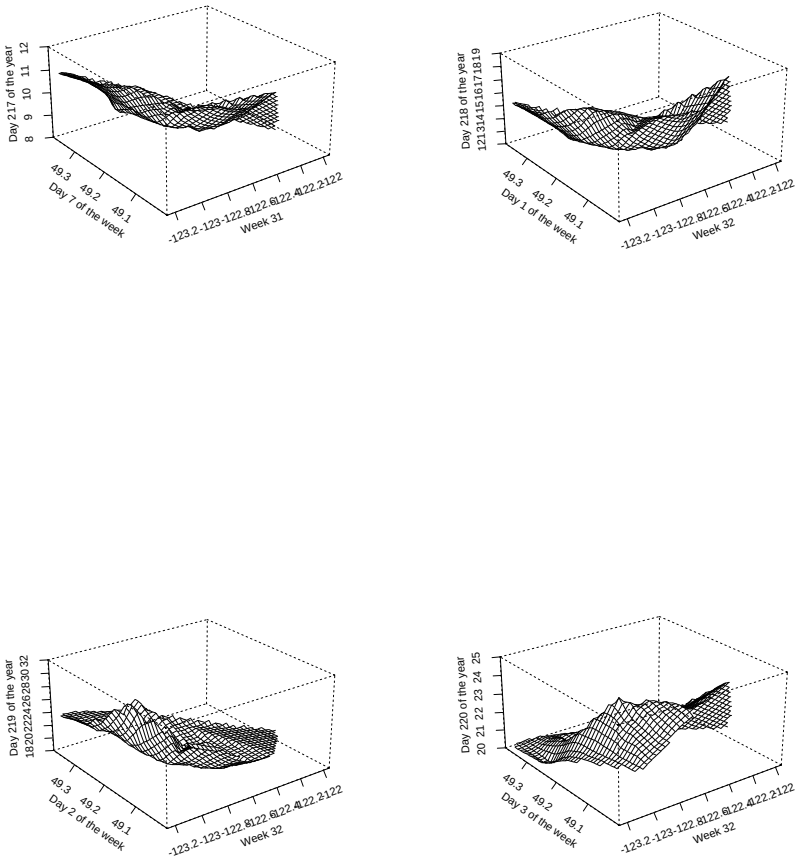



FIGURE 3. Interpolated $PM_{10}$ Field For Selected Summer Days.

spline) mapping $f_k$ of the geographic co-ordinates associated with the sites, $\mathbf{g}_i = (g_{i1}, g_{i2})$. The degree of smoothness and fit depend on the so-called "smoothing parameter". The $\{\mathbf{d}_i\}$ are then replaced by the fits $\{\mathbf{f}(\mathbf{g}_i)\}$,

where $\mathbf{f} = (f_1, f_2)$.
Selecting a large value of that parameter will result in a poorer fits between the $\{\mathbf{f}(\mathbf{g}_i)\}$ and the $\{\mathbf{d}_i\}$'s that would yield the best fits for the empirically estimated $\{\Psi_{ij}\}$ However, these smooth (splines) will more faithfully maintain the character of the geography and generally make the fitted covariance surface more readily interpretable. Selecting small values of the smoothing parameter can in fact lead to splines that twist the G-plane into unrecognizable form while ensuring a good fit to the estimated D-plane co-ordinates. Ultimately that parameter must be selected as a judicious compromise between the just-described competing objectives.
Once $\mathbf{f}$ has been specified, the required extension of the hypercovariance $\Psi_g$ estimated for the gauged sites must be extended to the full hypercovariance $\Psi$ for all sites. This step is readily completed by representing the g co-ordinates of ungauged sites i and j by $\mathbf{d}_i = \mathbf{f}(\mathbf{g}_i)$ and $\mathbf{d}_j = \mathbf{f}(\mathbf{g}_j)$. Then $\Psi_{ij}$ may be estimated by $\zeta(||\mathbf{d}_i - \mathbf{d}_j||)$.
In practice the SG method is implemented through the so-called "variogram" in exactly the same way as described above for the covariance. To illustrate this process, for any pair (i,j) of the 10 Vancouver $PM_{10}$ monitored sites. first estimate the 45 $\{\Psi_{ij}\}$ or rather the corresponding variograms in the obvious way. For a fixed level of smoothness, a *zeta* and smooth g co-ordinates mappings can now be found to obtain a good fit against the 45 estimated $\{\Psi_{ij}\}$'s.
If no smoothing is used one can see that plot in the right hand panel of Figure 2. The scatter plot shows 45 plotted variogram estimates and the best fitting variogram plotted against the fitted d co-ordinate distances.
That picture shows the surface must essentially be heavily distorted to achieve isotropy. Its interpretation is difficult. So an alternative is offered in Figure 3. There with a small amount of smoothing a flatter surface results.
After developing and applying the spatial predictive distribution, the ED ambient levels of $\mathrm{PM}_{10}$ can be convolved with the conditional predictive distribution determined by pCNEM/ $\mathrm{PM}_{10}$ to yield a predictive distribution for the exposure of a random individual conditional on the measurements at the 10 ambient $\mathrm{PM}_{10}$ monitors.
To conclude this section, we turn to the 3rd component of the health impact assessment model. There we model the level of adverse health outcomes as a function of unmeasured exposure. The model is developed for applications like that contemplated here by Zidek et al (1998a). Its use has been demonstrated by Zidek et al (1998b). The model takes an approach that has become quite standard in environmental epidemiology, the so-called "time-series" approach. It assumes a random response $\{Y_{kt}\}$ for a set of clusters k and times t. In our application these will be daily counts for residents of ED (census subdivisions) k on days t. In our as yet incomplete application of out model, the particular outcomes remain to be determined but could include such things as mortality attributed to cardio-vascular

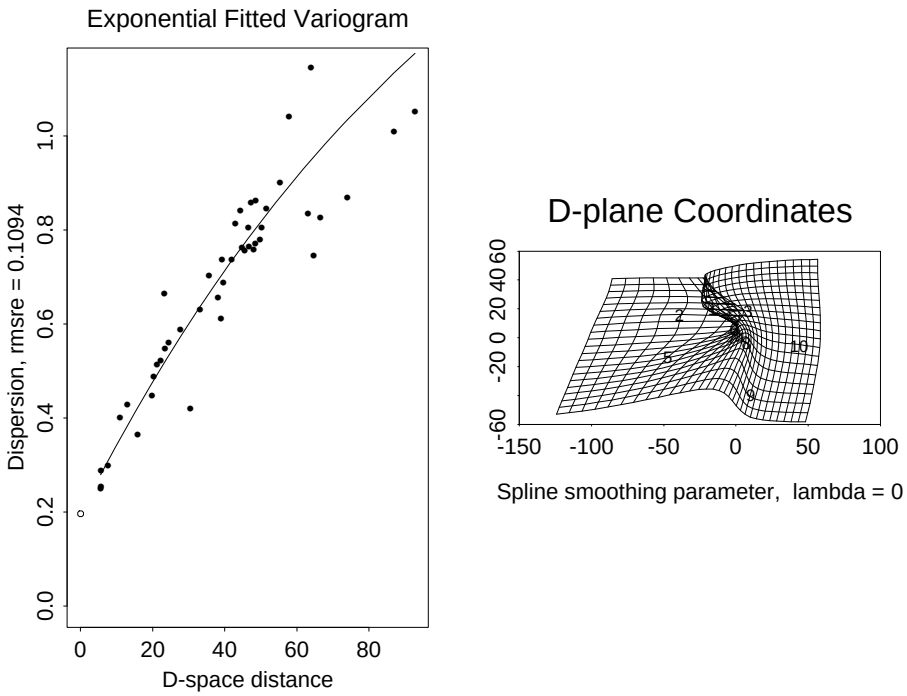



FIGURE 4. The Variogram Fit With No Smoothing.

causes or morbidity as measured by hospital admissions.
$Y_{kt}$'s distribution is assumed to be conditional on a random vector of predictors $X_{kt}$ including unmeasured exposure to $PM_{10}$ of ED k residents. (The latter would be obtained by averaging with appropriate weights based on demographic data of the predicted hourly level of exposure for a randomly selected individual in that ED.) However the X's would also include such things as meteorological variables to eliminate potential confounding. As well, the model will include random effects for cluster k.
Inference is simplified in the theory of Zidek et al (1998) since only the first and second moments of the $X_{kt}$ need to be specified. That simplicity is gained by adapting and using the GEE approach of Zeger et al (1988) as extended by Burnett and Krewski (1994). That approach gives a basis for inference about the effects of the hazards reflected in the X's. Results from our study will be presented in a future report.

## References

Brown, P.J., Le, N.D., Zidek, J.V. (1994). Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics* 22, 489-509.

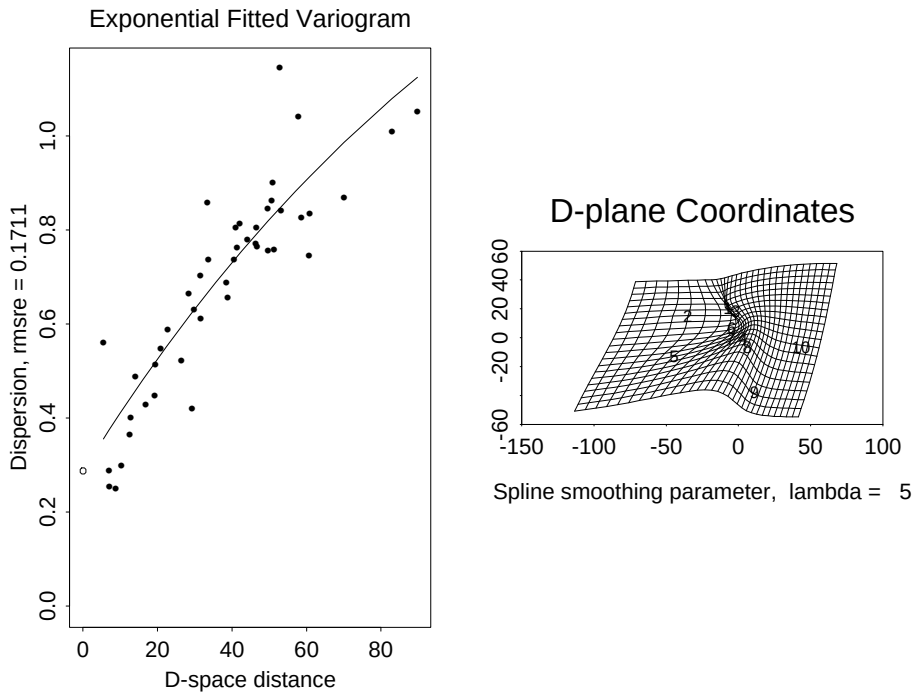



FIGURE 5. The Variogram Fit With Moderate Smoothing.

Burnett, R. and Krewski, D. (1994). Air pollution effects on hospital admission rates: a random effects modelling approach, *Can Jour Statist*, 22, 441-458.

Carroll R.J., Ruppert D., Stefanski L.A. (1995)*Measurement error in nonlinear models*. London: Chapman and Hall.

Johnston, T.R., Capel, J.E., Wijnberg and McCurdy, T. (1992). Estimation of Carbon Monoxide Exposures and Associated Carboxyhemoglobin Levels in Denver Residents Using A Probabilistic Version of NEM. Presentation at the 85th Annual Meeting of the Air and Waste Management Association.

Johnston, T.R. (1984). A study of personal exposure to carbon monoxide in Denver, Colorado. EPA-600/54-84-014, US Environmental Protection Agency, Research Triangle Park.

Johnston, T.R. (1987). A study of human activity patterns in Cincinnati, Ohio. Electric Power Research Institute, Palo Alto.

Le, N.D. and Zidek, J.V. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *Journal of Multivariate Analysis*, 43, 351-74.

Le, N.D. Sun, W. and Zidek, J.V. (1997). Bayesian Multivariate Spatial Interpolation with Data Missing-by-Design. *Jour Roy Statist Soc, Series B*, 59, 501-510..

Le N.D., Sun L. and Zidek J.V. (1999). Bayesian Spatial Interpolation and Backcasting Using Gaussian-Generalized Inverted Wishart Model. Department of Statistics. Technical Report #185. University of British Columbia.

MacIntosh D.L., Xue, J., Ozkaynak, H., Spengler, J.D. and Ryan, P.B. (1995). A population-based exposure model for benzene. Submitted.

Pierce, D.A., Stram, D.O., Vaeth, M. and Schafer, D.W. (1992), The errors-in-variables problem: considerations provided by radiation dose-response analyses of the a-bomb survivor data. *Jour Amer Statist Assoc*, 87, 351-359.

Sun, Li, Zidek, J.V., Le, N.D. and Özkaynak, H. (2000). Interpolating Vancouver's Daily Ambient PM10 Field. *Environmetrics*.. To appear.

Sun, W. (1994). Bayesian multivariate interpolation with missing data and its applications. Ph.D. Thesis, University of British Columbia.

Sun W., Le N.D., Zidek J.V. and Burnett R. (1998). Assessment of a Bayesian Multivariate Interpolation Approach for Health Impact Studies. *Environmetrics*, 9, 565-586.

Sun, W. (1998). Comparison of a CoKriging Method With a Bayesian Alternative. *Environmetrics*, 9, 445-457.

Sun, W., Le, N.D., Zidek, J.V. and Burnett, R. (1998) Assessment of a Bayesian multivariate spatial interpolation approach for health impact studies. *Environmetrics*, 9, 565-586.

Sun, W. (1998). Comparison of a CoKriging Method With a Bayesian Alternative. *Environmetrics*, 9, 445-457.

Sun, W. (1998). Comparison of a CoKriging Method With a Bayesian Alternative. *Environmetrics*, 9, 445-457.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data; a generalised estimating equation approach. *Biometrics*, 44 1049-1060.

Zidek J.V., Wong H., Le N.D. and Burnett R. (1996). Causality, measurement error and multicollinearity in epidemiology. *Environmetrics*, 7, 441-451.

Zidek J.V., Le N.D., Wong H. and Burnett R.T. (1998a) Including structural measurement errors in the nonlinear regression analysis of clustered data. *Can J Statist*, 26, 537-548.

Zidek, J.V., White, R., Le, N.D., Sun, W. and Burnett, R.T. (1998b). Imputing Unmeasured Explanatory Variables in Environmental Epidemiology With Application To Health Impact Analysis of Air Pollution. *Environmental and Ecological Statistics*, 5, 99-115.

# Aids for Modeling the Covariance Structure of Longitudinal Data: Alternative Specifications and Graphical Diagnostics

Dale L. Zimmerman[1]

[1] Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242, U.S.A.

**Abstract:** Parametric covariance modeling is an important aspect of a mixed linear models approach to the analysis of longitudinal data. Compared to choosing an appropriate model for the mean structure, however, the choice of a covariance structure too often is made somewhat arbitrarily, without a full understanding of what the structure implies about the variances, correlations, and partial correlations of the observations and without the benefit of graphical diagnostics. This article examines the behavior of the variances, correlations, and partial correlations of several structures and presents two graphical diagnostics useful for choosing plausible covariance structures to fit to the data.

**Keywords:** Antedependence; Mixed Model; Partial Correlations; Random Coefficients; Scatterplot Matrix.

## 1 Introduction

This article considers the modeling of continuous longitudinal data, i.e. repeated measurements of a continuous response variable taken over time on each of $m$ subjects. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$ be the response vector of $n_i$ measurements on the $i$th subject and let $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$ be the corresponding vector of measurement times. Suppose that we also observe a $p$-vector of covariates, $\mathbf{x}_{ij}$, associated with $y_{ij}$, and put $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$. In the past 20 years or so, an approach for analyzing these data that is based on mixed linear models has become quite popular and has been implemented in major statistical software packages, e.g. PROC MIXED of SAS (SAS Institute Inc. 1996). The model is given by

$$\mathbf{y}_i \sim \text{independent } N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i(\boldsymbol{\theta})), \tag{1}$$

where $\boldsymbol{\beta}$ is a $p$-vector of fixed, unknown, and typically unrestricted parameters; $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ is a $n_i \times n_i$ covariance matrix; and $\boldsymbol{\theta}$ is a $q$-vector of unknown parameters, restricted to a parameter space $\boldsymbol{\Theta}$ that is either the set of all $\boldsymbol{\theta}$-vectors for which all $\boldsymbol{\Sigma}_i$ are positive definite or some subset of that set. Note that the $\boldsymbol{\Sigma}_i$'s are assumed to be of the same basic form, i.e. the same

parametric covariance function determines the elements of each, but generally they may not be equal nor of the same dimensions; an exception occurs when measurement times are common across subjects.

Estimation of the parameters of model (1) is typically carried out by the methods of maximum likelihood or residual maximum likelihood. Application of these methods to a particular longitudinal data set requires the analyst to specify the model's mean structure and covariance structure. Often, the study's objectives largely dictate the choice of mean structure; for example, if the goal of the study is to see how the growth of animals is affected over time by different treatments, then the initial specification of mean structure would probably include treatment effects, one or more time effects (e.g. linear and quadratic terms), and effects for time-by-treatment interaction. Graphical techniques, such as a plot of subject profiles over time or a residual plot, are routinely used to inform or confirm this choice. However, an appropriate covariance structure is often more difficult to specify than a mean structure. Typically, the study's objectives do not dictate this choice; nevertheless, the choice is important because an appropriately parsimonious choice of covariance structure can substantially improve the efficiency of inferences made about the mean structure and provide better estimates of standard errors of estimated mean structure parameters (Diggle, Liang, and Zeger, 1994). The problem is not a shortage of possible structures. PROC MIXED, for example, allows the user to choose from no less than twenty distinct covariance structures. Rather, the problem is to determine which structures, among the many possibilities, should be fit to the data and, subsequently, to assess whether the structure judged to fit best (according, typically, to some numerical goodness-of-fit criterion such as AIC or BIC) actually provides a reasonable fit. Arbitrarily choosing some models to fit can be inefficient and increases the risk that the model selected as "best-fitting" really does not fit the data very well. A better, more informed approach would be to choose a set of models to fit on the basis of first examining and comparing, through graphical techniques and otherwise, the behavior of the sample variances, correlations, and partial correlations to the behavior of these quantities under each of a large number of assumed covariance structures.

This article examines the behavior of the variances, correlations, and partial correlations of several structures and presents two graphical diagnostics useful for choosing plausible covariance structures to fit to the data.

## 2 Alternative Specifications of Structure

A covariance structure can be specified and/or parameterized in several ways. Most structures have an *equation specification*, in which $y_{it}$ is expressed as the sum of a fixed component and one or more random components. The random components may include values of the response at earlier

times. An equation specification has the advantage of interpretability (by indicating how data with this structure could arise) and, for this reason, also provides the simplest and least computationally intensive method for simulating data with the structure. For diagnostic purposes, however, two other specifications may be more useful: a *variance-correlation specification*, which simply characterizes the structure in terms of the variances of, and correlations among, the observations; and an *partial variance-correlation specification*, which characterizes the structure in terms of the partial variances of, and partial correlations among, the observations conditional on all other observations.

To provide scope for the subsequent discussion, here we present a sampling of covariance structures, giving as many of these three specifications as are reasonably concise and illuminating. For brevity we do not attempt to provide an exhaustive survey; for additional structures see Jennrich and Schluchter (1986), Wolfinger (1996), and Núñez-Antón and Zimmerman (2000), for example. For these representations we assume that measurements are taken at equally-spaced time points $t_1 = 1, \ldots, t_n = n$. For the equation specification, we let $y_t$ be a generic subject's response at time $t$, $\mu_t = E(y_t)$, and $\{\epsilon_t : t = 1, \ldots, n\}$ be independent normal random variables with zero means and variances $\sigma_t^2$. For the variance-correlation specification, we put $\sigma_{tt} \equiv \mathrm{var}(y_t)$, and $\rho_{tu} \equiv \mathrm{corr}(y_t, y_u)$ for $t > u$. Note that $\sigma_t^2$ and $\sigma_{tt}$ represent different quantities. In the case of the partial variance-correlation specification, for reasons that will become clear subsequently we present only the partial correlations given the observations at intervening times, for which we write $\rho_{tu \cdot t+1,\ldots,u-1} \equiv \mathrm{corr}(y_t | y_{t+1}, \ldots, y_{u-1}, y_u | y_{t+1}, \ldots, y_{u-1})$ for $u - t \geq 2$.

*First-order Autoregressive, AR(1)*

$$\begin{aligned}
&y_1 = \mu_1 + \epsilon_1, \quad y_t = \mu_t + \rho(y_{t-1} - \mu_{t-1}) + \epsilon_t \qquad (t = 2, \ldots, n)\\
&\sigma_1^2 = \sigma^2 > 0, \quad \sigma_t^2 = \sigma^2(1 - \rho^2) \qquad (t = 2, \ldots, n), \qquad |\rho| < 1\\
&\sigma_{tt} = \sigma^2, \qquad \rho_{tu} = \rho^{u-t}, \qquad \rho_{tu \cdot t+1,\ldots,u-1} = 0 \;\forall\, t, u
\end{aligned}$$

Thus, variances and same-lag correlations are constant; correlations are given by an exponentially decreasing (in modulus) function of elapsed time; and partial correlations adjusted for intervenors are zero.

*Wiener, WI*

$$\begin{aligned}
&y_t = \mu_t + \sum_{j=1}^{t} \epsilon_j, \qquad \sigma_t^2 = \sigma^2 > 0 \;\forall\, t\\
&\sigma_{tt} = t\sigma^2, \qquad \rho_{tu} = (t/u)^{1/2}, \qquad \rho_{tu \cdot t+1,\ldots,u-1} = 0 \;\forall\, t, u
\end{aligned}$$

Thus, variances increase linearly; correlations are a decreasing function function of elapsed time; same-lag correlations increase over time; and partial correlations adjusted for intervenors are zero.

*First-order Antedependence, AD(1)*

$$y_1 = \mu_1 + \epsilon_1, \quad y_t = \mu_t + \phi_t(y_{t-1} - \mu_{t-1}) + \epsilon_t \quad (t = 2, \ldots, n)$$
$$\phi_t\text{'s arbitrary}, \quad \sigma_t^2 > 0 \; \forall \, t$$
$$\sigma_{11} = \sigma_1^2, \quad \sigma_{tt} = \phi_t^2 \sigma_{t-1,t-1} + \sigma_t^2 \quad (t = 2, \ldots, n)$$
$$\rho_{tu} = \begin{cases} \phi_{t+1}(\sigma_{tt}/\sigma_{t+1,t+1})^{1/2} & \text{if } u - t = 1 \\ \prod_{j=t}^{u-1} \rho_{j,j+1} & \text{otherwise} \end{cases} \qquad \rho_{tu \cdot t+1,\ldots,u-1} = 0 \; \forall \, t, u$$

Thus variances and lag-one correlations are essentially arbitrary (meaning that they are arbitrary subject to the restriction that the covariance matrix be positive definite); higher-lag correlations are functions of the lag-one correlations and decrease (in modulus) monotonically as elapsed time increases; and partial correlations adjusted for intervenors are zero.

*kth-order Antedependence, AD(k)*

$$y_1 = \mu_1 + \epsilon_1, \quad y_t = \mu_t + \sum_{l=1}^{\min(k,t-1)} \phi_{t,t-l}(y_{t-l} - \mu_{t-l}) + \epsilon_t \quad (t = 2, \ldots, n)$$
$$\phi_{t,t-l}\text{'s arbitrary}, \quad \sigma_t^2 > 0 \; \forall \, t$$
$$\rho_{tu \cdot t+1,\ldots,u-1} = 0 \text{ if } u - t > k + 1$$

Though it is not transparent from these representations, the variances are essentially arbitrary as are the same-lag correlations up to and including lag $k$; higher-lag correlations are complicated functions of those at lower lags; and partial correlations adjusted for at least $k$ intervenors are zero.

*Variable-order Antedependence, VAD($k_1, \ldots, k_n$)*

$$y_1 = \mu_1 + \epsilon_1, \quad y_t = \mu_t + \sum_{l=1}^{\min(k_t,t-1)} \phi_{t,t-l}(y_{t-l} - \mu_{t-l}) + \epsilon_t \quad (t = 2, \ldots, n)$$
$$\phi_{t,t-l}\text{'s arbitrary}, \quad \sigma_t^2 > 0 \; \forall \, t$$
$$\rho_{tu \cdot t+1,\ldots,u-1} = 0 \text{ if } u - t > k_t + 1$$

The variances are essentially arbitrary as are the same-lag correlations up to and including lag $k_t$; higher-lag correlations are complicated functions of those at lower lags; and partial correlations adjusted for at least $k_t$ intervenors are zero.

*qth-order Moving Average, MA(q)*

$$y_t = \mu_t + \sum_{j=0}^{q} \alpha_j \epsilon_{t-j}, \qquad \alpha_0 = 1, \;\; \alpha_1, \ldots, \alpha_q \text{ arbitrary}$$

$$\sigma_t^2 = \sigma^2 > 0, \qquad \sigma_{tt} = \sigma^2 \sum_{j=0}^{q} \alpha_j^2 \;\forall\, t$$

$$\rho_{tu} = \begin{cases} \frac{\alpha_{u-t} + \alpha_1\alpha_{u-t+1} + \alpha_2\alpha_{u-t+2} + \cdots + \alpha_{q-u+t}\alpha_q}{1+\alpha_1^2+\cdots+\alpha_q^2} & \text{if } u-t = 1, \ldots, q \\ 0 & \text{if } u-t \geq q+1 \end{cases}$$

Thus, variances are constant and correlations vanish beyond a finite, constant elapsed time. The partial correlations adjusted for intervenors are nonzero in general.

*Compound Symmetry, CS*

$$y_t = \mu_t + b + \epsilon_t, \qquad b \sim N(0, \sigma_b^2), \qquad \sigma_t^2 = \sigma^2 > 0 \qquad (t = 1, \ldots, n)$$
$$\sigma_b^2 \geq 0, \qquad \sigma_{tt} = \sigma_b^2 + \sigma^2$$
$$\rho_{tu} = \sigma_b^2/(\sigma_b^2 + \sigma^2), \qquad \rho_{tu\cdot t+1,\ldots,u-1} = \sigma_b^2/[(u-t)\sigma_b^2 + \sigma^2].$$

Thus, variances and all correlations are constant. Partial correlations adjusted for intervenors are given by a decreasing function of the number of intervenors.

*Linear Random Coefficients, LRC*

$$y_t = \mu_t + b_0 + b_1 t + \epsilon_t, \qquad \sigma_t^2 = \sigma^2 > 0 \qquad (t = 1, \ldots, n)$$
$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \sim N\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{00} & \sigma_{01} \\ \sigma_{01} & \sigma_{11} \end{pmatrix} \right), \qquad \sigma_{00} \geq 0,\; \sigma_{11} \geq 0,\; \sigma_{00}\sigma_{11} - \sigma_{01}^2 \geq 0$$
$$\sigma_{tt} = \sigma_{00} + 2t\sigma_{01} + t^2\sigma_{11} + \sigma^2$$
$$\rho_{tu} = \frac{\sigma_{00} + \sigma_{01}(t+u) + \sigma_{11}tu}{\sqrt{\sigma^2 + \sigma_{00} + 2\sigma_{01}t + \sigma_{11}t^2}\sqrt{\sigma^2 + \sigma_{00} + 2\sigma_{01}u + \sigma_{11}u^2}}$$

Thus, the variances are given by a concave-up function of time; specifically, the variances decrease if $t < -\sigma_{01}/\sigma_{11}$ and increase if $t > -\sigma_{01}/\sigma_{11}$ (assuming that $\sigma_{11} > 0$). The correlations can behave in a variety of ways, depending on the sign of $\sigma_{01}$ and the relative magnitudes of the covariance parameters. Note, however, that their behavior is tied to that of the variances. For example, the correlations are constant if and only if the variances are constant. Expressions for the partial correlations adjusted for intervenors can be given but are not illuminating.

Which of these specifications is most useful for which purposes? We have already noted the interpretability and data-generating capability of the equation specification. Another relative advantage of an equation specification for some structures is that fewer parameter constraints are required

to ensure positive definiteness. For example, the autoregressive parameters $\{\phi_{t,t-l}\}$ of the AD($k$) equation specification are unconstrained, something which is not the case for the parameters of the other two specifications.
As noted previously, however, for diagnostic purposes the other two specifications may be more useful. For example, plots of the sample variances against time and/or the sample correlations against elapsed time are easy to interpret and can be used to evaluate the plausibility of various structures. We take this idea further in the next sections, where we propose plotting the sample correlations against the fitted correlations assuming a particular structure. We also introduce a graphical diagnostic that permits an effective diagnosis of some structures from the behavior of the sample partial correlations adjusted for intervening observations.

## 3 Graphical Diagnostics

The variance-correlation specification defines a covariance structure in terms of its variances and correlations. Many diagnostics can convey useful information about the variances' behavior over time; a series of boxplots of the marginal distributions of responses at time $t_i$, plotted against $i$, is an example. Consequently we focus on diagnostics for the correlation structure only. Diggle et al. (1994) and Dawson, Gennings, and Carter (1997) review the few existing methods for diagnosing correlation structure. The most widely used method is the ordinary scatterplot matrix (OSM), which is a two-dimensional array of pairwise scatterplots of standardized responses (or of certain derived quantities such as residuals). The OSM is a graphical equivalent of the sample correlation matrix.
Although the OSM has a useful role in diagnosing correlation structure, alternative procedures can be complementary or even more informative in some cases. Here, I introduce two new graphical diagnostics called the correlation plot and the Partial Regression-on-Intervenors Scatterplot Matrix (PRISM), The acronym PRISM has some semantic substance: as a prism separates visible light into its components, so a PRISM can separate the dependence structure among within-subject responses into components which are easier to understand.
The correlation plot is simply a scatterplot of the sample correlations against the fitted correlations implied by a candidate structure. A perfect match of the sample correlations to the fitted correlations would manifest itself as a straight line from the point (0,0) to (1,1). To illustrate, we consider fitting the RCL and AD(1) structures to the speech recognition data described by Núñez-Antón and Zimmerman (2000). These data were taken at four measurement times, so each plot contains only six points. Figure 1 displays the correlation plots corresponding to the RCL and AD(1) structures. The AD(1) appears to be best, but both structures match the sample correlations reasonably well. On this basis I would recommend fitting both

to the data. Correlation plots corresponding to some other structures (not shown) do not fit nearly as well as these.

Two comments need to be made about correlation plots. First, the more complicated the structure, the better the plot will tend to appear. An extreme example is a completely unstructured covariance matrix, for which the correlation plot will be a straight line. Thus, decisions about which models to fit should keep model parsimony in mind as well as the extent of the match between sample and model correlations. Second, there are some important differences among structures in how the model correlations can be obtained. For some structures, such as WI, the model correlations are completely determined. For others, some correlations may depend on other correlations, in which case I recommend first matching the "independent" correlations perfectly to the sample correlations and then computing the "dependent " correlations. This is what was done for the AD(1) correlation plot in Figure 1, and in this case yields the maximum likelihood estimators of the correlations (see Byrne and Arnold, 1983). For still other structures, the model may need to be fit to the data to yield estimated correlations under the model; the estimated correlations for the RCL correlation plot were the maximum likelihood estimates.

Before constructing the PRISM it is advisable to standardize the response variables to prevent possible differences in the variances of responses over time from obfuscating the correlation structure. For simplicity of presentation, assume that there are no treatments or other observed covariates, and that measurement times $t_1, t_2, \ldots, t_n$ (not necessarily equally-spaced) are common across subjects, save possibly for a relatively small proportion of missing data. Let $\bar{y}_{\cdot j}$ and $s_{\cdot j}$ denote the sample mean and sample standard deviation, respectively, of the nonmissing responses at time $j$. Define the standardized responses

$$z_{ij} = \frac{y_{ij} - \bar{y}_{\cdot j}}{s_{\cdot j}}$$

for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Dawson et al. (1997) show that the sample correlation structure of the standardized observations is identical to that of the original observations; the same type of identity holds for the sample partial correlation structure as well.

The PRISM is constructed as the upper right "wedge" of a rectangular array of certain partial regression (or added variable) plots. The main diagonal of plots are ordinary scatterplots of $z_j$ (the collection of nonmissing standardized responses at time $j$) against $z_{j+1}$ ($j = 1, \ldots, n-1$). The second diagonal of plots are partial regression plots of standardized responses lagged two times apart, adjusted for the standardized response at the intervening time. The third diagonal of plots are partial regression plots of standardized responses lagged three times apart, adjusted for the standardized responses at the two intervening times. In general, the plot in row $j$ and column $k$ ($k \geq j$) is the partial regression plot of standardized response

variables $z_j$ and $z_{k+1}$ adjusted for standardized responses at the intervening times $t_{j+1}, t_{j+2}, \ldots, t_k$. Thus, the PRISM is the graphical equivalent of a matrix of certain partial correlations; specifically, the $(j,k)$th plot in the array displays points whose ordinary correlation is the partial correlation between $z_j$ and $z_{k+1}$ adjusted for all standardized responses at intervening times $t_{j+1}, t_{j+2}, \ldots, t_k$. Random scatter in the $(j,k)$th plot indicates that $z_j$ and $z_{k+1}$ are conditionally independent, given the intervening responses; departures from random scatter indicate conditional dependence.

The PRISM is particularly effective at identifying Markovian correlation structures, i.e. those structures for which observations taken sufficiently far apart in time are independent conditional on observations at intervening times. To illustrate, we consider simulated longitudinal data having (a) an AD(1) covariance structure, and (b) a variable-order AD structure, VAD(0,1,1,1,2,1). Note that the AD(1) is also a VAD(0,1,1,1,1,1). In both cases, the data are a random sample of size $m = 100$ from a six-dimensional normal distribution with mean vector $\mathbf{0}$, scaled to have unit variances.

It is very difficult to distinguish these two structures on the basis of their OSMs, which are not shown. Both OSMs reveal a persistence of correlation as one moves away from the main diagonal and an increase in correlation strength as one moves down any particular diagonal. However, the PRISMs (Figures 2 and 3) clearly distinguish the structures, for the plot in row 3, column 4 exhibits random scatter in the case of the AD(1) but in the case of the VAD. Of course, this difference is due to the latter's nonzero value of $\mathrm{corr}(z_3|z_4, z_5|z_4)$.

These two examples clearly demonstrate the ability of the PRISM to identify and distinguish between autoregressive and other Markovian structures. For non-Markovian dependence, however, the OSM may be more informative than the PRISM. For example, MA(1) and CS structures are more easily diagnosed from the OSM than from the PRISM. Neither the OSM nor the PRISM is particularly adept at identifying random coefficients models such as LRC.

The superiority of the PRISM for diagnosing Markovian structures and the superiority of the OSM for diagnosing some other structures suggest that a better diagnosis of correlation structure can occur when the OSM and PRISM are used in tandem than when only one is used. Using the PRISM in addition to the OSM for specifying the correlation structure of longitudinal data would parallel a similar practice in time series analysis. There, good statistical practice includes an examination of both the sample autocovariance function (ACF) and sample partial autocovariance function (PACF). In fact, in the case of equally-spaced longitudinal data from a stationary process, the superposition of plots of the OSM along diagonals yields $n-1$ scatterplots whose correlations coincide with those given by the ACF, and similarly superimposing plots of the PRISM yields $n-1$ scatterplots whose correlations coincide with those given by the PACF.

## 4 Conclusions

The thesis of this article is that analysts of longitudinal data should examine the sample variances, correlations, and partial correlations of their data and compare them to what is implied of the variances, correlations, and partial correlations of a covariance structure under consideration for fitting to the data. Two graphical diagnostics, the correlation plot and the PRISM, were introduced that should facilitate these comparisons. Through the use of these diagnostics, some inappropriate models for the covariance structure may be eliminated from consideration, thereby reducing the number the analyst needs to fit. Moreover, they may permit an assessment of whether the model(s) that minimize a numerical goodness-of-fit criterion actually provide a good fit.

### References

Byrne, P.J. and Arnold, S.F. (1983). Inference about multivariate means for a nonstationary autoregressive model. *Journal of the American Statistical Association*, **78**, 850-855.

Dawson, K.S., Gennings, C., and Carter, W.H. (1997). Two graphical techniques useful in detecting correlation structure in repeated measures data. *The American Statistician*, **51**, 275-283.

Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Jennrich, R.L. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**, 805-820.

Núñez-Antón, V. and Zimmerman, D.L. (2000). Modeling nonstationary longitudinal data. *Biometrics*, **56**, 93-99.

SAS Institute Inc. (1996). *SAS/STAT Software: Changes and Enhancements through Release 6.11*, Cary, NC: SAS Institute Inc.

Wolfinger, R.D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Health*, **1(2)**, 205-230.

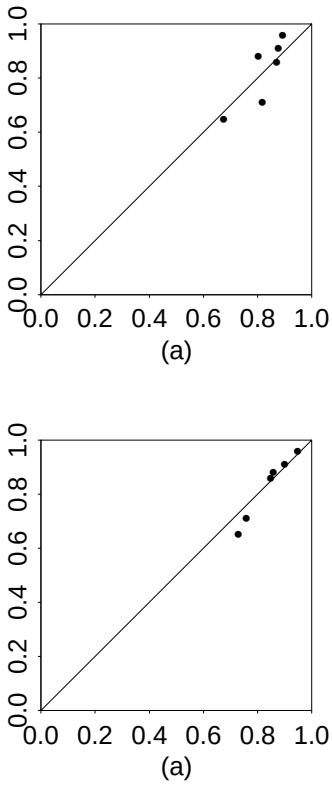



FIGURE 1. Correlation plots for the speech recognition data: (a) RCL, (b) AD(1).

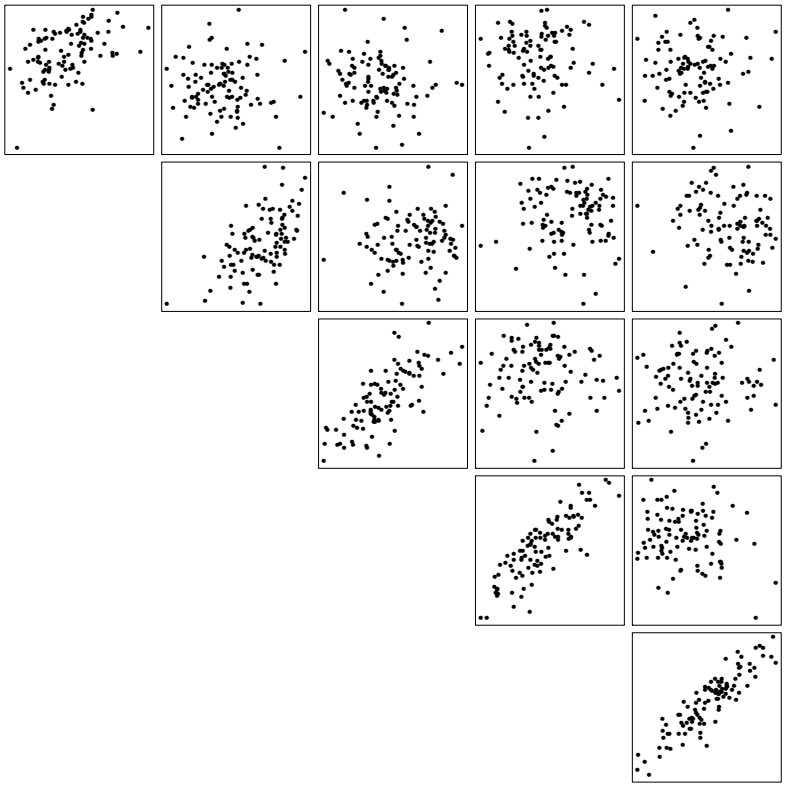


FIGURE 2. PRISM for simulated data having an AD(1) covariance structure with $(\phi_2, \phi_3, \phi_4, \phi_5, \phi_6) = (0.5, 0.6, 0.7, 0.8, 0.9)$, $\sigma_1^2 = 1$, $\sigma_t^2 = 1 - \phi_t^2$ $(t = 2, \ldots, 6)$.

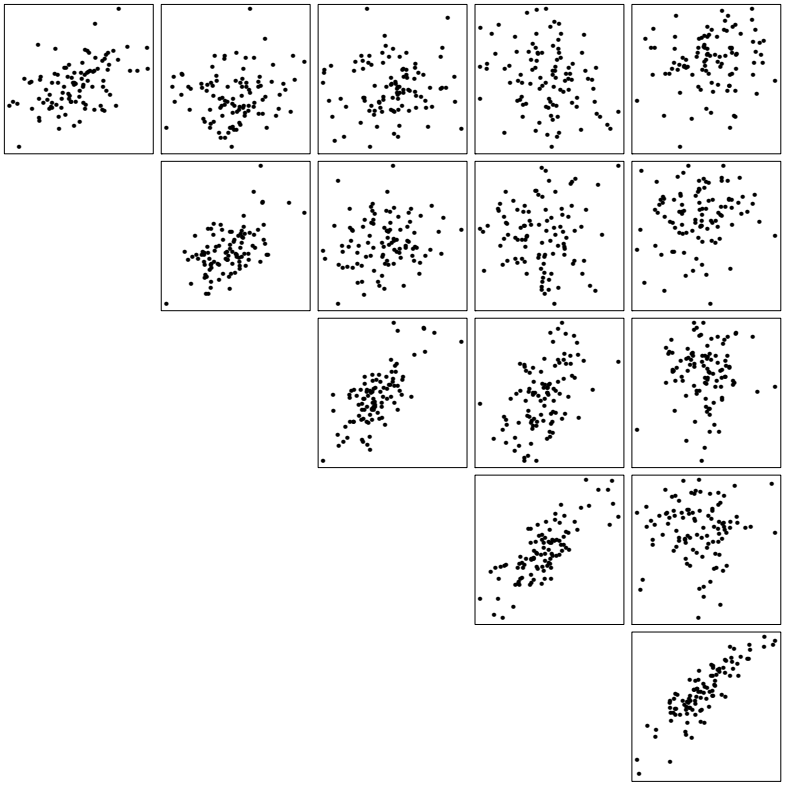


FIGURE 3. PRISM for simulated data having a VAD(0,1,1,1,2,1) covariance structure with $\phi_{21} = 0.5, \phi_{32} = 0.6, \phi_{43} = 0.7, \phi_{54} = 0.8, \phi_{53} = 0.7, \phi_{65} = 0.9$, and $\sigma_{tt} = 1$ $(t = 1, \ldots, 6)$.

# Drought analysis based on a compound Poisson model

Jesús Abaurrea[1], Ana Carmen Cebrián [1]

[1] Dpto. Métodos Estadísticos, Universidad de Zaragoza,
Ed. Matemáticas. Pedro Cerbuna, 12.
Zaragoza 50009
E-mail adress: acebrian@posta.unizar.es

**Abstract:** This paper develops a stochastic model for drought events. A Poisson cluster process is used to represent their occurrence and a vector series of random variables, composed by duration, deficit and maximum intensity, to describe their magnitude. The signal used for defining drought is monthly moving annual rainfall. Useful parameters for the design and planning of water resource systems can be derived from the model.

**Keywords:** Drought analysis, Poisson cluster process, threshold methods, extreme value theory.

## 1 Introduction

The aim of this paper is to develop a stochastic model to analyze the occurrence and severity of meteorological drought.
Nowadays, most water resource planners rely on mathematical indexes to decide when to start implementing measures in response to drought. Description of drought characteristics by probability distributions provides measures such as inter-drought recurrence time, expected duration or mean deficit, which are helpful in water resource management. The characterization of the largest drought event to occur in a given period of time also requires to be based on a drought model such as the one proposed.
Herein, only long-term meteorological drought has been analyzed, but the methodology employed can be applied to study other aspects and definitions of the phenomenon. Six Spanish monthly rainfall series of about one hundred years long have been studied to check the model. All the analyses have been programmed using S-Plus.

## 2 Drought definition

In general, drought can be defined as a deficiency of precipitation over an extended period of time resulting in a water shortage for some activities.

In practice, however, drought definition must reflect differences caused by climate, needs and disciplinary approach; so, no single definition works in all circumstances. Operational definitions that allow identifying the beginning, end and degree of severity of the drought are based on 'Excess over threshold' methods, where a stochastic process, $s(t)$, related to precipitation or some other variable that describes the hydric state of the system, is compared to a threshold, $u1(t)$, which represents a critical level for the process. A drought will occur when $s(t)$ is below $u1(t)$.
Our approach to defining drought is based on the decile method developed by Gibbs and Maher (1967). Drought threshold is defined by these authors as the tenth percentile, $p10$, of the precipitation series. In this work, we have used as signal $s(t)$ the monthly series of moving annual precipitation. Since this variable represents an annual amount, it will reflect water resource deficiencies in processes based on long-term precipitation, such as reservoir levels. Drought effects in other fields, agriculture for example, should be based on series with a shorter accumulation period. The monthly updating allows a frequent drought intensity evaluation, whatever accumulation period we use. As there is no seasonal component in the signal used, a constant critical level can be defined.
A similar approach in defining drought periods could be based on the Standardized Precipitation Index, SPI. This drought indicator was introduced by McKee et al. (1993) and it is often used in the USA. Since it is a standardized measure and its mean is always zero, we can define the same threshold for any location and accumulation period. The most frequent threshold is -1.5 but other values can be used. In particular, drought series obtained from SPI, using the tenth normal percentile as threshold, are virtually equal to the ones resulting from the decile method application.
In order to describe drought magnitude we have used three variables: duration or length, L, deficit, D, and maximum intensity, IM.

## 3 Modelling drought process

### 3.1 Compound Poisson process

Extreme value theory, Davison and Smith (1990), Embretchs et al. (1997), asserts that:

- The occurrence of high level excesses, in independent or short term dependent stationary stochastic processes, behaves asymptotically as a Poisson process, PP.
- The limit distribution of the excess amount is Generalized Pareto, GP.

Thus, after a preliminary examination of Poisson properties of dry period series defined with different threshold values, we tried to model the

drought process as a compound Poisson process $CPP(\lambda; F_1, F_2, F_3)$ where occurrence is modelled by a Poisson process and independent and equally distributed random vectors $(L, D, IM)$ are associated to each event. Since PP is a point process and every drought has a length, the ocurrence of each event has to be associated to an instant. Initial, half and maximum intensity instants are the most reasonable alternatives.
Checking of the model hypothesis in $p10$-series showed that Poisson properties of the occurrence were satisfied, but a lack of independence was detected in the corresponding magnitude vectors, specially in the deficit and maximum intensity series. This dependence is caused by the clustering of dry spells. During a prolonged dry period it is often observed that the hydrologic signal exceeds by a small amount the critical level for a short period of time, thus dividing a large drought into a number of minor dry spells. Such a cluster of dry spells should be treated as one drought as long as the short separation non-dry periods don't eliminate the dry period impact.

### 3.2 Compound cluster Poisson process

One simple model representing this structure is the Poisson cluster process, PClP. In this model, droughts -represented as clusters- occur according to a Poisson process and are formed by a random number of points -which in this case correspond to dry spells- that form a subsidiary process, figure 1. Although we checked that dry period process defined with the $p10$-threshold was a PP, the process resulting from a random grouping of points doesn't always conserve the Poisson character. The following property, that provides justification for the proposed model when sufficiently extreme thresholds are used, has been proved.

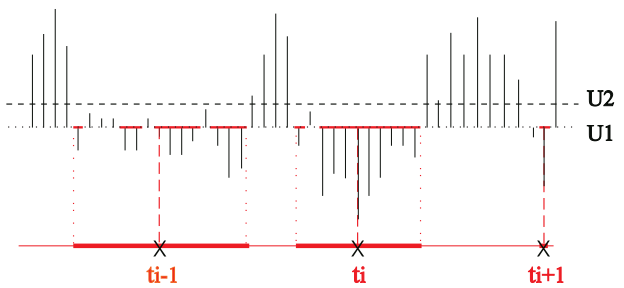



FIGURE 1. Cluster Poisson process diagram.

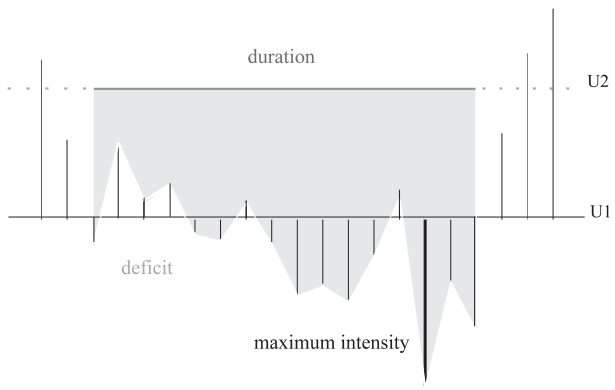



FIGURE 2. Vector of variables corresponding to a drought event.

**Property**: If a dry spell process defined with a threshold $u$ is a PP, the process resulting from clustering its points will converge, when the threshold value decreases, to a PP.

*Cluster definition*

One inconvenience of the model is how to determine the groups of dry spells which form each cluster. After trying out several criteria, the cluster composition has been fixed applying an empirical rule based on inter-event period duration and the degree of recovery of the system between dry spells. We have considered that two events belong to the same cluster if inter-event time is less or equal to six months and no intensity value in that period reaches the thirtieth percentile, $p30$, which represents a normal rainfall value according to the Gibbs and Maher classification.

*Final model*

The final model is a compound Poisson cluster process, $\mathrm{CPClP}(\lambda; F_1, F_2, F_3)$, which combines a PClP occurrence model with the above described vector of variables, shown in figure 2. Deficit is defined as the cumulative sum of differences to the normal rainfall value, threshold $u2 = p30$. This definition avoids problems relating to the influence in deficit of inter-event period excesses.
Model estimation is finished by fitting adequate probability distributions to the magnitude vector. Five positive, continuous distributions -Exponential, Weibull, Gamma, Lognormal and Generalized Pareto- are tested to model deficit and maximum intensity series. In the case of duration, two additional discrete distributions -Geometric and Negative Binomial- are also considered.

## 4 Results

In order to check the model, it was applied to six Spanish rainfall series. The results obtained in the examination of CPClP hypothesis are satisfactory for threshold values between $p10$ and $p6$ depending on the location:

- Burgos, Daroca and Huesca $p10$-processes and Murcia $p7$-process can be considered CPClP.
- Madrid $p6$-process fits also a CPClP but it needs a seasonal magnitude vector, as can be seen in the bubble plot, figure 3.
- San Fernando $p10$-process should be fitted by a non-homogeneous CPClP since both, occurrence process and magnitude vector, show seasonal behaviour, figure 4.

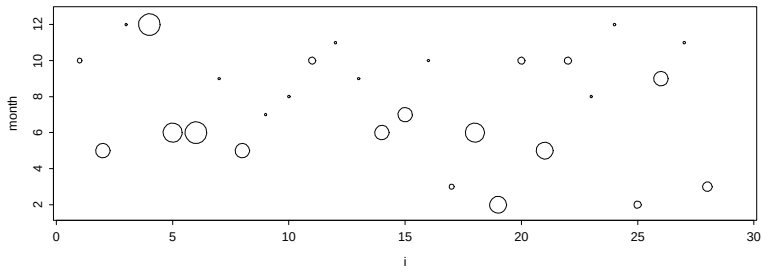



FIGURE 3. Duration bubble plot from Madrid.

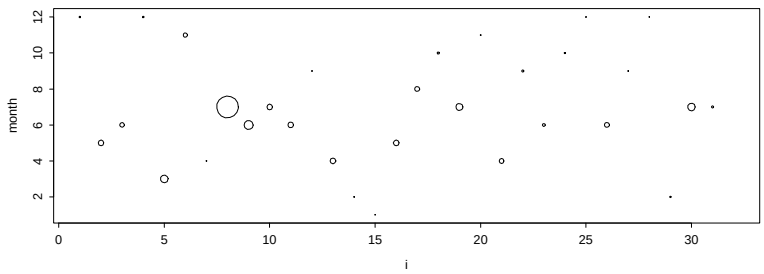



FIGURE 4. Duration bubble plot from San Fernando.

Finally, we summarize the results from the Huesca series fit; duration requires a no-memory distribution such as shifted Exponential or Geometric; deficit can be fitted by Exponential or Lognormal distributions, and maximum intensity by a Generalized Pareto. Some useful parameters for each variable are shown in table 1.

| | |
|---|---|
| Mean L | 7.79 months |
| St. dev. | 6.79 |
| $p50$<br>$p25 - p75$ | 5.7<br>3.0-10.4 |
| Return value | 50 years: 17.5<br>100 : 22.2<br>200 : 26.9 |

| | |
|---|---|
| Mean L | 7.79 months |
| St. dev. | 6.27 |
| $p50$<br>$p25 - p75$ | 5<br>2-9 |
| Return value | 50 years: 16<br>100 : 20<br>200 : 24 |

| | |
|---|---|
| Mean D | 907.6 l. |
| St. dev. | 1377.5 |
| $p50$<br>$p25 - p75$ | 499.3<br>238.9-1043.7 |
| Return value | 50 years: 2196.1<br>100 : 3229.0<br>200 : 4521.0 |

| | |
|---|---|
| Mean IM | 67.9 l. |
| St. dev. | 47.9 |
| $p50$<br>$p25 - p75$ | 59.7<br>27.3-101.8 |
| Return value | 50 years: 143.1<br>100 : 160.6<br>200 : 173.0 |

TABLE 1. Estimated parameters, Duration (L): Shifted Exponential (left) and Geometric (right), Deficit (D): Lognormal, Maximum Intensity (IM): Generalized Pareto; Huesca, $p10$-process.

## References

Cebrián, A.C. (1999). *Análisis, modelización y predicción de episodios de sequía.* Non-published Ph.D. dissertation.

Davison, A.C. and Smith, R. L. (1990) Models for exceedances over high thresholds, *J. R. Statist. Soc. B*, **52**, 3, pp 393-442.

Embrechts, P. et al. (1997) *Modelling extremal events*, Springer.

Gibbs, W.J. and Maher, J.V. (1967). Rainfall deciles as drought indicators. *Bureau of Meteorology Bulletin. Melbourne. Australia*, **48**.

McKee, T.B., Doesken, N.J. and Kleist, J. (1993). The relationship of drought frequency and duration to time scales. *Preprints, 8th Conference on Applied Climatology*, Anaheim, CA, pp 179-84.

Zelenhasic, E. and Salvai, A. (1987). A method of streamflow drought analysis, *Water Resour. Res.*, **23**, 1, pp 156-68.

# Bootstrapping Multiparameter Models, with Applications to Clustered Binary Data

Marc Aerts[1], Gerda Claeskens[2], Geert Molenberghs[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B3590 Diepenbeek, Belgium
[2] Centre for Mathematics and its Applications, School of Mathematical Sciences, Australian National University, Canberra, ACT 0200, Australia

**Abstract:** It is shown how a one-step semiparametric bootstrap procedure can be applied to multiparameter models in different situations: for testing hypotheses, for the construction of simultaneous confidence intervals based on local polynomial smoothers and for improved estimation and bias correction. The method is illustrated on models for clustered binary data.

**Keywords:** Bootstrap, Clustered Binary Data, Local Polynomial Smoothing, Multiparameter Models, Testing Hypotheses.

## 1 Introduction

The bootstrap is a well established statistical methodology nowadays. There are several papers and books showing a multitude of examples where the bootstrap can be implemented and applied succesfully, see e.g. Davison and Hinkley (1997). Here we are interested in applying the bootstrap to clustered binary data, typically modelled by multiparameter likelihood models. There has been considerable interest in bootstrapping generalized linear models (see e.g. Moulton and Zeger 1989) but, to our knowledge, there are not many results on applying the bootstrap to multiparameter models in general. Of course, for fully specified likelihood models, one can always apply the parametric bootstrap. Such an approach has been generalized to pseudolikelihood models and applied to clustered binary data in Aerts and Claeskens (1999a). But often the "true likelihood" is unknown and one might expect a parametric bootstrap to break down if the likelihood model of the data is grossly misspecified. Therefore, a semiparametric bootstrap approach might be preferable. Such a robust method is presented here and it is shown how it can be applied to testing hypotheses, the construction of confidence intervals and to multiparameter local likelihood models. It should be stressed that although we focus attention to *clustered binary response* data, the domain of application of these methods is much broader.

## 2 A One-Step Bootstap Procedure

Let $\boldsymbol{Y}_i$, $i = 1, \ldots, n$ be independent response variables of length $m$ with (unknown) joint density or discrete probability function (pdf) $g(\boldsymbol{y}_i; \boldsymbol{x}_i)$ where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})$ and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$, the latter representing a vector of $p$ explanatory variables. In the context of clustered binary data, $m$ corresponds to the size of the cluster.
In general, parametric inference is based on an $r$ dimensional score function $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$, where the "true" parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ is defined as the solution $\boldsymbol{t}$ to $\sum_{i=1}^n E[\boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \boldsymbol{t})] = \boldsymbol{0}$ where all expectations are w.r.t. the true pdf $g(\boldsymbol{y}_i; \boldsymbol{x}_i)$. Solving the system of equations $\sum_{i=1}^n \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \boldsymbol{t}) = \boldsymbol{0}$ leads to the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\boldsymbol{\theta}$.
Within classical maximum likelihood $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t}) = (\partial/\partial \boldsymbol{t}) \log f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ and, for clustered binary data, $f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ represents e.g. the beta-binomial distribution or the conditional model of Molenberghs and Ryan (1999) (MR-model). Note that, in this setting, the assumed pdf $f(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ might not contain the true structure $g(\boldsymbol{y}; \boldsymbol{x})$. Effects of likelihood misspecification are examined in Molenberghs, Declerck and Aerts (1998). But $\boldsymbol{\psi}(\boldsymbol{y}; \boldsymbol{x}, \boldsymbol{t})$ might also represent the pseudolikelihood scores (see Geys, Molenberghs and Ryan 1999) or generalized estimating equations.
We propose to resample the score and the differentiated score values. Based on a linear approximation, we define a bootstrap replicate of $\widehat{\boldsymbol{\theta}}_n$ as

$$\widehat{\boldsymbol{\theta}}_n^* = \widehat{\boldsymbol{\theta}}_n - \left( \sum_{i=1}^n \dot{\boldsymbol{\psi}}_i^*(\widehat{\boldsymbol{\theta}}_n) \right)^{-1} \sum_{i=1}^n \boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n) \tag{1}$$

where $(\boldsymbol{\psi}_i^*(\widehat{\boldsymbol{\theta}}_n), \dot{\boldsymbol{\psi}}_i^*(\widehat{\boldsymbol{\theta}}_n))$, $i = 1, \ldots, n$ is a sample with replacement from the set $\left\{ \left( \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_n), (\partial/\partial \boldsymbol{\theta}) \boldsymbol{\psi}(\boldsymbol{Y}_i; \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}_n) \right), i = 1, \ldots, n \right\}$. A similar linearization idea is used in simulation approaches for the bootstrap, as the linear bootstrap and the one-step bootstrap. For linear models $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the idea of resampling scores has also been proposed by Hu and Zidek (1995). Inspired by higher order approximations, the linear bootstrap (1) can be considerably improved by a one-step quadratic bootstrap (see Aerts and Claeskens 1999b).

## 3 Application 1: Hypothesis Testing

Although testing hypotheses is also of great interest in settings with clustered binary data, bootstrap tests have never been studied and applied extensively in this situation. One of the main reasons for this is that for the bootstrap to work, data have to be generated under the restrictions imposed by the specific null hypothesis. Aerts and Claeskens (1999b) show how valid Wald and score tests can be based on the one-step bootstrap by

replacing $\widehat{\boldsymbol{\theta}}_n$ by $\widehat{\boldsymbol{\theta}}_n^{(0)}$ in the rhs of definition (1). This null estimate $\widehat{\boldsymbol{\theta}}_n^{(0)}$ reflects the null hypothesis and the second term of (1) represents the random fluctuation of the bootstrap replicate $\widehat{\boldsymbol{\theta}}_n^*$ around the estimator $\widehat{\boldsymbol{\theta}}_n^{(0)}$.
As an example, consider simulated data as they appear in developmental toxicity studies with rodents. We selected dose levels 0, 0.25, 0.5, 1 and an equal number of 15 litters, assigned to each dose group. 500 datasets were generated from the beta-binomial distribution with $\text{logit}(\pi(d)) = \theta_{10}+\theta_{11}d$, $\text{FisherZ}(\rho(d)) = \theta_{20}$ under the null hypothesis that $H_0 : \theta_{11} = 0$ (no dose effect). Here, for a pregnant rodent exposed to dose $d$, $\pi(d)$ is the probability that an individual fetus is malformed and $\rho(d)$ represents the intra-litter correlation. For each run, the scores are resampled 1000 times in each dose group separately, denoted by $B_1/D$ for the linear and $B_2/D$ for the quadratic one-step bootstrap method. Resampling the complete set of scores is denoted by $B_i/A$ $(i = 1, 2)$. Finally, $B_{it}/D$ corresponds to resample the data in each dose group. The data were fitted using the pseudolikelihood model and the robust Wald and robust score statistics, testing for no dose effect, were calculated. Some results are shown in Table 1 ($*$ denotes the proportion of significant tests (at 5%) which differs significantly from 5%).

| $\theta_{10}$ | | $\chi^2$ | $B_1/D$ | $B_2/D$ | $B_{it}/D$ | $B_1/A$ | $B_2$/A |
|---|---|---|---|---|---|---|---|
| -4.0 | $W_n$ | 10.55* | 10.76* | 6.96 | 10.76* | 9.28* | 6.54 |
| | $S_n$ | 6.12 | 6.75 | — | — | 5.70 | — |
| -2.5 | $W_n$ | 7.80* | 6.60 | 5.80 | 8.20* | 6.00 | 5.20 |
| | $S_n$ | 7.40* | 5.60 | — | — | 5.20 | — |

TABLE 1. Simulated type I errors (as %), significance level 0.05. Data are generated with the beta-binomial model (with $\theta_{20} = 0.2$) and fitted using the pseudolikelihood model. $H_0 : \theta_{11} = 0$.

.

## 4 Application 2: Bootstrapping Local Likelihood Estimators

Aerts and Claeskens (1997) and Claeskens and Aerts (1999) studied local polynomial likelihood in the context of clustered binary data. Definition (1) can be modified by including kernel weights ($K((x_i - x)/h)$ for $p = 1$ and $K$ a density) and an extra term in the rhs representing a bias correction. Details and consistency results for this local version of the one-step bootstrap are given in Claeskens and Aerts (1999). There it is also indicated how simulation of the bootstrap distribution allows for the construction of simultaneous confidence intervals in a finite number of grid points.

As an illustration, consider data from the Wisconsin diabetes study. Both eyes of each of 720 younger onset diabetic persons were examined for the presence of macular edema. See Klein, Klein, Moss, Davis, and DeMets (1984) for more details. So the response data are $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ with $y_{ij}$ the binary response value of eye $j = 1, 2$ of person $i$. We will study the probability of macular edema as a function of the patient's systolic blood pressure, hereby taking the clustered nature of the data into account, as indeed the response values of both eyes are likely to be correlated. The simultaneous and pointwise 90% confidence intervals for the probability of macular edema and for the intra-person correlation, are given in Figure 1.

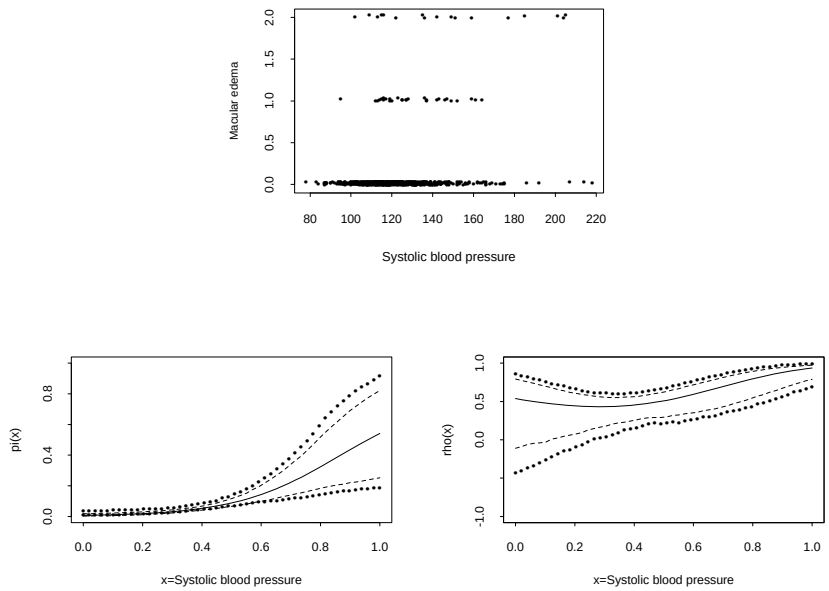



FIGURE 1. The Wisconsin diabetes data (top panel), Simultaneous and pointwise 90 % confidence intervals for the probability of macular edema (left bottom panel) and the intra-person correlation (right bottom panel).

## 5 Application 3: Bias Correction and Double Bootstrap

Although the ML estimator $\widehat{\boldsymbol{\theta}}_n$ is asymptotically unbiased, the quadratic one-step bootstrap procedure can be used for finite sample bias correction.

In practical applications a large number, say $B$, resamples are taken, resulting in a set of $B$ bootstrap estimators $\widehat{\boldsymbol{\theta}}_n^{*1}, \ldots, \widehat{\boldsymbol{\theta}}_n^{*B}$. From this set a bias corrected estimator is defined as $\widehat{\boldsymbol{\theta}}_n^{bc} = 2\widehat{\boldsymbol{\theta}}_n - \frac{1}{B}\sum_{i=1}^{B} \widehat{\boldsymbol{\theta}}_n^{*i}$.
Simulations show that this bias correction might even decrease the variance. Using a double bootstrap procedure, Aerts, Claeskens and Molenberghs (1999) study the distribution of $\widehat{\boldsymbol{\theta}}_n^{bc}$ and define a bootstrap based variance estimator for $\widehat{\boldsymbol{\theta}}_n^{bc}$.
Table 2 shows that the quadratic one-step bootstrap slope estimator is quite able to estimate the finite sample bias. The settings in this simulation were as follows. We generated 2000 data sets of size $n = 10$ and $n = 25$, for each value of $x$, from a logistic regression model $\text{logit}\{P(Y = 1)\} = \beta_0 + \beta_1 x$, with $(\beta_0, \beta_1)$ equal to (-1,-1), (-2.5, 1) or (-2.5, 2), and $x = 0, 0.25, 0.5$ and 1. For each of these 2000 data sets we constructed 1000 one-step quadratic bootstrap replicates, the latter were used to obtain the bias corrected estimates $(\hat{\beta}_0^{bc}, \hat{\beta}_1^{bc})$.
An important observation is that the bias correction even decreases the variance, as the simulated standard deviation $\sigma(\hat{\beta}_0^{bc})$ and $\sigma(\hat{\beta}_1^{bc})$ are, for all settings in this study, smaller than the corresponding simulated values of $\sigma(\hat{\beta}_0)$ and $\sigma(\hat{\beta}_1)$, respectively.

| | $\beta_0 = -1$ | | $\beta_0 = -2.5$ | | $\beta_0 = -2.5$ | |
|---|---|---|---|---|---|---|
| | $\beta_1 = -1$ | | $\beta_1 = 1$ | | $\beta_1 = 2$ | |
| | $n = 10$ | $n = 25$ | $n = 10$ | $n = 25$ | $n = 10$ | $n = 25$ |
| $E(\hat{\beta}_1)$ | -1.265 | -1.070 | 0.889 | 1.027 | 2.152 | 2.096 |
| $E(\hat{\beta}_1^{bc})$ | -1.043 | -0.988 | 0.852 | 1.001 | 1.959 | 2.022 |
| $\sigma(\hat{\beta}_1)$ | 1.511 | 0.795 | 1.555 | 0.908 | 1.303 | 0.779 |
| $\sigma(\hat{\beta}_1^{bc})$ | 1.314 | 0.747 | 1.297 | 0.836 | 1.119 | 0.735 |
| $\frac{MSE(\hat{\beta}_1^{bc})}{MSE(\hat{\beta}_1)}$ | 0.734 | 0.875 | 0.701 | 0.846 | 0.728 | 0.878 |

TABLE 2. Simulated mean, standard deviation and mean squared error values of original and bias corrected slope estimators.

**Acknowledgements:** This project was partly supported by the NATO Collaborative Research Grant 950648. The research of Gerda Claeskens was supported by the Fund for Scientific Research - Flanders (Belgium) (FWO). We thank Professor R. Klein of the University of Wisconsin, Madison, for kindly providing this data set (NIH grant EY 03083, Wisconsin Diabetic Retinopathy Study).

## References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimators in multiparameter likelihood models, *Journal of the American Statistical Association*, **92**, 1536–1545.

Aerts, M. and Claeskens, G. (1999a). Bootstrapping pseudolikelihood models for clustered binary data, *Annals of the Institute of Statistical Mathematics*, **51**, 515–530.

Aerts, M. and Claeskens, G. (1999b). Bootstrap tests for misspecified models, with application to clustered binary data", Submitted.

Aerts, M., Claeskens, G. and Molenberghs, G. (1999). A note on the quadratic bootstrap and improved estimation in logistic regression, Technical Report, Limburgs Universitair Centrum, Diepenbeek.

Claeskens, G. and Aerts, M. (2000). Bootstrapping local polynomial estimators in likelihood-based models, *Journal of Statistical Planning and Inference*, **86**, 63–80.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge: Cambridge University Press.

Geys, H., Molenberghs, G. and Ryan, L.M. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology, *Journal of the American Statistical Association*, **94**, 734–745.

Hu, F. and Zidek, J. (1995). A bootstrap based on the estimation equations of the linear model, *Biometrika*, **82**, 263–275.

Klein, R., Klein, B.E.K., Moss, S.E., Davis, M.D. and DeMets, D.L. (1984). The Wisconsin epidemiologic study of diabetic retinopaty: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years, *Archives of Ophthalmology,* **102**, 520–526.

Molenberghs, G., Declerck, L. and Aerts, M. (1998). Misspecifying the likelihood for clustered binary data, *Computational Statistics and Data Analysis,* **26**, 327–349.

Molenberghs, G. and Ryan, L.M. (1999). Likelihood inference for clustered multivariate binary data, *Environmetrics*, **10**, 279–300.

Moulton, L.H. and Zeger, S.L. (1989). Analyzing repeated measures on generalized linear models via the bootstrap, *Biometrics*, **45**, 381–394.

# Nonparametric Regression Estimators in Biased Sampling Models

J.T. Alcalá[1], J.A. Cristóbal[1] and J. Ojeda[1]

[1] Department of Statistical Methods, University of Zaragoza, 50009 Zaragoza, Spain

**Abstract:** Little attention has been paid to obtaining a nonparametric estimator of a regression function in biased sampling models. In this paper, we extend local polynomial nonparametric regression to this context, and study the asymptotic properties of the proposed estimator.

**Keywords:** Nonparametric regression; Local polynomial smoothers; biased sampling.

## 1 Introduction.

In analyzing data taken from the natural world, attention frequently has to be given to the biased sampling phenomenon, in which each observation $Y = y$ is given a different chance $w(y)$ of being recorded. In this paper we concentrate on the nonparametric estimation of a regression function of one response variable $Y$ on a covariate $X$, where the data are observed by way of biased sampling in the variable $Y$. Some examples motivating this study are encountered in the analysis of discovery data; in this context, Nair and Wang (1989) have carried out an analysis with multivariate petroleum resource data, considering a parametric situation when variables such as the volume of a pool and mean formation depth, among others, are analyzed. Some others references on the analysis of oil and gas discoveries can also be seen in this work, and the more useful weight functions in these problems are given by $w(y) = y^a$. Other examples are provided in aerial surveys. Thus, Brown (1972) discussed an aerial survey problem in traffic streams research; if one records the velocities of vehicles in a highway section at time $t$ (for example through aerial measures) one will tend to observe an unduly high proportion of slower vehicles. An appropriate weight function is now $w(y) = 1/y$. On the other hand, Rao (1989) considers a situation related to the measurement of some bones recovered in a cemetery. When a femur is not broken, three measures are possible: length, width of the upper part and width of the lower part; but if it is broken, only one width is feasible. Therefore, the data are incomplete, and it would be interesting to obtain an estimation of the regression function of the length on the widths. However,

the longer the bone, the lower the chance of having it in its entirety, and thus, recording length variable is biased with a decreasing weight function. In the context of the nonparametric estimation of a regression function, a Nadaraya-Watson type estimator is considered by Ahmad (1995), and a local polynomial fit is used in Cristóbal and Alcalá (2000) to construct other estimators in the length biased case. In this work, we consider the last approach when there is a general weight $w(x, y)$, which can jointly depend on the $x$ and $y$ variables.

## 2 The Estimator and its Asymptotic Properties.

Let us suppose a random vector with distribution function $F(x, y)$ and let $f_{XY}(x, y)$ be its density function. Let $f^w_{XY}(x, y)$ be the weighted density function that the biased sample data follow,

$$f^w_{XY}(x, y) = \frac{w(x, y) f_{XY}(x, y)}{W}, \quad W = \int w(x, y) f_{XY}(x, y) dx dy < \infty, \tag{1}$$

where $w(x, y) > 0$ (a.s) is the weight function. Conditional densities from weighted distributions are also weighted densities, for example

$$f^w_{Y|X}(y|x) \quad = \quad \frac{w(x, y) f_{Y|X}(y|x)}{c(x)}, \text{with } 0 < c(x) = \mathrm{E}[w(X, Y)|X = x],$$

where E(.) is the expectation with the non-biased distribution, in contrast to $\mathrm{E}^w(.)$, which is an expectation calculated from the weighted distribution (1).

Let $m(x) = \mathrm{E}[Y|X = x]$ be a regression function from the original population, which is unknown but which is assumed to be sufficiently smooth. Let $\beta_j = m^{(j)}(x)/j!, j = 0, \ldots, p$ be the coefficients of a Taylor expansion of $m(.)$ about the point $x$ up to degree $p$, which are assumed to exist and which are finite.

Given a biased sample $\{(X_i, Y_i)\}_{i=1}^n$ from $F^w$, a nonparametric maximum likelihood estimation of $F(x, y)$ is studied in Gill, Vardi and Wellner (1988). This estimator $\hat{F}_n$ assigns a weight proportional to $1/w(X_i, Y_i)$ in each data point, in such a way that the weights total one. The main idea is to incorporate this weight into the local least squares function, i.e. to minimize in $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ the following function

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^p \beta_j (X_i - x)^j \right)^2 \frac{K_h(X_i - x)}{w(X_i, Y_i)}.$$

where $K_h(u) = K(u/h)/h$, $K(.)$ is usually a symmetric p.d.f. and $h$ is the bandwidth. The solution to this "bias-corrected" weighted least squares

problem gives us the proposed estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}_w\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}_w\mathbf{y}, \tag{2}$$

with $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, where $\mathbf{X}$ denotes the design matrix (as in Fan and Gijbels (1996)), and

$$\mathbf{W}_w = \mathrm{diag}(K_h(X_1 - x)/w(X_1, Y_1), \ldots, K_h(X_n - x)/w(X_n, Y_n)) \tag{3}$$

In particular, we have

$$\hat{m}_n(x) = \hat{\beta}_0 = \frac{\det(\mathbf{X}^T\mathbf{W}_w\mathbf{X}^{(1)}(\mathbf{y}))}{\det(\mathbf{X}^T\mathbf{W}_w\mathbf{X})}, \tag{4}$$

where $A^{(j)}(u)$ means matrix $A$ but with column $j$ replaced by vector $u$. This expression shows the regression estimator as a ratio of two symmetric statistics in the observations. Symmetric statistics are convenient, because we can obtain approximations of their moments and their asymptotic properties are well established; furthermore a ratio statistic allows for an asymptotic analysis similar to the Nadaraya-Watson estimator for direct data (see Cristóbal and Alcalá (1998) for properties of these symmetric statistics). Observe that the new estimator is not linear in the response variable and that this complicates its analysis.
Following the usual notation in local polynomial fitting (see Fan and Gijbels (1996)), we can write the matrices $S = (\mu_{i+j-2})_{1\le i,j\le p}$ with elements $\mu_l = \int u^l K(u)du$, and $S^* = (\nu_{i+j-2})_{1\le i,j\le p}$ with elements $\nu_l = \int u^l K^2(u)du$. Furthermore, let the vectors $c_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$ and $\tilde{c}_p = (\mu_{p+2}, \ldots, \mu_{2p+2})^T$. Let $\mathbf{e}_1$ be the vector with 1 in its first position and 0 in the rest. Let $\mathrm{Var}_w[(Y - m(X))/w(X,Y)|X = x] = \nu_w^2(x)$ be the conditional variance associated with the proposed estimator.

**Theorem 1** *Let us suppose that $f_X(x) > 0$ and that $m^{(p+1)}(.)$, $f_X(.)$, $c(.)$ and $\nu_w^2(.)$ are continuous in a neighborhood of $x$. Let us further suppose that $h \to 0$ and $nh \to \infty$. Therefore, the asymptotic conditional variance of $\hat{m}_n(x)$ is*

$$\mathrm{Var}_w(\hat{m}_n(x)|\mathcal{X}) = (nh)^{-1}\frac{c(x)\nu_w^2(x)W}{f_X(x)}\mathbf{e}_1^T S^{-1}S^*S^{-1}\mathbf{e}_1\{1 + o_P(1)\}.$$

*The asymptotic bias for the odd case $p$, when $nh^{p+2} \to \infty$, is:*

$$\mathrm{Bias}_w\{\hat{m}_n(x)|\mathcal{X}\} = h^{p+1}\beta_{p+1}\mathbf{e}_1^T S^{-1}c_p\{1 + o_P(1)\}.$$

*Furthermore, for the case of an even $p$, if $f_X'(.)$ and $m^{(p+2)}(.)$ are continuous in a neighborhood of $x$, and $nh^{p+3} \to \infty$, we have that:*

$$\mathrm{Bias}_w\{\hat{m}_n(x)|\mathcal{X}\} = h^{p+2}\left\{\beta_{p+1}\frac{f_X'(x)}{f_X(x)} + \beta_{p+2}\right\}\mathbf{e}_1^T S^{-1}\tilde{c}_p\{1 + o_P(1)\}.$$

For the case $p = 0$, we obtain the estimator proposed by Ahmad (1995). When $p = 1$, the estimator proposed in (4) has the same conditional asymptotic bias as the local linear estimator acting on non-biased data:

$$\mathrm{ABias}_w\{\hat{m}_n(x)|\mathcal{X}\} = \frac{h^2}{2}\mu_2 m''(x).$$

In general, the asymptotic conditional bias has the same expression as is obtained when we use direct sampled data from the population.
One of the most relevant weighted distributions is the length-biased distribution. This case corresponds to $w(x, y) = y$ and then it is immediate that

$$\begin{aligned} c(x) &= m(x); \quad W = \mu_Y \\ \nu_w^2(x) &= m(x)\tau(x) - 1, \end{aligned}$$

where $0 < \tau(x) = \mathrm{E}[Y^{-1}|X = x] < \infty$ and $\mu_Y$ is the marginal mean of $Y$. The asymptotic variance of the proposed estimator is now

$$\mathrm{Var}_w(\hat{m}_n(x)|\mathcal{X}) = (nhf_X^w(x))^{-1}m^2(x)(m(x)\tau(x) - 1),$$

where $f_X^w(x) = m(x)f_X(x)/\mu_Y$ (the weighted marginal density function of $X$). This expression agrees with that obtained by Cristóbal and Alcalá (2000). Moreover, note that $m^2(x)(m(x)\tau(x) - 1)/n$ is the asymptotic conditional length-biased variance of the sample harmonic mean of the variable $Y|X = x$.

## 3 Application to Unemployment Time data.

In this section, we apply the proposed estimators to a real data-set corresponding to unemployment time (in days) in Zaragoza (Spain). The period considered is from 1st January 1990 to 30th April 1998. We randomly selected a sample of 93 people that have a consecutive unemployment time period covering a fixed date, specifically 1st April 1994 (precisely in the middle of the period analyzed). Data are obtained from the National Institute for Employment (INEM); see Olave *et al.* (1998) for a survival analysis of an extended set of this data. For each person, we recorded the length of their time period of unemployment, their age and sex. The long-time unemployed are not included in this study. Observe that the sampling mechanism used for selecting the sample gives higher probability to longer periods and lower probability to shorter periods. Therefore, we assume that data form a length biased sample in the response variable.
In Figure 1 we represent a local linear estimator without any modification against length-bias (dashed line), and a local linear estimator incorporating weights for compensating the length-bias (solid line). We can see the different behaviour of both estimators particularly around 35-45 years old,

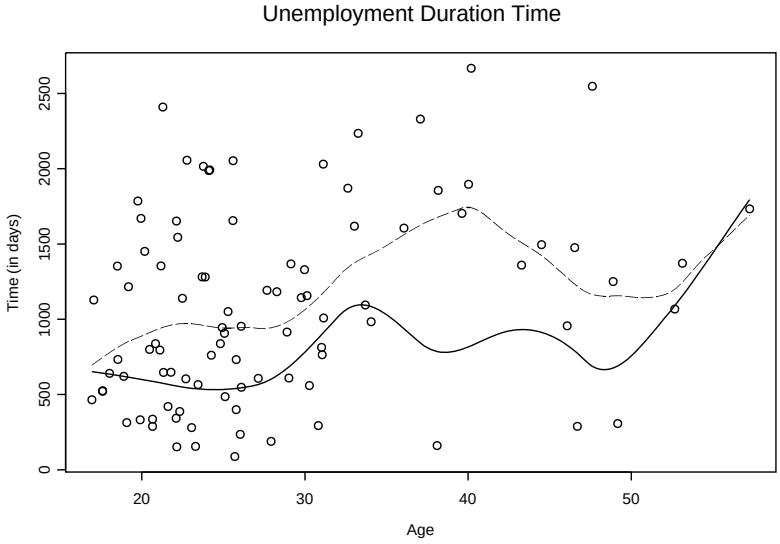



FIGURE 1. Unemployment Duration Time. Local linear estimator without weights for length-bias (dashed line) and local linear estimator incorporating the proposed weights (solid line)

where the ordinary local linear estimator overestimates the duration of unemployment.

In Figure 2, we estimate regression curves separately by sex (47 males and 46 females), with both estimates being length-bias corrected. We can see the similar mean duration time until the age of 35, and how this estimated mean time then increases rapidly for females. However, this increase is delayed for males until the age of 45 and is more moderated.

The different smoothing parameters have been selected in order to obtain a good visual impression. A detailed study of the performance of selection criteria for the smoothing parameter is beyond the scope of this paper and will be addressed elsewhere.

**Acknowledgements:** This work has been partially supported by grant No. PB98-1587. We are also grateful to Dra Pilar Olave for permission to use data on unemployment time.

## References

Ahmad, I.A. (1995). On multivariate kernel estimation for samples from weighted distributions, *Statist. Prob. Lett.*, **22**, 121–129.

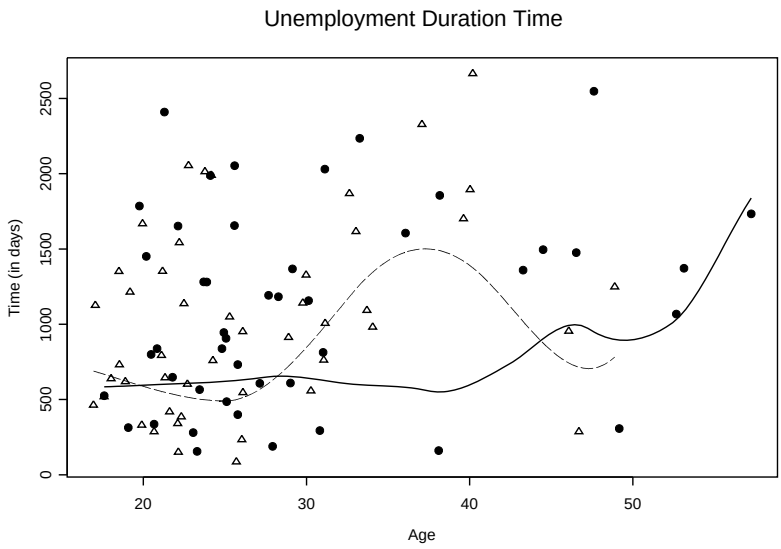



FIGURE 2. Unemployment Duration Time by Sex. Local linear estimator for men (•) in solid line and for women (△) in dashed line. Both estimates are adjusted for length-biased samples.

Brown, M. (1972). Low density traffic streams. *Addv. Appl. Prob.*, **4**, 177-192.

Cristóbal, J.A. and Alcalá, J.T. (1998). Error Process Indexed by Bandwidth Matrices in Multivariate Local Linear Smoothing, *J. Mult. Anal.*, **66**, 207–236.

Cristóbal, J.A. and Alcalá, J.T. (2000). Nonparametric Regression Estimators for Length Biased Data, *J. Statist. Plann. Infer.*, to appear.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Gill, R.D., Vardi, Y. and Wellner J.A. (1988). Large Sample Theory of Empirical Distributions in Biased Sampling Models, *Ann. Statist.* , **16**, 1069–1112.

Nair W.N. and Wang, P.C.C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics*, **31**, 423–436.

Olave, P., Salvador, M., Miguel, J.A. and Muñoz, L. (1998). Fecundity and unemployment in Spanish women. A nonparametric study., *Minist. Trabajo y Asuntos Sociales. España.*

Rao, C.R. (1989) *Statistics and Truth (2nd. ed.)*, World Scientific, Singapore.

# Underidentification?

Manuel Arellano[1], Lars Peter Hansen[2] and Enrique Sentana[1]

[1] CEMFI, Casado del Alisal, 5, 28014 Madrid, Spain
[2] Department of Economics, University of Chicago, 1126 East 59th Street, Chicago, IL 60637, USA

**Abstract:** We study the identification of an econometric model that is linear in parameters from a generalized-method-of-moments perspective. We regard underidentification as a set of over-identifying restrictions imposed on an augmented structural model. Therefore, our proposal is to test for underidentification by testing for overidentification in the augmented model using standard methods that are available in the literature. As examples we consider intertemporal asset pricing and dynamic panel data models.

**Keywords:** Instrumental variables; GMM; Identification test

In instrumental variables estimation of an econometric model it is useful to have a statistical test designed to ascertain whether the model is underidentified. Indeed Koopmans and Hood (1953, page 184) wrote:

> "It is ... natural to abandon without further computation the set of restrictions strongly rejected by the (likelihood ratio) test. Similarly, it is natural to apply a test of identifiability before proceeding with the computation of the sampling variance of estimates ... and to forego any use of the estimates, if the indication of nonidentifiability is strong."

While it was recognized in the early econometric literature on simultaneous equations systems that underidentification is testable, to date such tests are uncommon in econometric practice. Nevertheless, many econometric models of interest often imply a large number of moment restrictions relative to the number of unknown parameters and are therefore seemingly overidentified. However, this situation is often coupled with informal evidence that identification may be at fault. In those cases, an identification test may provide a useful diagnostic of the extent to which estimates are well identified.

We study the identification of an econometric model that is linear in parameters. We adopt a generalized-method-of-moments (GMM) perspective and write the model as:

$$E(\Psi_t)\alpha = 0 \tag{1}$$

where $\alpha$ is a $k+1$-dimensional unknown parameter vector in the null space of the population matrix $E(\Psi_t)$ where $\Psi_t$ is an $r$ by $k+1$ matrix constructed from data. We suppose the order condition ($r \geq k$) is satisfied, but not necessarily the rank condition. Thus the maximal possible rank of the matrix $E(\Psi_t)$ is $\max\{r, k+1\}$. The model is said to be *identified* when the null space of $E(\Psi_t)$ is precisely one dimensional. In this case the parameter vector of interest is obtained by imposing a normalization that selects one element from the null space. The selection rule can restrict one of the components of $\alpha$ to be one, or it might require that $|\alpha| = 1$ together with a sign restriction on one of the nonzero coefficients. Identification follows when the matrix $E(\Psi_t)$ has rank $k$. When $r > k$ and the model is identified, it is said to be *over-identified* because the rank of the matrix $E(\Psi_t)$ now must not be full. Instead of having maximal rank $k+1$, $E(\Psi_t)$ has reduced rank $k$. This implication is known to be testable and statistical tests of overidentification are often conducted in practice.

The model is said to be *under-identified* when the rank of $E(\Psi_t)$ is less than $k$. In this case the null space of $E(\Psi_t)$ will have more than one dimension. A single normalization will no longer select a unique element from the parameter space. Instead there exists another solution $\alpha^*$ not proportional to $\alpha$ such that

$$E(\Psi_t)\alpha^* = 0. \tag{2}$$

To test for the lack of identification, we ask whether there exists another normalized vector $\alpha^*$ that satisfies (2). We approach this question by thinking of (1) and (2) as emerging from a new *augmented model*. We attempt to determine $(\alpha, \alpha^*)$ simultaneously and ask whether they satisfy the combined over-identifying moment restrictions. If they do, then we may conclude that the original econometric relation is *not identified* or equivalently is under-identified. Thus by building an augmented equation system, we may pose the null hypothesis of underidentification as a hypothesis that the augmented equation system is over-identified. Rejections of the over-identifying restrictions for the augmented model provide evidence that the original model is indeed identified. Posed in this way, underidentification can be tested simply by applying appropriately an existing test for overidentification. For instance, a standard test for overidentification such as that of Sargan (1958) (and extended by Hansen, 1982) is potentially applicable to the augmented model.

As we will see, there are two complications that must be considered in this implementation. First, under the null hypothesis of underidentification, we are compelled to extend the normalization to extract multiple, linearly independent elements from the null space of $E(\Psi_t)$. For instance, if $(\alpha, \alpha^*)$ satisfies (1) and (2), then so does any pair of linear combinations of $\alpha$ and $\alpha^*$. Since the parameter estimates of the augmented model are of no particular interest to us, it is of little consequence which rule is used to achieve identification. Any convenient normalizations will suffice, and it is known

how to construct GMM estimators that are insensitive to normalization. Second, when we duplicate the moment relations to achieve identification of the augmented model, we may introduce some redundancy into the system. As a consequence, sometimes we will be compelled to use less than the full $2r$ moment conditions from the augmented system when testing for underidentification. We will provide some guidance as to when to expect redundancy in the moment conditions.

We initially consider identification testing in the context of a single structural equation, including comparisons to other approaches. We discuss the relationship of our method with the minimum eigenvalue tests suggested by Koopmans and Hood (1953) and Sargan (1958), and the reduced form approach proposed by Cragg and Donald (1993). We also deal with cross-equation restrictions, discussing two examples motivated in the estimation of an intertemporal asset pricing model and a translog share equation system. Finally, we consider identification testing in autoregressive models with individual effects for short panels. This is an example of a system of equations in which the valid instruments differ for different equations, and the model has a nonstandard reduced form. We provide empirical illustrations and Monte Carlo simulations for the asset pricing and the panel data examples.

**Acknowledgements:** We thank Javier Alvarez, Raquel Carrasco, and Francisco Peñaranda for able research assistance, and John Campbell for kindly allowing us to use his data.

## References

Cragg, J.G. and Donald, S.G. (1993). Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, 9, 222-240.

Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50, 1029-1054.

Koopmans, T.C. and Hood, W.C. (1953). The Estimation of Simultaneous Linear Economic Relationships. In: Hood, W.C. and T.C. Koopmans (eds.): *Studies in Econometric Method*, Cowles Commission Monograph No. 14, Chapter 6, Wiley.

Sargan, J.D. (1958). The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica*, 26, 393-415.

# A Generalization for Skewness of the Basic Stochastic Volatility Model

Francesco Bartolucci[1], Giovanni De Luca[2] and Nicola Loperfido[3]

[1] Dipartimento di Scienze Statistiche - Perugia - IT
[2] Dipartimento di Economie Società e Istituzioni, Sezione Statistica, via dell'Artigliere 19, 37129 Verona, Italy. E-mail: gdeluca@chiostro.univr.it
[3] Istituto di Economia - Urbino - IT

**Abstract:** The basic stochastic volatility (SV) model does not take into account the possible skewness of a time series. In order to overcome this drawback, we introduce a generalisation of such a model based on the assumption that the observations follow the Skew Normal distribution rather than the Normal one. The degree of skewness is regulated by an appropriate parameter; when this parameter is equal to zero, the proposed model is equivalent to the basic SV model. It turns out especially appropriate for daily stock returns.

**Keywords:** Maximum likelihood estimation; Quadrature; Stock returns.

## 1 Introduction

In the analysis of stock return time series, it is frequently observed some degree of skewness: the fall of the price in correspondence of bad news is usually more pronounced than the increase of the price in correspondence of good news. In other words, the same piece of news with an opposite relevance (increase or decrease of the interest rate of $x\%$, increase or decrease of the dividends of $x\%$) does not have the same effect on the price of the assets. Moreover, it is observed that, when volatility increases, stock prices are expected to fall. So, volatility tends to amplify the negative effect of a bad piece of news on the stock prices while it tends to mitigate the positive effect of a good piece of news. This effect is known as *volatility feedback* and has been discussed by several authors such as French, Schwert and Stambaugh (1987) and Campbell and Hentschel (1992). It can explain the prevalence of negative log-returns with respect to positive log-returns, i.e. a certain degree of negative skewness, also known as *contemporaneous asymmetry*. A different idea is that of *predictive asymmetry* which relies on the observation that future volatility is higher after a fall of stock market than after a rise.
These aspects are not taken into account in the basic stochastic volatility (SV) model. So, our aim is to propose a generalization of such a model to

overcome these limitations. This is obtained making use of the *Skew Normal* distribution which has been recently introduced by Azzalini (1985) and is a generalization for skewness of the Normal distribution. This distribution is introduced in following section whereas the assumptions of the proposed model are described in 3rd section. Then, some indications concerning the estimation of the parameters and further possible developments are given, respectively, in the 4th and 5th section.

## 2 The Skew Normal distribution

The Skew Normal distribution is defined by the density

$$f(z;\lambda) = 2\phi(z) \cdot \Phi(\lambda z) \qquad -\infty < z, \lambda < +\infty \tag{1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard Normal density function and distribution function, respectively. When (1) is the density of a random variable $Z$ we write $Z \sim SN(\lambda)$. The skewness of this distribution increases as the parameter $\lambda$ increases while it tends to -0.995 when $\lambda$ tends to $-\infty$, and to 0.995 when $\lambda$ tends to $\infty$. An important results is that for $\lambda = 0$ this distribution reduces to the standard Normal distribution. Moreover, despite skewness, it shares many properties with the Normal distribution: it is unimodal, its support is the real line and its square follows the chi-square distribution.
In this paper we deal with a generalized version of the (1) which is defined by a linear transformation (Azzalini, 1985). In this way we may obtain a distribution with skewness chosen on basis of $\lambda$ and an arbitrary value of the mean $\mu$ and the variance $\sigma^2$. When $Z$ has a distribution of this kind we write $Z \sim SN(\mu, \sigma^2, \lambda)$.
Recent developments of the Skew Normal distribution have been dealt with by several authors such as Azzalini and Capitanio (1999) and Azzalini and Dalla Valle (1999). In particular, these two papers discuss a multivariate extension of (1) and a possible statistical applications.

## 3 The model

The basic stochastic volatility model is based on the assumption that the log-return at time $t$, $y_t$, is defined as

$$\begin{aligned} y_t &= \exp(h_t/2)\varepsilon_t \\ h_t &= \gamma_0 + \gamma_1 h_{t-1} + \sigma_\eta \eta_t \end{aligned}$$

where $\{h_t\}$ is a latent process which represents the flow of news which arrives on the market and the random variables $\varepsilon_t$ and $\eta_t$ are white noises

with zero mean and unit variance. In particular, when $\varepsilon_t$ and $\eta_t$ are normally distributed and independent of one another we have the *basic* SV model which is also known as the *log-normal SV* model.
Although the basic SV model is very popular, it does not take into account the possible skewness in the log-returns. Both the conditional and the unconditional distribution of the log-returns are symmetric, and this is often in contrast with reality. This restriction is common to many financial time series models which usually capture the changing volatility and fat tails of real time series, but do not encounter the observed contemporaneous asymmetry.
The aim of this paper is to propose a SV model which takes into account the possible asymmetry of the log-returns, consistently with what is observed. This model, which may be called *Skew* SV model, is based on the following assumptions:

$$\begin{aligned} y_t &= \exp(h_t/2)\zeta_t \\ h_t &= \gamma_0 + \gamma_1 h_{t-1} + \sigma_\eta \eta_t \end{aligned}$$

where $\zeta_t \sim SN(0,1,\lambda)$, namely $\zeta_t$ has Skew Normal distribution with mean 0, variance 1 and skewness parameter $\lambda$, $\eta_t$ is still distributed as a standard normal and is independent of $\zeta_t$.
It is worthwhile to stress that our model captures the more relevant empirical stylized facts of stock return time-series, i.e. the conditional distribution has zero mean, changing volatility and some degree of skewness. The unconditional distribution of observables turns out to be a mixture of a log-normal and a Skew Normal distribution, which gives rise to a skewed distribution with zero-mean, variance equal to

$$\mathrm{E}[y_t^2] = \mathrm{E}[\zeta_t^2]\mathrm{E}\{\exp(h_t)\} = \exp(\mu_h + \sigma_h^2/2),$$

and third moment given by

$$\mathrm{E}[y_t^3] = \mathrm{E}[\zeta_t^3]\mathrm{E}\left\{\exp\left(\frac{1}{2}h_t\right)\right\}^3 = \frac{1}{2}(4-\pi)\left(\delta\sqrt{2/\pi}\right)^3 \exp\left(\frac{3}{2}\mu_h + \frac{9}{8}\sigma_h^2\right),$$

where $\delta$ is defined as $\lambda/\sqrt{1+\lambda^2}$ and assumes the sign of $\lambda$.

## 4 Parameter estimation

A natural way to estimate the parameters of the basic SV model, consists of maximizing the likelihood obtained by marginalizing on the latent variable, namely

$$L(\boldsymbol{\theta}|\mathbf{y}, h_0) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left\{ \prod_{t=1}^{T} f_Y(y_t|h_t) f_H(h_t|h_{t-1}, \boldsymbol{\theta}) \right\} dh_T \cdots dh_1 \quad (2)$$

where $\mathbf{y} = (y_1, \cdots, y_T)'$ and $f_Y(y_t|h_t)$ is the conditional density of $y_t$ given $h_t$ whereas $f_H(h_t|h_{t-1}, \boldsymbol{\theta})$ is that of $h_t$ given $h_{t-1}$ and the value of the parameter vector $\boldsymbol{\theta} = (\gamma_0, \gamma_1, \sigma_\eta, \lambda)'$. However, computing and maximising with respect to $\boldsymbol{\theta}$ the $T-$dimensional integral in (2) is really challenging and so the approach of Bartolucci and De Luca (2000) used for the basic SV model may be easily extended to deal this likelihood. In practice, this approach consists of applying a quadrature method to $L(\boldsymbol{\theta}|\mathbf{y}, h_0)$ so that it may be approximated, with the required precision, by the function $\tilde{L}(\boldsymbol{\theta}|\mathbf{y}, h_0)$ defined as

$$a^T \sum_{i_1=1}^{n} f_Y(y_1|x_{i_1}) f_H(x_{i_1}|h_0, \boldsymbol{\theta}) \sum_{i_2=1}^{n} f_Y(y_2|x_{i_2}) f_H(x_{i_2}|x_{i_1}, \boldsymbol{\theta}) \cdots$$
$$\cdots \sum_{i_T=1}^{n} f_Y(y_T|x_{i_T}) f_H(x_{i_T}|x_{i_{T-1}}, \boldsymbol{\theta}) \tag{3}$$

where $\{x_k\}$, with $k = 1, \ldots, n$, is a set of quadrature points chosen as

$$x_k = r + (k-1)a \quad \text{with} \quad a = (s-r)/n$$

while $r$ and $s$ are two finite integration limits which replaces the infinite ones. $\tilde{L}(\boldsymbol{\theta}|\mathbf{y}, h_0)$ may be formulated using matrix notation and largely available mathematical package, as MATLAB, may be used to quickly compute such a function. Moreover, it is possible to compute also the first and second derivatives of $\tilde{L}(\boldsymbol{\theta}|\mathbf{y}, h_0)$. Obviously, the higher $n$ and the wider the interval $[r, s]$, the better the accuracy of the approximation. Measures of this accuracy are provided, for the basic SV model, in Bartolucci e De Luca (2000). Even if a rigorous proof is not available yet, it seems that the same results hold in the present context. For the maximization of $\tilde{L}(\boldsymbol{\theta}|y, h_0)$ the Newton Raphson algorithm may be used; we experimented that as initial values for this algorithm it is convenient to use, for $\gamma_0$, $\gamma_1$ and $\sigma_\eta$, their maximum likelihood estimates $\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}_\eta$ based on the basic SV model. In fact, it can be proved that the distribution of such estimators is invariant under the more general Skew SV model, namely it does not depend on $\lambda$. Moreover, in order to avoid the problem of multiple maxima, it is appropriate to perform the maximization with different initial values for $\lambda$. The advantage of the maximum likelihood approach consists of the possibility of computing the first and second derivative of the log-likelihood, so that standard errors, $t$-statistics and confidence intervals may be associated to the estimates of the parameters.

Using this maximum likelihood approach, we studied the time series of some U.S. daily stock log-return in the period from 02/01/1996 to 10/03/2000. As an illustration, some simple descriptive statistics were calculated for one of these time series, the log-returns of Bank One, which consists of 1,061 observations. The mean, standard deviation and skewness index $\gamma_1$, are equal, respectively, to 0, 0.022 and -1.471. These statistics suggest that the

inclusion in the model of a possible skewed distribution might improve the fit of the data. In Table 1 we report the estimates of the parameters, with the standard errors (s.e.) and the 95% confidence intervals (CI). All the parameters are significatively different from zero, including the skewness parameter $\lambda$. This enables us to conclude that the *Skew* SV model improves the description of this time series. The same conclusion holds for a number of time series.

| Parameter | Estimate | s.e. | C.I. | |
|---|---|---|---|---|
| $\gamma_0$ | -0.782 | 0.259 | -1.288 | -0.275 |
| $\gamma_1$ | 0.902 | 0.032 | 0.838 | 0.965 |
| $\sigma_\eta$ | 0.320 | 0.059 | 0.204 | 0.436 |
| $\lambda$ | -0.822 | 0.327 | -1.462 | -0.181 |

TABLE 1. Estimation results for Bank One daily log-returns from 02/01/1996 to 10/03/2000.

## 5 Further developments

The Skew Normal distribution is a generalization for skewness of the Normal one (obtained with $\lambda = 0$), so it is clear that the proposed model is a generalization of the basic SV model. The basic SV model is nested in the new formulation, so it is possible to verify if the use of the Skev SV model implies a significative reduction in the deviance using the likelihood ratio test. Unfortunately, dealing with its distribution is problematic, so future research will focus on it.

Moreover, if it might be too restrictive to assume the same skewness for the whole time series at least two alternatives may be followed: selection of two or more sub-periods (e.g. according to economic considerations) and use of a different parameter $\lambda$ for each of them, or choice of a functional form for $\lambda_t$ depending on $h_t$. The former is actually flexible and permits to test hypothesis as $H_0 : \lambda_1 = \lambda_2$ where $\lambda_1$ and $\lambda_2$ are the skewness parameters associated to two different sub-periods. The latter seems the more appropriate to deal with real time series. However, the choice of a suitable functional form for $\lambda_t$ is not easy since it could give rise to difficulties in the parameter estimation. For instance, some results are already available for a model in which the constant parameter $\lambda$ is replaced by a time-changing parameter $\lambda_t$, whose expression is given by $\nu e^{h_t/2}$ which graduates the skewness of the Skew Normal distribution and $\nu$ is a further parameter which may be positive or negative. These results are:

1. The sampling distribution of the maximum likehood estimates of $\sigma, \gamma_0, \gamma_1$ does not depend on $\nu$.

2. The probability that a future observation differs from zero more (less) than a fixed value, conditionally on past observations, does not depend on $\nu$.

The trade-off between a more realistic model and the computational problems will be the object of future research.

## References

Azzalini, A. (1985), A Class of Distribution which includes the Normal Ones, *Scandinavian Journal of Statistics*, **12**, pp. 171-178.

Azzalini, A. and Capitanio, A. (1999), Statistical applications of the multivariate skew normal distribution, *Journal of the Royal Statistical Society* **B 61**, pp. 579-602.

Azzalini A. and Dalla Valle, A. (1996), The multivariate skew-normal distribution, *Biometrika*, **83**, pp. 715-726.

Bartolucci, F. and De Luca, G. (2000), Maximum Likelihood Estimation of a Latent Variable Time Series Model, *Applied Stochastic Models in Business and Industry, forthcoming.*

Campbell, J.Y. and Hentschel, L. (1992), No News is a Good News: an Asymmetric Model of Changing Volatility in Stock Returns, *Journal of Financial Economics*, **31**, pp. 281-218.

French, K.R., Schwert, W.G. and Stambaugh, R.F. (1987), Expected Stock Returns and Volatility, *Journal of Financial Economics*, **19**, pp. 3-29.

# Subspace algorithm cointegration analysis – An application to interest rate data

Dietmar Bauer[1] and Martin Wagner[2]

[1] Institute for Econometrics, Operations Research and System Theory, Vienna University of Technology, Argentinierstrasse 8, A-1040 Vienna.
[2] Department of Economics, University of Berne, Gesellschaftsstrasse 49, CH-3012 Berne.
E-mail addresses: Dietmar.Bauer@tuwien.ac.at, Martin.Wagner@vwi.unibe.ch

**Abstract:** In this paper the application of so called subspace methods for the specification and estimation of cointegrated systems is examined. This method, which is based on the state space representation, is suited for the analysis of general cointegrated systems of order one, i.e. is not limited to autoregressive models, as is e.g. Johansen's method. To assess the empirical usefulness of the method we apply it to perform a cointegration analysis of the US term structure of interest rates.

**Keywords:** Cointegration analysis; Subspace algorithms; Term structure.

## 1 Introduction

Over the past 15 years cointegration analysis has become one of the most popular fields in modern econometrics. By now a variety of methods is available, but the majority of analyses is carried out using the methods developed by Johansen and his co-authors. This method however has one limitation: It is restricted to the analysis of VAR models. Although this assumption may be a good approximation in many cases, the possibility of a more general data generating process deserves some attention. This can e.g. be done using subspace methods, in particular the method presented in Larimore (1983), which is dealt with here, combined with a cointegration analysis as has been examined in Bauer and Wagner (1999a). This method is suited for estimation of cointegrated ARMA models. There are already a couple of related results available in the literature. E.g. Yap and Reinsel (1995) derive the ML estimate for cointegrated Gaussian ARMA systems integrated of order one and give the distribution of the estimates, which allows to test hypotheses. The method presented in this paper is based on the state space representation and derives consistent estimates of the cointegrating space as well as the transfer function. From the latter one can easily derive e.g. an ARMA representation if this is the preferred system representation. Also test procedures for the number of common trends are

derived. We apply the method on US data to test the expectations hypothesis of the term structure, which states that the yield to maturity at time $t$ of a $k$ period pure discount bond $r_{t,t+k}$ is related to the yield on a bond with one period to maturity $r_{t,t+1}$ via equation (1):

$$r_{t,t+k} = \frac{1}{k}\sum_{j=0}^{k-1} \mathbf{E}_t(r_{t+j,t+j+1}) + L(t,k) \tag{1}$$

with $L(t,k)$ being the risk premium and $\mathbf{E}_t$ denoting the conditional expectation given the information available at time $t$. Given that interest rates are often found to be integrated of order 1 without drift, equation (1) implies that $r_{t,t+k}$ and $r_{t,t+1}$ are cointegrated if the risk premia are stationary. As equation (1) holds for all $k$, in a system with $n$ interest rates $n-1$ cointegrating relations occur, thus only one common factor $r_{t,t+1}$ is driving the system. In applications often less than $n-1$ cointegrating vectors are found. The possible reasons for this include size distortions due to multiple tests, problems with high dimensional systems with many cointegrating relations, or simply an undermodelling due to the use of a too simple model like e.g. a low order VAR model. The method presented here does not suffer from the two above mentioned possible problems, thus it may constitute a valuable additional or complementary tool.
The structure of the paper is as follows: Section 2 starts with a very brief description of the method, Section 3 then applies the method to the US interest rate data and Section 4 concludes.

## 2 Description of the method

In this section we briefly describe the subspace algorithm cointegration analysis presented in Bauer and Wagner (1999a). The starting point of our method is the state space representation of finite-dimensional, time invariant, discrete time systems

$$x_{t+1} = Ax_t + K\varepsilon_t, \qquad y_t = Cx_t + E\varepsilon_t \tag{2}$$

where $y_t, t = 0, 1, \ldots, T$ denotes the $s$-dimensional observed series. $\varepsilon_t$ denotes an ergodic, strictly stationary white noise sequence with zero mean, nonsingular innovation variance and finite fourth moments. For detailed assumptions in a martingale difference framework see Bauer and Wagner (1999a). Furthermore we restrict ourselves to systems that are strictly minimum-phase, i.e. the eigenvalues of $(A - KE^{-1}C)$ have an absolute value smaller than one. The system poles, i.e. the eigenvalues of $A$, are restricted to be inside the open unit disc or at $z = 1$. The geometric multiplicities of the eigenvalues at $z = 1$ are restricted to be equal to one, this assumption corresponds to an order of integration of one. The results of Bauer and Wagner (1999b) imply that $y_t = C_1K_1\sum_{j=1}^{t-1}\varepsilon_t + k_{st}(L)\varepsilon_t$,

where $k_{st}(L) = E + LC_{st}(I - LA_{st})^{-1}K_{st}$ is a stable and strictly minimum-phase transfer function, $L$ denoting the backward shift operator. In order to achieve identifiability (for details see Bauer and Wagner, 1999b) $C_1$ is chosen to be part of an orthonormal matrix, i.e. $C_1 \in \mathbf{R}^{s\times r}, C_1'C_1 = I_r$. Therefore there exists a matrix $C_2$ with $C_2'C_2 = I_{s-r}$ and $C_2'C_1 = 0$, i.e. $C_2$ is in the orthogonal complement of $C_1$. This representation coincides with Granger's. The first component corresponds to the common trends and the columns of $C_2$ span the cointegrating space. Therefore the cointegrating rank is equal to $s-r$ and the number of common trends equals the number of eigenvalues of $A$ at one.

The basis of the algorithm is found in the interpretation of the state vector: For given positive integers $f$ and $p$ define $Y_{t,f}^+ = [y_t', y_{t+1}', \dots, y_{t+f-1}']'$ and $Y_{t,p}^- = [y_{t-1}', y_{t-2}', \dots, y_{t-p}']'$. Further let $E_{t,f}^+ = [\varepsilon_t', \varepsilon_{t+1}', \dots, \varepsilon_{t+f-1}']'$. Let $\mathcal{O}_f = [C', A'C', \dots, (A^{f-1})'C']'$ and $\mathcal{K}_p = [K, (A - KE^{-1}C)K, \dots, (A - KE^{-1}C)^{p-1}K]$. Finally define $\mathcal{E}_f$ as the matrix, whose $i$-th block row is equal to the matrix $[CA^{i-1}K, \cdots, CK, E, 0]$. Then it follows from the system equations (2), that

$$Y_{t,f}^+ = \mathcal{O}_f\mathcal{K}_pY_{t,p}^- + \mathcal{O}_f(A - KE^{-1}C)^px_{t-p} + \mathcal{E}_fE_{t,f}^+$$

Here for notational simplicity $y_t = 0, t < 0, x_t = 0, t \leq 0$. Now the subspace algorithm can be described as follows:

1) In a first step regress $Y_{t,f}^+$ on $Y_{t,p}^-$ to obtain an estimate $\hat{\beta}_{f,p}$ of $\mathcal{O}_f\mathcal{K}_p$.

2) Typically $\hat{\beta}_{f,p}$ has full rank, whereas $\mathcal{O}_f\mathcal{K}_p$ is of rank $n$ for $f, p \geq n$. Thus approximate $\hat{\beta}_{f,p}$ by a rank $n$ matrix with decomposition $\hat{\mathcal{O}}_f\hat{\mathcal{K}}_p$.

3) Use the estimate $\hat{\mathcal{K}}_p$ to estimate the state as $\hat{x}_t = \hat{\mathcal{K}}_pY_{t,p}^-$. Given the estimate of the state, the system matrices $(A, K, C, E)$ can be estimated by OLS using the system equations.

In Bauer and Wagner (1999a) consistency for the estimates of the cointegrating space, the system order and the transfer function estimates is derived. Also a method for the estimation of the dimension of the cointegrating space is developed, which is based on estimated singular values: The approximation in step 2 of the procedure outlined above is performed using the singular value decomposition of $(\hat{\Gamma}_f^+)^{-1/2}\hat{\beta}_{f,p}(\hat{\Gamma}_p^-)^{1/2} = \hat{U}_n\hat{\Sigma}_n\hat{V}_n' + \hat{R}$, where $\hat{\Gamma}_f^+$ denotes the sample covariance of $Y_{t,f}^+$ and $\hat{\Gamma}_p^-$ the sample covariance of $Y_{t,p}^-$. Here $\hat{U}_n$ is the matrix, which contains the first $n$ right singular vectors as columns and $\hat{\Sigma}_n$ is a diagonal matrix, whose diagonal entries are the estimated singular values in decreasing order. $\hat{R}$ accounts for the neglected singular values. In the case, where there are $r$ common trends, the first $r$ singular values of the limit of $(\hat{\Gamma}_f^+)^{-1/2}\hat{\beta}_{f,p}(\hat{\Gamma}_p^-)^{1/2}$ are equal to one. In Bauer and Wagner (1999a) the asymptotic distribution of the singular values is derived and a test procedure based on the asymptotic distribution is suggested.

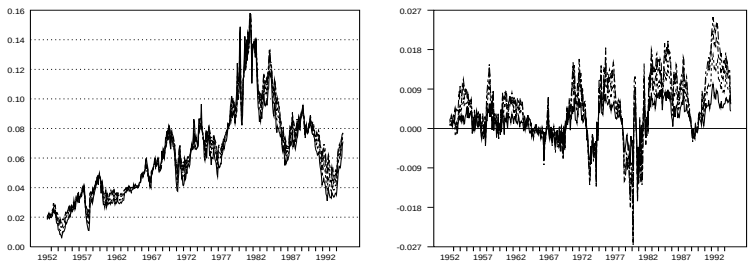



FIGURE 1. Left part of this figure: interest rates for the different maturities, from one to five years. Right part: the four spreads $r_{t,t+i} - r_{t,t+1}$.

## 3 An application to interest rate data

The interest rate data we use are the yields on 1 to 5 year US government bonds, they are displayed in the left part of Figure 1. The returns are computed from the bond prices underlying the analysis of Fama and Bliss (1987). The data range from June 1952 to December 1994 with monthly frequency.

Unit root tests performed for all 5 series lead to the conclusion that all of them are integrated of order one. Furthermore univariate unit root testing also leads to the conclusion that all the spreads $r_{t,t+i} - r_{t,t+1}$ for $i = 2, \ldots, 5$ are stationary. These spreads are displayed in the right picture in Figure 1 and are seen to quite resemble stationary time series. Hence univariate investigations lend strong support to the expectations hypothesis of the term structure. Clear results like those just mentioned are only obtained for the US market, e.g. for German data the evidence is not that supportive for the expectations hypothesis.

The results obtained by applying the subspace procedures confirm the theory and preliminary investigations. The order estimate according to AIC is equal to 5, therefore $f = p = 10$ are chosen. In Table 1 we present the first 3 estimated singular values $\hat{\sigma}_i$, the test statistic $T(1 - \sum_{j=1}^{r} \sigma_j^2 / r)$, critical values obtained by 2 different bootstrapping procedures and the asymptotic critical values, which are generated using the estimated parameters. The critical values for the bootstraps have been generated by 1000 replications of the estimated model under the null hypothesis, where the first procedure re-samples the estimated residuals and the second uses Gaussian residuals with the same covariance matrix.

The results are as expected: One common trend is found. The gap between the second and third estimated singular value leads to the order estimate $n = 2$. Note that the critical values obtained from the asymptotic theory, are differing substantially from the bootstrapped critical values. This is evidence in favor of using the latter. Also note that the tests for the number of

| $i$ | $\hat{\sigma}_i$ | $T(1-\sum_{j=1}^{r}\hat{\sigma}_j^2/r)$ | c.v.(true) | c.v.(N) | c.v. (asymp.) |
|---|---|---|---|---|---|
| 1 | 0.999 | 0.741 | 10.713 | 11.326 | 59.509 |
| 2 | 0.954 | 23.385 | 20.263 | 20.189 | 36.727 |
| 3 | 0.766 | 85.999 | – | – | – |

TABLE 1. The first 3 estimated singular values and the statistic described in Bauer and Wagner (1999a). Third column: value of the statistic. Fourth column: bootstrapped critical value, using the distribution of the residuals. Fifth column: uses normal innovations having the same covariance matrix. Last column: asymptotic values. The underlying test is one-sided.

| Component number | 2 | 3 | 4 | 5 | Hausdorff |
|---|---|---|---|---|---|
| lower bound (true) | 0.908 | 0.731 | 0.624 | 0.450 | 0 |
| upper bound (true) | 1.170 | 1.692 | 1.786 | 2.274 | 0.147 |
| lower bound (normal) | 0.971 | 0.953 | 0.940 | 0.929 | 0 |
| upper bound(normal) | 1.130 | 1.319 | 1.331 | 1.370 | 0.082 |
| estimated values | 1.0375 | 1.0648 | 1.0857 | 1.0974 | 0.0332 |

TABLE 2. Bootstrapped confidence regions for the entries of the column of $C$ corresponding to the common trend. Here *true* stands for the procedure, which re-samples the estimated residuals, whereas *normal* uses normal innovations with the same covariance matrix.

common trends are only computed up to order 2, since the state dimension is an upper bound for the number of common trends. Also the estimated eigenvalues, which are $z = 0.9121$ and $z = 0.9992$, confirm the hypothesis of only one common trend.
Finally also the hypothesis that the spreads are forming a basis for the cointegrating space can be tested. This hypothesis is equivalent to the hypothesis that the common trend is $C_1 = [1, 1, 1, 1, 1]'$. Again bootstrapping methods can be used to generate confidence intervals, see Table 2. We test this hypothesis using the Hausdorff distance. The test is therefore one-sided and the hypothesis is rejected if the Hausdorff distance between the estimated and the hypothetical space is larger than the critical value. Also for the individual components of the vector, after normalizing the first component to 1, confidence intervals around 1 can be generated. None of the null hypotheses can be rejected, and again the different distributions for the bootstrapping procedures lead to very similar results.
For comparison we have also applied the Johansen procedure to this data set. For all specifications the conclusion is always a 4-dimensional cointegrating space. However the hypothesis that the spreads span that space is rejected throughout, although tests for the individual spreads to be contained in the cointegrating space lead to an acceptance of these hypotheses.

Finally note that the state space system is preferred to the autoregressive system using the AIC criterion. The best autoregressive model leads to a

value of $-64.3731$, whereas the state space model with two states results in $-64.5548$, which is an upper bound of the AIC value, since the subspace estimate is not guaranteed to equal the ML estimate in finite samples.

## 4 Conclusions

In this paper we dealt with the application of so called subspace methods to the estimation of cointegrated systems. The methods have been applied to the Fama-Bliss data set. The results of the analysis are a confirmation of the expectations hypothesis of the term structure. Also the structure of the cointegrating space has been investigated, showing that the spreads between the interest rates seem to be stationary, in accordance with the univariate statistics. Note, however that the assumptions in the multivariate setting of course are different, as we also model the interdependencies between the various interest rates. The application demonstrates, that the state space model leads to good models as measured by the AIC. Furthermore the testing of hypotheses on the common trends is also easily incorporated in this framework. It has to be noted however, that similar results have not been achieved for other data sets and that the procedures seem up to now lack a profound theoretical justification when exogenous inputs and nonzero means and time trends are present. Also the accuracy for small data sets seems to be poor in some cases, as has been noted when analyzing German interest rate data.

## References

Bauer, D. and Wagner, M. (1999a). Estimating cointegrated systems using subspace algorithms. Submitted to Journal of Econometrics.

Bauer, D. and Wagner, M. (1999b). Unit root analysis in a state space framework: Canonical form and maximum likelihood analysis. Mimeo.

Fama, E.F. and Bliss, R.R. (1987). The information in long-maturity forward rates. *American Economic Review*, **77**, 680-692.

Johansen, S. (1995). *Likelihood-Based Inference on Cointegration in the Vector Autoregressive Model.* Oxford: Oxford University Press.

Larimore, W.E. (1983). System identification, reduced order filters and modelling via canonical variate analysis. In Rao, H.S. and Dorato, P. (Eds.) *Proc. 1983 American Control Conference 2.* Piscataway, NJ: IEEE Service Center, 445-451.

Yap, S.F. and Reinsel, G.C. (1995). Estimating and Testing for Unit Roots in a Partially Nonstationary Vector Autoregressive Moving Average Model. *Journal of the American Statistical Association*, **90**, 253-267.

# Structural Equation Models with Neural Network Techniques - The Idea of the Black Box

Jörg Betzin[1]

[1] Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany, betzin@cs.tu-berlin.de

**Abstract:** Neural networks and structural equation models are two methods of evaluating path models, whereas the first one looks in prediction of the outputs and the second one in the covariance structure of the path model. We would like to describe some differences and commonalities between these methods and introduce partial least square modeling as a connection between both of them.

**Keywords:** structural equation models; neural networks; path models

## 1 Introduction

The idea of this paper stems from a presentation by Leo Breimann held at the "Workshop on Semi- and Nonparametric Modeling" in Munich last year. He offered two statements for the relation between statistics and machine learning:

1. The people in machine learning are interested in the prediction tool and because of that they often obtain much better results than statistical procedures provide.

2. On the other hand the learning procedures are mostly working like a black box with a focus only on the output of the machine and its error rate.

The second statement, he explained, required to have more details about this black box. The following paper will be a contribution in this direction. One of the most popular models in machine learning are neural networks (NN). We look upon a network as the modeling of causal relationships between different variables via a set of parametrized equations with the target to specify the parameters of the model.
Nothing else is done by the *structural equation models* (SEM) in the statistical framework. And in both cases the target in general will be achieved by searching for a minimum of a loss or discrepancy function.

The only difference is the loss function. While the neural networkers are looking at the error rate of the output, the statisticians are looking at the inherent relationship structure between the variables.
The following exponations will describe three types of modeling a causal chain system with their different loss functions. On the one side a neural network, on the other side a structural equation model, known as LISREL, and "between" of them the *partial least square* (PLS) model (the LISREL and the PLS model are termed "structural equation model" throughout the remainder of this paper).
Due to space constraints we restrict ourselves to a particular case of neural networks. The same holds true for SEMs. In this way the abstract ideas are not clouded by technicalities. Also there is no discussion about identifiability of a model and other very important aspects of modeling.

## 2 The Path Model

The idea is to describe a system of relationships in a so called path model. This means a graphical model where the constructions under investigation are connected by directed paths. For example, we can describe a multiple regression model as a path model like in Fugure 1 (the described model is a special regression model with two uncorrelated sets of input variables).

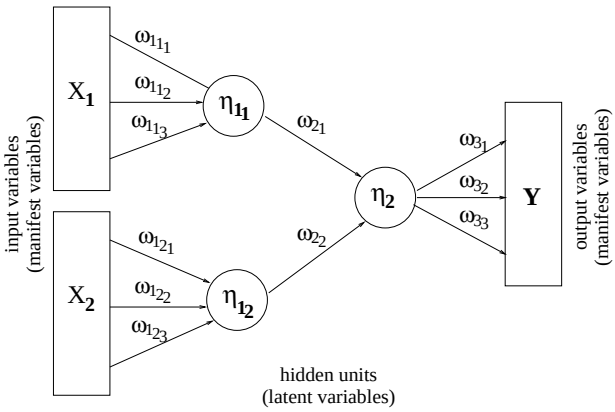



FIGURE 1. Path model for regression.

More generally we describe a path model in terms of NN's as some observable input and output variables (graphically represented as rectangles) connected by so-called hidden units (characterized as circles), whereas the connections are formulated in terms of parametrized equations. In the framework of SEMs the observable variables are called manifest variables (MV) and the hidden units latent variables (LV).
As an example we would like to handle with a kind of a multiple regression problem, represented in the above chart. There are much more complex

models conceivable, but in the interest of clarity of the general idea we will deal with this "simple" model.
Names and symbols (bold faced capitalized Latin letters characterize matrices of observations and bold faced small letters are vectors in the parameter or variable set, whereas the special meaning should be clear from the context; small roman letters are scalars from the corresponding vectors):
$\mathbf{X}$ set of input MV
$\mathbf{Y}$ set of output MV
$\omega$ weight vectors (the parameters of interest in the model)
$\eta$ vector of the hidden units or latent variables

### 2.1 Neural networks approach

The above problem will be handled here as a multilayer feedforward network with a backpropagation learning algorithm (for details see e. g. Bishop, 1995), which is represented as a set of functions

$$f(\omega, .) : \mathcal{R}^{\mathcal{P}} \longrightarrow \mathcal{R}^{\mathcal{Q}}$$

mapping the input $\mathbf{x} \in \mathbf{X} \subseteq \mathcal{R}^{\mathcal{P}}$ to the output $\mathbf{y} \in \mathbf{Y} \subseteq \mathcal{R}^{\mathcal{Q}}$. In our example the function $f(\omega, \mathbf{x})$ is a composition of three functions $f_1, f_2, f_3$ working in the single layers of the model. In particular we have:

$$\begin{aligned} f_{1_m}(\omega_{1_m}, \mathbf{x}_m) &= \sum_{p=1}^{P_m} \omega_{1_{mp}} x_{mp} &= \eta_{1_m} \\ f_2(\omega_2, \eta_1) &= \sum_{m=1}^{2} \omega_{2m} \eta_m &= \eta_2 \\ f_3(\omega_3, \eta_2) &= \omega_3 \eta_2 &= \mathbf{y} \end{aligned} \tag{1}$$

(The indices are described as follows: $m = 1, 2$ the two sets of input variables; $P$ the whole number of input variables; $P_m$ number of input variables in set $m$; $Q$ the number of output variables.)
If we characterize the $q$th output of the network by $f^q$ we get the components of $f(\omega, \mathbf{x})$:

$$f^q(\omega, \mathbf{x}) = \omega_{3q} \left( \sum_{m=1}^{2} \omega_{2m} \left( \sum_{p=1}^{P_m} \omega_{1_{mp}} x_{mp} \right) \right) \tag{2}$$

which leads, in the actual case, because there are linear (identitiy) activation functions in all hidden units, to a feedforward network with seemingly no real hidden units:

$$f^q(\omega, \mathbf{x}) = \left( \sum_{m=1}^{2} \left( \sum_{p=1}^{P_m} \omega_{3q} \omega_{2m} \omega_{1_{mp}} x_{mp} \right) \right) = \sum_{p=1}^{P} \omega_{q_p} x_p \tag{3}$$

If we try to determine the parameters $\omega$ with a minimal error rate in form of a least square loss function, in the case of data vectors $\mathbf{y}_q$ and $\mathbf{x}_p$, we get (with $\|.\|^2$ being the squared Euclidian norm):

$$\omega = \arg\min_{\omega} \sigma(\omega) = \arg\min_{\omega} \sum_{q=1}^{Q} \left\| \mathbf{y}_q - \sum_{p=1}^{P} \omega_{q_p} \mathbf{x}_p \right\|^2 \tag{4}$$

which is actually a special kind of canonical correlation between $\mathbf{X}$ and $\mathbf{Y}$, without consideration of the inner structure of the model. This is the *black box* we mentioned above. (In models with other than linear activation functions (see (2)) the situation may be different.)

### 2.2 The LISREL model

The aim of structural equation modeling in statistics is the analysis of the covariance structure between the different sets of the manifest variables. The LISREL model consist of two parts. The *measurement model* describes the relations between the manifest variables and the connected latent variables in form of a factor analytic model, while the *structure model* specifies the relations between the latent variables as a system of interdependent (linear) equations corresponding to the underlying path model (see e. g. Brachinger, 1996). In this way we have a measurement equation system (MES):

$$\begin{array}{rcl} \mathbf{x} & = & \mathbf{H}_1 \Omega_1 + \delta \\ \mathbf{y} & = & \eta_2 \Omega_3 + \varepsilon \end{array} \tag{5}$$

where $\mathbf{H}_1 = (\eta_{1_1}, \eta_{1_2})$ and a structural equation system (SES):

$$\eta_2 = \mathbf{H}_1 \Omega_2 + \zeta \tag{6}$$

with some additional assumptions we do not refer to here and the parameter matrices

$$\Omega_1 = \begin{pmatrix} \omega_{1_1} & 0 \\ 0 & \omega_{1_1} \end{pmatrix}, \quad \Omega_2 = \omega_2^{|} = (\omega_{21}, \omega_{22}), \quad \Omega_3 = \omega_3 = \begin{pmatrix} \omega_{31} \\ \vdots \\ \omega_{3Q} \end{pmatrix}.$$

Because the latent variables ($\eta_1$ and $\eta_2$) are not observable, the parameters in (5) and (6) are not directly estimable, so the aim is to look at the covariance matrix of the LISREL model and try to derive estimations for $\omega$.
If we characterize by

$$\Phi = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{pmatrix} := \begin{pmatrix} \mathrm{Var}[\eta_{1_1}] & \mathrm{Cov}[\eta_{1_1}, \eta_{1_2}] \\ \mathrm{Cov}[\eta_{1_2}, \eta_{1_1}] & \mathrm{Var}[\eta_{1_2}] \end{pmatrix},$$

from the covariance matrix

$$\Sigma := \mathrm{Cov}\left((\mathbf{x}^{|}, \mathbf{y}^{|})^{|}\right) = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{xy}}^{|} & \Sigma_{\mathbf{yy}} \end{pmatrix}$$

we get for our actual problem (for detailed exponation see e.g. Bollen, 1989):

$$\begin{array}{rcl} \Sigma_{\mathbf{xx}} & = & \Omega_1 \Phi \Omega_1^{|} + \mathrm{Cov}[\delta] \\ \Sigma_{\mathbf{xy}} & = & \Omega_1 \Phi \Omega_2^{|} \Omega_3^{|} \\ \Sigma_{\mathbf{yy}} & = & \Omega_3 \left(\Omega_2 \Phi \Omega_2^{|} + \mathrm{Var}[\zeta]\right) \Omega_3^{|} + \mathrm{Cov}[\varepsilon] \end{array} . \tag{7}$$

If we summarize in the vector $\theta$ the parameters $\omega$ and the variance or covariance components in (7) (mean $\Phi$ and the covariances of the error terms) we can describe $\Sigma$ as a function of $\theta$ ($\Sigma = \Sigma(\theta)$).
Again we will remark, that we don't speak about the conditions to the model for estimability of $\theta$, but if these assumptions hold, the empirical covariance matrix $\mathbf{S}$ is an unbiased estimator for $\Sigma$. The estimation problem now is to find estimations $\widehat{\theta}$ for $\theta$ such that $\widehat{\Sigma} = \Sigma(\widehat{\theta})$ is a "good" approximation of $\mathbf{S}$. The "goodness" of the approximation we will again measure by a loss function (tr(.) is the trace of the argument matrix):

$$\sigma(\theta) = \frac{1}{2}\mathrm{tr}\left((\mathbf{S} - \Sigma(\theta))^2\right). \tag{8}$$

### 2.3 The PLS model

Like the LISREL model the PLS model consists of a measurement and a structure model. But the PLS-Model is more prediction-oriented and the desideratum of the developer H. Wold (Wold, 1975) was to allow the construction of a model with only few distribution assumptions to the variables. Therefore the MES and SES have the same form as in (5) and (6) in the LISREL model, but the idea of estimating the unknown parameters is different. The above remarked prediction ability needs estimates for the latent variables (not so in LISREL) and the latent variables are constructed as linear combinations of the corresponding manifest variables

$$\begin{array}{rcl} \eta_{1_1} & = & \mathbf{x}_1 \lambda_{1_1} \\ \eta_{1_2} & = & \mathbf{x}_2 \lambda_{1_2} \\ \eta_2 & = & \mathbf{y} \lambda_2 \end{array} . \tag{9}$$

The first step in the PLS-Algorithmus is the estimatation of the latent variables, whereas the MES and SES will be evaluated in an ordinal regression framework for (5) and (6) afterwards.
To complete the first step it is necessary to put in the structure model. This is done by the construction of so-called instrumental variables

$$\eta_m^* = \sum_{m' \in C(m)} \eta_{m'} \rho_{mm'} \tag{10}$$

and an iteration algorithm, which provides "new" weights $\lambda$ for (9) from the instrumental variables by least square estimation. On the other hand the instrumental variables in (10) are evaluated by the "known" latent variables from (9). The summation in (10) is over all latent variables which are directly connected with the "actual" latent variable $\eta_m$ (this is the index set $C(m)$) and the weights $\rho_{mm'}$ characterize the strength of this connection (for example as a correlation coefficient). The iteration algorithm will be done only for a "partial" part (hence "partial" least squares) of the model (for every latent variable in turns) while the other latent variables are taken as known (for more details see e. g. Lohmöller, 1989).
For taking correlation weights in (10) and a special least square criterion in weight estimation in (9) we get the following optimization function, to maximize for the latent variables (see e. g. Mathes, 1993):

$$\rho(\lambda) = \sum_{m=1}^{M} \sum_{m' \in C(m)} (\mathrm{corr}(\eta_m, \eta_{m'}))^2 \tag{11}$$

where $M$ is the number of all latent variables in the model.
In the second step of the PLS-Algorithm the structure coefficients $\omega$ are estimated by ordinary least squares regression of the MES or SES, which leads to a loss function that looks very similar to that of the neural network in (4), but is in fact very different. In the case of the output variables we get (the case of the input variables will not be discussed here):

$$\begin{aligned} \sigma(\omega_3) &= \textstyle\sum_{q=1}^{Q} \|\mathbf{y}_q - \omega_{3q}\eta_2\|^2 \\ &= \textstyle\sum_{q=1}^{Q} \|\mathbf{y}_q - \omega_{3q}(\eta_1\Omega_2)\|^2 \\ &= \textstyle\sum_{q=1}^{Q} \left\|\mathbf{y}_q - \omega_{3q}\sum_{m=1}^{2}\sum_{p=1}^{P_m}(x_{mp}\lambda_{1_{mp}})\omega_{2_m}\right\|^2 \\ &= \textstyle\sum_{q=1}^{Q} \left\|\mathbf{y}_q - \sum_{m=1}^{2}\sum_{p=1}^{P_m} x_{mp}(\lambda_{1_{mp}}\omega_{2_m}\omega_{3q})\right\|^2 \end{aligned} \tag{12}$$

The loss function for the outputs is never more than a loss function for the parameters of $\omega$ for the hidden unit, which is directly connected with the output, whereas the other parameters in (12) are fixed at there previous estimations. This seems nothing more than a scale adjustment of this unit relative to the single output variables.
In this way we would like to characterize the PLS model as a minmax strategy. Where the manifest variables ar modelled with minimum loss from the latent variables (the hidden units), which are carrying the information about the structural relations in the model. But the latent variables are evaluated as linear combinations of the manifest variables (which carry the information of the measurement model) with maximal adjustment in the structure model (for example measured by correlation coefficients).

## 3 Discussion

First of all, the models for NNs, LISREL, and PLS, which are used in this contribution are very simple because we would like to show the general idea and not the complex nature of the different modeling tools. There are a lot of other, more complex models with different estimations for all of the three procedures (for more details see the list of references).
But already on the referred simple model we can see some differences and commonalities. As we have remarked above the three models describe the same path but with different goals.
Moreover the LISREL model, in the basic form, needs a lot of "hard" distribution assumptions. If they hold we are able to test hypotheses about the model itself. This is not true for the neural network. But are the distribution assumptions realy fulfilled? This question is very hard to decide. Because of that, in the SEM framework the LISREL model belong to the so-called "hard models", whereas the PLS model was developed especially for the task with unknown distribution assumptions, for the case "*when theoretical knowledge is scarce*" (Wold, 1975). This is one of the bridges between LISREL and NNs.
The other bridge is the the purpose of the LISREL approach "*to study the* ***structure*** *of the observables as reflected by their dispersion (variance-covariance) matrix*" (Jöreskog/Wold, 1982, p.266) on the one side and the prediction of outcomes from the model in NNs on the other side, while again the PLS model serves both aspects of the model.
All of the three models have their advantages and disadvantages. And the LISREL model is much better in describing the structure of the path model, whereas the NN is the best way to describe a prediction model. But both the statistician and the neural networker can learn from each other.
If the path is right in construction, the neural networker is in general not able to interprate the weights of the model. In addition, if the prediction is bad, or the generalization ability, it is very hard to adapt the model, because there is only little knowledge about the inner workings of the model. In this case the SEM can possibly provide some ideas about the connections of the hidden units and their adaptation to obtain better results.
On the other hand, if we are not sure about any distribution assumptions the estimated weights in the LISREL model can be completly wrong. In this case the PLS model and the NN can provide results to evaluate the validity of the model.
One great advantage of the neural networkers is here great experience in use of nonlinear functions in the path model. This task is not very developed both in the LISREL and the PLS approach. Here the statisticians should find a fruitfull area to carry over these methods to their models.
There are some more aspects like robustness of the methods, the role of activation functions (used in NNs) in SEMs, the relationship between fit indices in SEMs and error rates in NNs. And there are some (first) results

in single areas of them but not a general connection. It should thus be an interesting and fruitful task to examine further other relations between these models both in theory and practice.

## References

Bishop, Ch. M. (1995). *Neural Networks for Pattern Recognition.* Oxford: Clarendon Press.

Bollen, K. A. (1989). *Structural Equation Models* New York et al.: John Wiley and Sons.

Brachinger, H. W. (1995). *Linear Structural Relationships - Eine Einführung in LISREL.* in: Fahrmeir, L.; Hamerle, A.; Tutz, G. (Eds.): *Multivariate statistische Verfahren*, Berlin, New York: de Gruyter.

Jöreskog, K. G.; Wold, H. (Eds.) (1982). *Systems Under Indirect Observations (Part I).* Amsterdam et al.: North-Holland Publishing Company.

Lohmöller (1989). *Latent Variable Path Modeling with Partial Least Squares.* Heidelberg: Physica-Verlag.

Mathes, H. (1993). *Global Optimization Criteria of the PLS-Algorithm in Recursive Path Models with Latent Variables.* in: Haage, K.; Bartholomew, D. J.; Deistler, M. (Eds.): *Statistical Modeling and Latent Variables*, Netherlands: Elsevier Science Publishers, 229-248.

Wold, H. (1975). *From hard to soft modeling.* in: Wold, H. (Ed.): *Group report: Modeling in complex situations with soft information*, Third Congress of Econometrics Research Report, Uppsala.

# Rational expectations and switching regime models: theory and application to the term structure of interest rates

Arielle Beyaert[1] and Juan José Pérez-Castejón[1]

[1] Fac. de Economía y Empresa. Universidad de Murcia. 30100 - Murcia (Spain). E-mail: arielle@um.es (A. Beyaert) and jjpc@um.es (J.J. Pérez-Castejón)

**Abstract:** In order to evaluate the efficiency of the monetary transmission mechanism, we develop the formulas for testing rational expectations theory in the term structure of interest rates with VAR models of stochastically switching regimes in which all the parameters are regime-dependent. These formulas are obtained for the strict version of rational expectations and for the case where measurement errors are assumed in the expectations relationship. They are extensible to other contexts which involve I(1) variables linked by rational-expectations behaviors. The testing procedure is separately implemented on Spanish and US interest rates. Measurement errors must be assumed to accept the theory.

**Keywords:** interest rates; term structure; rational expectations; Markov switching regimes; non linearity.

## 1 Introduction

The so-called "monetary transmission mechanism" allows the monetary authorities to control only the short term interest rates for the implementation of the monetary policy. This mechanism is most effective if the hypothesis of Rational Expectations (REH) governs the term structure. This explains why this theory has been extensively tested over the last twenty years, although it has very often been rejected. However, this rejection may be due to the linearity of the models used in the testing. In this paper, we develop a method of testing the Rational Expectations Hypothesis (REH) for the term structure of interest rates in VAR models that allow for unobservable Markov switching regimes. These models tackle the non-linearities of the relationships originated in stochastic changes of the economy. In what follows, we present a summary of our original paper, explaining what we have obtained, without detailed formulas or empirical results. We acknowledge the support of the Spanish R&D National Plan under CICYT Research Grant SEC97-1253.

## 2 The model

Let $R_{t,n}$ be the interest rate at time t of an asset with maturity within n periods from t, and let $r_t$ be the interest rate of an asset maturing in $(t+1)$. According to the Rational Expectations Theory of the term structure of interest rates, when $R_{t,n}$ and $r_t$ correspond to assets which both have a short maturity, typically measured in terms of days, weeks or months, Shiller, Campbell and Schoenholtz (1983) suggest a formula which, in terms of the spread between the longer-term and shorter-term rates, is:

$$S_{t,n} = R_{t,n} - r_t = \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) E_t\left(\Delta r_{t+i}\right) + k = E_t\left(S^*_{t,n}\right) + k \qquad (1)$$

where $E_t$ represents the rational expectations of the agents conditional on the information available at moment t and k is a constant liquidity premium. Campbell y Shiller (1987) develop a procedure to test the present-value relation of the type expressed in (1) for the case when n is infinite. Whether n is finite or not, it is easy to show that if the interest rates are I(1) and the expectations theory is true, $S_{t,n}$ is I(0) and there exists a ECM model that relates $\Delta r_t$ and $S_{t,n}$. From this model it is possible to derive a bivariate VAR defined on $\Delta r_t$ and $S_{t,n}$. In this VAR, all the parameters are constant. The possibility of behavioural modifications of economic agents caused by political, institutional or economic changes are not considered. We allow for such modifications by introducing the possibility of stochastic changes of regime, generalising the approach of Hamilton (1988) and Sola and Driffill (1994).
Denoting $Z_t = (S_t - \mu_{S,x_t}\ ,\ \Delta r_t - \mu_{\Delta r,x_t})'$ where the letter $\mu$ refers to the conditional mean, our model is :

$$Z_t = \sum_{i=1}^{p} B^{(i)}_{x_{t-i}} Z_{t-i} + u_t \qquad (2)$$

In this model, $x_t$ is an unobservable variable that takes a value 0 or 1, according to the state of the economy at date t. It is governed by a first-order Markov process, with transition probabilities $p_{ij},\ i,j = 0,1$. Conditional on $x_t$, the distribution of the errors $u_t$ is $N(0, \Omega_{x_t})$.
This model extends the Hamilton (1988) approach in two directions: it allows the short rate $r_t$ to depend on the past values of the long rate $R_t$ and it allows not only the means, the variances and covariances, but also the autoregressive coefficients to vary with the state . Only the first extension has been contemplated by Sola and Driffill (1994), who centre their study on the specific and simplest case in which the maturity of the longer rate $R_t$ is twice the maturity of the short rate $r_t$ (n=2). We also deal with other values of n.

The 12+8p parameters of this model are estimated by maximum likelihood, applying numerical optimisation techniques. A 5-steps filter process is used, similar to the one described for instance in Hamilton (1988) and Sola and Driffill (1994). This procedure slows down very fast as p increases. It is therefore essential to make a sensible selection of the starting values given to the parameters in the estimation algorithm. The procedure we use in this paper is identical to the one described in detail in Beyaert and Pérez-Castejón (2000).

## 3 The restrictions implied by the expectations theory

The rational expectations equation (1) indicates that agents calculate their expectations conditional on the available information at time t. In the context of switching-regime models, the relevant information at time t would be $H_t = \{Z_t, \cdots, Z_{t-p+1}, x_t, \cdots, x_{t-p+1}\}$ if the states of the economy were directly observable by the agents. As the states are not observable, the relevant information is reduced to $h_t = \{Z_t, \cdots, Z_{t-p+1}\}$. Obtaining the restrictions on the elements of $B^{(i)}_{x_{t-i}}$ in (2) implied by equation (1) is, however, untractable. But conditioning on $H_t$, and using the law of iterated expectations, it is possible to obtain sufficient conditions for equation (1) to be true. We obtain the specific expressions for these conditions. They can be tested using a non-linear Wald test, or alternatively a LR test. The LR test is considered only when $p = 1$ and $n = 2$, because it is extremely difficult to estimate the restricted model for higher values of these parameters. Note also that these conditions can be extended with minor modifications to other contexts in which rational expectations theory linking I(1) variables has to be tested.

## 4 Application on the term structure of interest rates of Spain and the USA

We have estimated model (2) and tested rational expectations both on Spanish and US interest rates. However, for brevity, we summarize here only the results referring to Spain.

### 4.1 The Spanish inter-bank interest rates

The data correspond to the interest rates of the Spanish inter-bank money market between January 1986 and May 1995 on a weekly basis. Three different terms have been considered: one week, two weeks and four weeks. The short-term rate $r_t$ is in all cases the one-week rate, the other two interest rates play the role of the longer term rate $R_t$, so that $n = 2, 4$. In Prats-Albentosa and Beyaert (1998), a linear model accepted the hypothesis for $n = 24$, but rejected it for $n = 2, 4$ and 12; the rejection is very

strong in the case of $n = 2$ (two weeks). It is our purpose to check whether these conclusions change for $n = 2$ and $n = 4$ when a non-linear model of the type of (2) is adjusted, using the sufficient conditions that we have obtained to test REH.

The use of VAR models to test REH is justified if the interest rates exhibit a unit root, whereas the spread is I(0). The first step of the empirical study starts, therefore, with the testing of the presence of a unit root in the series $r_t$, $R_{t,n}$ , and $S_{t,n}$ , $n = 2, 4$. We use Phillips-Perron tests, according to which all the interest rates exhibit one unit root, whereas the spreads are I(0). However, the unit root may stem from the existence of a structural break in the mean of the series. To check for that possibility, we apply recursive ADF tests, as developed by Banerjee, Lumsdaine and Stock (1992). The results confirm that a structural break, if present, is not the only source of non-stationarity of the series.

Model (2) has been estimated for n=2 and n=4 as described above. The maximum value of p we have considered is 5. To check the empirical validity of the models, we apply specification tests which constitute bivariate extensions of those developed by Hamilton (1996). They are based on the scores of the likelihood function with respect to the parameters at time t. We use them to test for autocorrelation and ARCH effects in the residuals. For $p$ as low as 1, there is no symptoms of autocorrelation, although there are clear symptoms of non modelled ARCH effect. But it must be noted that the heteroskedasticity still present in these models is far below the heteroskedasticity detected in linear VAR models with the same data. Moreover, these linear models require much higher values of $p$ to avoid autocorrelation. We also apply Hansen (1996) test which allows testing the validity of the linear model against the non-linear model (2). This test tackles the problem of the existence of unidentified parameters under the null hypothesis, which rules out the application of standard Likelihood Ratio tests. It requires Monte-Carlo simulations in every application, in order to obtain the critical values. The results indicate an overwhelming domination of the non-linear version over the linear one.

The next step in the construction of the model consists of simplifying it down in order to increase efficiency. Looking at the estimated values, the means of the models seem to be equal in both states and some autoregressive coefficients seem to be non significant. We test these simplifying restrictions both jointly and separately, and reestimate the model under the accepted restrictions.

Besides the simplification restrictions, it is worth mentioning that the probability $p_{00}$ of staying in state 0 is high and systematically above state 1. This fact is reinforced by the relative size of variances: those of state 0 are small and far below those of state 1. The former may be qualified as a "low-variance high-persistence stable state" , whereas the latter would be more a "high-variance unstable state".

From these estimations, the so-called "smoothed probabilities" may be in-

ferred: on the basis of the estimated vector of parameter and the full sample of T observations, an inference is drawn about the historical state the process was in at some date t. An analysis of these probabilities, which we cannot reproduce here due to limited space, provides an additional proof of the usefulness of our model. They indeed show that the model is able to perceive the changing characteristics of the market. For brevity, let us mention the main feature of these smoothed probabilities: they correctly reflect the much higher stability of the market from the middle of 1989 onwards, as well as the instability period that extends from October 1992 to the third term of 1993. All these dates coincide with specific events that affected the market: June 1989 is the date of the entrance of Spain in the European Monetary system; the period from the fall of 1992 to the fall of 1993 coincides with the crisis of the EMS: it is marked by the "monetary turmoil" of 1992, and the depreciation of the Peseta in September and November of that same year and of May of 1993; the EMS crisis ended in August 1993 with the enlargement of the fluctuation bands, although the Spanish monetary authorities purposely let increase the rates to very high levels in the fall of 1993, which is captured by the model.
As far as the expectations hypothesis is concerned, they are tested for the case $p = 1$, and for $n = 2, 4$. They are rejected for both values of $n$, although Granger causality from $S_t$ to $\Delta r_t$, required by (1), is satisfied.
It is often the case that the strict version of REH is rejected, but a weaker version that makes allowance for a random error term in the REH relation is accepted. This error is usually attributed to measurement or specification errors, lack of information, and so on . Although it is relatively easy to deduce the theoretical expression of these restrictions once it has been done for the case where no measurement errors are assumed, their practical implementation has to be reduced to very small values of p and n. We obtain the new expression for the restrictions under measurement errors and we test them only in the simplest possible case: n=2 and p=1, which corresponds to the two-week model. In this case, REH is accepted with a p-value of 2.72%.

### 4.2 Conclusions

The switching-regime models developed in our paper overwhelmingly dominate linear models applied to the same data. The formulas that we develop to test the rational expectations hypothesis of the term structure of interest rates are then applied, and the results are supportive of REH if measurement errors are allowed in the expectations building process.

## References

Banerjee, A., Lumsdaine, R.L. and Stock, J.H. (1992) Recursive and sequential tests of the unit-root and trend-break hypothesis: theory and international evidence. *Journal of Business & Economic Statistics*, **10**(3), 271-87.

Beyaert, A. and Pérez-Castejón, J.J. (2000) Switching-regime models in the Spanish inter-bank market. *The European Journal of Finance,* **6**, 1-20.

Campbell, J. and Shiller, R. (1987). Cointegration and tests of present value models. *Journal of Political Economy*, **95**(5), 1062-88.

Hamilton, J.D. (1996) Specification testing in Markov-switching time-series models. *Journal of Econometrics*, **70**, 127-157.

Hamilton, J.D. (1988) Rational-expectations econometric analysis of changes in regime. *Journal of Economic Dynamics and Control*, **12**, 385-423.

Hansen, B. (1996) Erratum: the likelihood ratio test under non-standards conditions: testing the Markov switching model of GNP. *Journal of Applied Econometrics*, **11**, 195-98.

Prats-Albentosa, M.A. and Beyaert, A. (1998) Testing the expectations theory in a market of short-term financial assets. *Applied Financial Economics,* **8**, 101-9.

Shiller, R., Campbell, J. and Schoenholtz, K. (1983). Forward rates and future policy: interpreting the term structure of interest rates. *Brookings Papers on Economic Activity*, **1**, 173-217.

Sola, M. and Driffill, J. (1994) Testing the term structure of interest rates using a stationary vector autoregression with regime switching. *Journal of Economics Dynamics and Control*, **18**, 601-28.

# Higuest-Density Forecast Regions: Some Evidence in the Spanish Stock Market

Natividad Blasco[1] and Rafael Santamaría[2]

[1] Department of Accounting and Finance, University of Zaragoza, Doctor Cerrada, 1-3, 50005 Zaragoza, Spain
[2] Department of Business Administration. Public University of Navarre, Campus de Arrosadia, 31006 Pamplona, Spain

**Abstract:** This paper proposes the use of highest-density regions in practical forecasting. We explore whether the information they provide is likely to be more effective compared to that offered by other easier methods of determining the forecast densities.

**Keywords:** Highest Density Regions (HDRs), Bootstrap Techniques, Conditional Heteroskedasticity, January Effect, Day of the Week Effect.

## 1 Introduction

Following the conventional view mentioned above, forecast regions can easily be constructed as symmetrical intervals about the mean. However, according to Hyndman (1995), this method cannot be appropriate in the non-linear or non-normal context since, under such circumstances, any skewness or other asymmetry, high kurtosis, multimodality or the relatively high importance of extreme values will not be properly reflected. Highest-density regions are, in this case, flexible enough to provide more accurate information about the forecast density.
The purpose of this paper is to construct HDRs for every close return of the General Index of the Madrid Stock Exchange in January 1998 and to assess the reliability of the information they provide compared to that offered by the simulation of the model proposed and that offered by the hypothesis of normal distribution defined by the historical standard deviation and mean. In so doing, we attempt to highlight the usefulness of processing and analysing historical information. In the next section the data and the model are presented; section 3 presents a brief description of the HDR construction techniques; section 4 reports the empirical results and section 5 concludes.

## 2 Data Description and Model Selection

The data set consists of daily close prices of the Madrid Stock Exchange General Index for the period January 1994 to December 1997 as well as the close prices corresponding to each trading day in January 1998 to control the out-of-sample forecasts. Daily returns ($X_t$) are calculated as the log differences of the General Index prices.
With respect to the choice of the model it should be noted that our forecasts concern daily close-to-close returns over the first month of 1998 and therefore, as has been widely discussed, both the day of the week effect as well as the January effect should be considered. The initial examination of the data reported in Table 1 supports the intuition of a positive effect on Fridays included in approximately the first half of January. Moreover, the examination of the autocorrelation function as well as the results offered for the Spanish stock market in Blasco, Del Rio y Santamaría (1997) reveal significant first order autocorrelation. To formally test for these characteristics some preliminary regressions were run with dummy variables for the different days of the week, regardless of whether they are in the first half of January. In addition, the robustness of the estimates both to autocorrelation and heteroskedasticity is also examined. Eliminating from the model those variables that are not significant and taking into account the conditional heteroskedasticity through a GARCH model (following the suggestions in Lamoreaux y Lastrapes (1990), among others), the model can be stated as follows:

$$\begin{aligned} x_t &= \mu + \phi x_{t-1} + \gamma V_{1mt} + \epsilon_t \\ \epsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \\ Z_t &\ i.i.d.\ N(0,1) \end{aligned} \tag{1}$$

where $x_t$ denotes return at time t, $V_{1mt}$ is a dummy variable for a Friday in the first half of January, et denotes the residual at time $t$ and $\sigma_t^2$ denotes variance at time $t$. This model supports the supposition outlined above.

## 3 Methodology

Let $\{X(_{ti})\}$ denote the time series of observed returns at times $t_1 < t_2 < ... < t_n$. The $100(1-\alpha)\%$ highest-density region $R\alpha$ summarizes the density $p_{m/n}(x)$ with $n < m$ given all observations up to and including tn. That is, $R\alpha = \{x : p_{m/n}(x) \geq f\alpha\}$, where $f\alpha$ is chosen such that $Pr(X_{m/n} \in R\alpha) = 1 - \alpha$.
To compute any forecast region it is necessary to first estimate the forecast density. We use a simulation procedure where error terms in the model are replaced by a bootstrap of the fitted residuals, as is suggested in some recent papers (e.g. Thombs y Schucany, 1990; McCullough, 1994; Douglas,

1996). The bootstrap resampling technique used in this paper is that recommended by Miguel and Olave (1998), given that the model produces heterocedastic residuals and therefore it is almost impossible to find a backward representation of the volatility. By drawing random samples with replacement from the centered standarized residuals we get the bootstrap residuals and the bootstrap estimates of close returns and variances for each of the 20 trading sessions in January 1998 as follows:

$$x^*_{t+j} = \mu + \sigma x_{t+j-1} + \gamma V_{1mt} + z^*_{t+j}\sigma^*_{t+j} \qquad (2)$$
$$\sigma^{*2}_{t+j} = \omega + \alpha z^{*2}_{t+j-1}\sigma^{*2}_{t+j-1} + \beta\sigma^{*2}_{t+j-1}$$

where $x^*_i = x_i$ , $z^*_i = z_i$ if $i \leq t$ and $\sigma^*_i = \sigma_i$ if $i \leq t+1$. We have repeated this procedure 1500 times and hence 1500 values of close return are calculated for each of the 20 trading sessions.

## 4 Empirical Results

Table 2 reports the parameter estimates of the model. All of them are significant at the 5% level. The sum $\alpha + \beta < 1$ although it is close to unity, supports the stationarity condition.

The HDR for each of the trading days is calculated at the 75% confidence level in order to make a clearer identification of the close return values that are most likely to occur. To compute the forecast density we have considered five decimal figures of the bootstrap close returns so that an investor would not notice any great difference between a 1,352% daily close return and a 1,359% daily close return, but he would be able to appreciate a difference between 1,35% and 1,34%. The higher the precision level in the HDR construction procedure, the greater the difference with respect to the conventional methods of constructing forecasting intervals.

To assess the relative performance of this forecasting procedure and its potential applicability we compare the probability assigned to the real close return data varying in a short range with that derived from two assumptions implying considerably lower computational cost. First (A1) we assume a normal behaviour defined by the historical mean and standard deviation. In so doing, it is not necessary either to propose a model or the HDR construction. Second, we assume (A2) the empirical distribution derived from the simulation of the model proposed, avoiding in this case the computational cost of HDR construction. Columns 4-5 of Table 3 contain our results on the accuracy of the HDR information. Specifically, Column 4 mainly deals with the usefulness of processing and analysing historical information. The results provide evidence in favour of this kind of work. Comparisons in column 5 are mainly aimed at studying whether 75% HDRs are good information selectors. Although in 4 out of the 20 cases the corresponding HDR has not been able to improve the results from the empirical distribution of the fitted model, it is worth mentioning that only the results for January 13th and 30th represent a significant disadvantage, given that the

empirical density of the simulation clearly outperforms the normal inferred probabilities.
To summarize, the probabilities assigned by the HDRs are, in average, a 73% higher than those arising from A1 and a 29% higher than those from A2. The range along which the ratio HDR-probability to normal-probability varies is set between 0,59 and 7,85. These figures change to 0,71 and 2,04 when the comparison deals with HDR versus simulation of the model.

## 5 Conclusions

In this paper we have proposed the use of HDR to obtain better information on the forecast density. As we have shown, the probability levels are significantly higher when HDRs are compared with other lower cost procedures and therefore, forecast could be more accurate using this technique. This specific application provides reliable results for the period under analysis.

## References

Blasco, N., del Río C., Santamaría, R. (1997) The random walk hypothesis in the Spanish Stock Market: 1980-1992. *Journal of Business Finance and Accounting* **vol. 24**, n. 5, 667-683.

Douglas, S. (1996) Bootstrap confidence intervals in a switching regression model, *Economic Letters*, **53**, 7-15.

Hyndman, R.J. (1995) Highest-density Forecast Regions for Non-linear and Non-normal Time series Models, *Journal of Forecasting*, **14**, 431-441.

Lamoureux, C.G., Lastrapes W.D. (1990) Persistence in variance, structural change and the GARCH model, *Journal of Business and Economic Statistics*, **8, 2**, 225-234.

McCullough, B.D. (1994) Bootstrapping Forecast Intervals: An application to AR(p) Models, *Journal of Forecasting*, **13**, 51-66.

Miguel, J., Olave, P. (1998) Forecast intervals in ARCH models: Bootstrap versus parametric methods, *Applied Economic Letters*, forthcoming.

Thombs, L.A.; Schucany, W.R. (1990) Bootstrap prediction intervals for autoregression, *American Statistical Association*, **85, 410**, 486-492.

**Table 1. Preliminary mean Values**

| Mean | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| Total | -0.000083 | 0.001245 | 0.000677 | 0.000694 | 0.000855 |
| January | -0.000412 | 0.004432 | 0.001497 | 0.000254 | 0.004215 |
| Rest of the year (except January) | -0.000052 | 0.000933 | 0.000601 | 0.000734 | 0.000555 |
| First half of January (approximately) | -0.000326 | 0.003682 | -0.00107 | -0.000912 | 0.009783 |
| Rest of the year (except first half of January) | -0.000158 | 0.001144 | 0.000750 | 0.000758 | 0.000473 |

Data set: January 1994-December 1997

**Table 2. Parameter Estimates**

(p-values are given in parentheses)

| Mean equation | | | Variance equation | | | loglikelihood |
|---|---|---|---|---|---|---|
| $\mu_{est}$ | $\phi_{est}$ | $\gamma_{est}$ | $\omega_{est}$ | $\alpha_{est}$ | $\beta_{est}$ | ML |
| 0.000635 | 0.11646 | 0.007713 | 0.000004 | 0.09949 | 0.86012 | 3210.42 |
| (0.02313) | (0.00064) | (0.0291) | (0.002) | (0.0000) | (0.0000) | |

**Table 3. Real close returns and Comparative Results**

| | DATES | RETURN | HDR vs A1 (Normal Distribution) | HDR vs A2 (Simulation of the fitted model) |
|---|---|---|---|---|
| Friday | 02/01/98 | 0.025272 | S | |
| Monday | 05/01/98 | 0.017313 | S | S |
| Wednesday | 07/01/98 | -0.005286 | S | S |
| Thursday | 08/01/98 | -0.007445 | I | S |
| Friday | 09/01/98 | -0.001367 | | I |
| Monday | 12/01/98 | -0.008781 | I | S |
| Tuesday | 13/01/98 | 0.013441 | I | |
| Wednesday | 14/01/98 | 0.008285 | S | I |
| Thursday | 15/01/98 | 0.00392 | S | S |
| Friday | 16/01/98 | 0.018278 | S | S |
| Monday | 19/01/98 | 0.002148 | S | S |
| Tuesday | 20/01/98 | 0.009605 | S | S |
| Wednesday | 21/01/98 | 0.00167 | S | S |
| Thursday | 22/01/98 | 0.002952 | S | S |
| Friday | 23/01/98 | -0.009369 | S | S |
| Monday | 26/01/98 | 0.011948 | | |
| Tuesday | 27/01/98 | 0.008349 | S | S |
| Wednesday | 28/01/98 | 0.003285 | S | S |
| Thursday | 29/01/98 | -0.001555 | S | S |
| Friday | 30/01/98 | 0.002461 | | |

Close returns are given to facilite the observation of Figure 1

S means Superiority of HDR; I means Indiference; blanks mean worse performance of HDR

# Null Intercept Measurement Error Models

Heleno Bolfarine, Reiko Aoki, Julio M. Singer

[1] Instituto de Matemática e Estatistica, Universidade de São Paulo, São Paulo, S.P., Brasil

**Abstract:** The paper considers models for analyzing dependent data sets where the covariates are measured with error and the intercept is considered to be known. The dependent structure arises from the fact that repeated observations are obtained from the same individual. Inference is approached via maximum likelihood estimation. A general structure for dependency is obtained by considering a mixed model for the observed values of the true covariate, thus providing a clustering structure for observations from the same individual. A simpler structure may result after testing for components of variance. Applications are considered to a data set related to dental plaque removal after toothbrushing.

**Keywords:** known intercept, dependent variables, maximum likelihood

## 1 Introduction: Measurement Error Models

In a recent paper, Singer and Andrade (1997) proposed regression models to analyze data from a pretest/posttest study designed to compare two types of toothbrushes with respect to the efficacy in removing dental plaque. In that study, a dental plaque index was obtained on each of 26 (14 female and 12 male) preschoolers, before and after toothbrushing, with a regular and an experimental (hugger) toothbrush. No intercepts were included in the proposed models, since null pretest dental plaque indices imply null expected posttest values. The models also allowed for correlated within individual measurements (bringing into perspective the fact that the same subjects were evaluated under two different experimental conditions) as well as for nonlinear relationships between the posttest dental plaque index (taken as the random response variable) and the pretest dental plaque index (considered as a fixed explanatory variable).

Given the symmetry between the pretest and the posttest variables and the fact that the dental plaque index is not measured precisely, we believe that measurement error models constitute an interesting alternative to analyze the data. Moreover, the amount of dental plaque certainly is evaluated imprecisely.

A simple linear regression model with measurement errors typically considered in the literature is defined by the equations

$$Y_i = \alpha + \beta x_i + e_i, \tag{1}$$

$$X_i = x_i + u_i, \tag{2}$$

with

$$(e_i, u_i, x_i)' \sim N_3(\mathbf{0}', \mathrm{diag}(\sigma_e^2, \sigma_u^2, \sigma_x^2)'), \tag{3},$$

where diag denotes diagonal matrix, $i = 1, \ldots, n$. It is well known that without further assumptions such a model is not identifiable and to bypass this inconvenience, we must make an assumption about the parameters which includes i) $\sigma_e^2$ or (and) $\sigma_u^2$ known, ii) $\lambda = \sigma_e^2/\sigma_u^2$ known, iii) $k_x = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ known and iv) $\alpha$ known. The last alternative, where the intercept is assumed known, is discussed mainly in Chan and Mak (1979) and Patefield (1985). Both papers consider maximum likelihood estimation emphasizing aspects of large sample inference. Taking $\alpha$ equals to zero, without loss of generality, the moment estimators of the elements of the parameter vector $\theta = (\beta, \mu_x, \sigma_x^2, \lambda, \sigma_u^2)$ are

$$\hat{\beta} = \overline{Y}/\overline{X}, \quad \hat{\mu}_x = \bar{X}, \quad \hat{\sigma}_x^2 = \bar{X} S_{XY}/\bar{Y},$$

$$\hat{\lambda} = \frac{\overline{Y}}{\overline{X}} \left( \frac{\overline{X} S_{YY} - \overline{Y} S_{XY}}{\overline{y} S_{XX} - \overline{X} S_{XY}} \right), \quad \text{and} \quad \hat{\sigma}_u^2 = \frac{\overline{Y} S_{XX} - \overline{X} S_{XY}}{\overline{Y}}, \tag{4}$$

where $\bar{Y} = \sum_{i=1}^n Y_i/n$, $S_{XY} = \sum_{i=1}^n (Y_i - \bar{Y}) X_i/n$, and similarly for the other sample moments. As noted by Chan and Mak (1979), these are maximum likelihood estimators provided that the variance estimators are nonnegative. In this paper, the the model defined by equations (1)-(3) is extended by considering

$$Y_{ij} = \beta_i x_{ij} + e_{ij}, \tag{5}$$

$$X_{ij} = x_{ij} + u_{ij}, \tag{6}$$

where $(e_{ij}, u_{ij}, x_{ij})'$ is distributed according to the three-variate normal distribution given in (3). This model is used with the data set presented in Singer and Andrade (1997) which is related to dental plaque removal by $p = 2$ different brands of toothbrushes that we call conventional ($i = 1$) and experimental ($i = 2$). Specifically, $Y_{ij}(X_{ij})$ are the amount of dental plaque observed in individual $j$ after (before) toothbrushing with toothbrush $i = 1, 2$. As it seems to be the case, the amount of plaque is measured imprecisely (with error) given the imprecisions associated with the measuring devices for the amount of dental plaque. Moreover, since the same individual is tested with both toothbrushes it seems necessary to consider models that can incorporate a certain amount of dependency among measurements from the same individual. To model dependency from observations coming from the same individual, we consider that the true covariate values are related through a random effects model given by

$$x_{ij} = \mu_x + a_j + \delta_{ij}, \tag{7}$$

with $a_j$ and $\delta_{ij}$ all independent, with $a_j \sim N(0, \sigma_x^2)$ and $\delta_{ij} \sim N(0, \sigma_\delta^2)$, $j = 1, \ldots, n$ and $i = 1, 2$. Thus, under the correlation structure (7), it can be verified that

$$\rho_{12} = corr(x_{1j}, x_{2j}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2 + \sigma_\delta^2}, \tag{8}$$

$j = 1, \ldots, n$. A special case of model (7) follows when $\sigma_\delta^2 = 0$, in which case, $x_{ij} = \mu_x + a_j$, specifying that the true unobserved covariate values (true dental plaque amount) are produced in both occasions by the same variable, in which case the correlation structure (8) reduces to $\rho_{12} = \text{corr}(x_{1j}, x_{2j}) = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$, $j = 1, \ldots, n$.
Section 2 presents the log-likelihood function under the dependent model specified by (5)-(7) and a disussion related to the derivation of the maximum likelihood estimators. To start numerical procedures for computing the MLEs, moment estimators can be considered. The inverse of the observed information matrix can be used as an estimate of the asymptotic covariance matrix of the MLEs. Section 3 is dedicated to an application of the results derived in the previous sections to the dental plaque data set which has motivated the consideration of the dependent model.

## 2 Maximum likelihood estimation in the depedent model

In this section, we consider maximum likelihood estimation in the dependent model. The likelihood function is presented from which the likelihood equations can be obtained. The inverse of the observed information matrix can be used for estimating the asymptotic covariance of the maximum likelihood estimators.
Under the model specified by equations (5)-(8), it follows that the observed vector $(X_{1j}, Y_{1j}, X_{2j}, Y_{2j})'$ is distributed according to the fourth-variate normal distribution with mean vector $(\mu_x, \beta_1\mu_x, \mu_x, \beta_2\mu_x)'$ and covariance matrix given by

$$\begin{bmatrix} \sigma_X^2 + \sigma_\delta^2 & \beta_1(\sigma_x^2 + \sigma_\delta^2) & \sigma_x^2 & \beta_2\sigma_x^2 \\ \beta_1(\sigma_x^2 + \sigma_\delta^2) & \beta_1^2(\sigma_x^2 + \sigma_\delta^2) + \lambda_1\sigma_u^2 & \beta_1\sigma_x^2 & \beta_1\beta_2\sigma_x^2 \\ \sigma_x^2 & \beta_1\sigma_x^2 & \sigma_X^2 + \sigma_\delta^2 & \beta_2(\sigma_x^2 + \sigma_\delta^2) \\ \beta_2\sigma_x^2 & \beta_1\beta_2\sigma_x^2 & \beta_2(\sigma_x^2 + \sigma_\delta^2) & \beta_2^2(\sigma_x^2 + \sigma_\delta^2) + \lambda_2\sigma_u^2 \end{bmatrix},$$

where $\sigma_X^2 = \sigma_x^2 + \sigma_u^2$. By using general properties of the multivariate normal distribution it follows that the log-likelihood function for $\theta =$

$(\beta_1, \beta_2, \mu_x, \sigma_x^2, \sigma_\delta^2, \sigma_u^2,$
$\lambda_1, \lambda_2)'$ can be written, disregarding unimportant constants, as

$$logL \propto -\frac{n}{2}Log[\sigma_u^4 J] + \frac{1}{J}\{\mu_x\left(b_2\sigma_\delta^2 + \lambda_2\sigma_u^2\right)\left(\lambda_1\sum_{j=1}^n X_{1j} + \beta_1\sum_{j=1}^n Y_{1j}\right) \quad (9)$$

$$+\mu_x\left(b_1\sigma_\delta^2 + \lambda_1\sigma_u^2\right)\left(\lambda_2\sum_{j=1}^n X_{2j} + \beta_2\sum_{j=1}^n Y_{2j}\right)$$

$$+\sigma_x^2[\lambda_1\lambda_2\sum_{j=1}^n X_{1j}X_{2j} + \beta_1\beta_2\sum_{j=1}^n Y_{1j}Y_{2j} + \beta_1\lambda_2\sum_{j=1}^n X_{2j}Y_{1j}$$

$$+\beta_2\lambda_1\sum_{j=1}^n X_{1j}Y_{2j}] + \left(\sigma_\delta^2 + \sigma_x^2\right)[(\beta_1\lambda_2\sum_{j=1}^n X_{1j}Y_{1j} + \beta_2\lambda_1\sum_{j=1}^n X_{2j}Y_{2j}$$

$$-\frac{b_2\sum_{j=1}^n Y_{1j}^2 + b_1\sum_{j=1}^n Y_{2j}^2}{2}] + \frac{\sigma_\delta^2\left(\sigma_\delta^2 + 2\sigma_x^2\right)}{\sigma_u^2}[b_2(\beta_1\sum_{j=1}^n X_{1j}Y_{1j}$$

$$-\frac{\beta_1^2}{2}\sum_{j=1}^n X_{1j}^2 - \frac{\sum_{j=1}^n Y_{1j}^2}{2}) + b_1(\beta_2\sum_{j=1}^n X_{2j}Y_{2j} - \frac{\beta_2^2}{2}\sum_{j=1}^n X_{2j}^2 - \frac{\sum_{j=1}^n Y_{2j}^2}{2})]$$

$$-\frac{1}{2}\{\left(\sum_{j=1}^n X_{1j}^2 + \sum_{j=1}^n X_{2j}^2\right)[\lambda_1\lambda_2\sigma_u^2 + \left(\sigma_\delta^2 + \sigma_x^2\right)\Phi]$$

$$+\left(\lambda_2\sum_{j=1}^n Y_{1j}^2 + \lambda_1\sum_{j=1}^n Y_{2j}^2\right)\left(\sigma_\delta^2 + \sigma_x^2 + \sigma_u^2\right) + n\mu_x^2[2\beta_1^2\beta_2^2\sigma_\delta^2$$

$$+\left(\sigma_u^2 + 2\sigma_\delta^2\right)\Phi + \lambda_1\lambda_2\sigma_u^2]\}\},$$

where $b_1 = \beta_1^2 + \lambda_1$, $b_2 = \beta_2^2 + \lambda_2$, $\Phi = \beta_1^2\lambda_2 + \beta_2^2\lambda_1 + \lambda_1\lambda_2$, and

$$J = \beta_1^2\beta_2^2\sigma_\delta^2\left(\sigma_\delta^2 + 2\sigma_x^2\right) + \lambda_1\lambda_2\sigma_u^2\left(\sigma_X^2 + \sigma_\delta^2\right) + \Phi\left[\sigma_\delta^2\left(\sigma_\delta^2 + 2\sigma_x^2\right) + \sigma_u^2\left(\sigma_\delta^2 + \sigma_x^2\right)\right].$$

To derive the maximum likelihood estimators of $\theta = (\beta_1, \beta_2, \lambda_1, \lambda_2, \mu_x, \sigma_x^2,$ $\sigma_u^2)'$, numerical procedures are required to solve the likelihood equations given by

$$\mathbf{U}(\theta) = \frac{\partial \log L(\theta)}{\partial\theta} = \mathbf{0}. \quad (10)$$

To start the procedures, moment estimators can be used. By equating population to sample moments the following (explicit) moment estimators are obtained:

$$\hat{\mu}_x = \frac{\overline{X}_1 + \overline{X}_2}{2},\ \hat{\beta}_1 = \frac{2\overline{Y}_1}{\overline{x}_1 + \overline{X}_2},\ \hat{\beta}_2 = \frac{2\overline{Y}_2}{\overline{X}_1 + \overline{X}_2},\ \hat{\sigma}_x^2 = \frac{S_{X1Y2}\left(\overline{X}_1 + \overline{X}_2\right)}{2\overline{Y}_2}$$

$$\hat{\sigma}_\delta^2 = \frac{S_{Y1Y1}\left(\overline{X_1} + \overline{X_2}\right)}{2\overline{Y_1}}, \quad \hat{\lambda}_1 = \frac{1}{\hat{\sigma}_u^2}\left(S_{Y1Y1} - \frac{4\overline{Y}_1^2(\hat{\sigma}_x^2 + \hat{\sigma}_\delta^2)}{(\overline{X}_1 + \overline{X}_2)^2}\right),$$

$$\hat{\sigma}_u^2 = S_{X1X1} - \left(\hat{\sigma}_x^2 + \hat{\sigma}_\delta^2\right) \quad \hat{\lambda}_2 = \frac{1}{\hat{\sigma}_u^2}\left(S_{Y2Y2} - \frac{4\overline{Y_2}^2\left(\hat{\sigma}_x^2 + \hat{\sigma}_\delta^2\right)}{\left(\overline{X_1} + \overline{X_2}\right)^2}\right),$$

where $S_{X1Y2}$, $S_{Y1Y1}$, $\bar{X}_1$ and so on, are as defined before. If the variance $\sigma_\delta^2 = 0$, then the likelihood equations and the observed information matrix present much simpler expressions. Moreover, simulation studies indicate that it is more likely to have convergence of the numerical algorithms used to solve the likelihood equations in the case $\sigma_\delta^2 = 0$ than otherwise. The hypothesis $H_0 : \sigma_\delta^2 = 0$ can be tested by using the score type statistics

$$Q_R = \mathbf{U}(\bar{\theta})\mathbf{I}_o^{-1}(\bar{\theta})\mathbf{U}(\bar{\theta})/n, \tag{11}$$

which in large sample sizes is approximately distributed according to the chisquare distribution with one degree of freedom (Sen and singer, 1993), where $\bar{\theta}$ is the maximum likelihood estimators of $\theta$ under $H_0$ and $\mathbf{I}_o$ denotes the observed information matrix obtained by computing the second derivative of $\log L(\theta)$ with respect to $\theta$.
Since the usual regularity conditions are satisfyed under the depedent model specified by equations (5)-(8), it follows, as $n \to \infty$, $\sqrt{n}(\hat{\theta}_M - \theta) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_F^{-1}(\theta)^{-1})$, where $\mathbf{I}_F(\theta)$ is the expected information of $\theta$ obtained by computing the expected value of the second derivative of $\log L(\theta)$ with respect to $\theta$. The expected Fisher information is derived in Aoki et al. (2000). The observed information $\mathbf{I}_o(\hat{\theta})$ evaluated at the MLE is a consistent estimator of $\mathbf{I}_F(\theta)$. For testing $H_0 : \beta_1 = \beta_2$ we can consider the Wald statistics

$$Q_w = (\mathbf{C}\hat{\theta} - \mathbf{C}\theta)'\mathbf{C}\mathbf{I}_o^{-1}(\hat{\theta})\mathbf{C}'(\mathbf{C}\hat{\theta} - \mathbf{C}\theta), \tag{12}$$

with $\mathbf{C} = (1, -1, 0, 0, 0, 0, 0)$. The altervative hypothesis can also be in the direction of one of the $\beta$'s being greater than the other, as for example, $H_1 : \beta_1 > \beta_2$. To test the hypothesis $H_0 : \sigma_\delta^2 = 0$, we can consider a score type statistics given by $Q_R = \mathbf{U}(\bar{\theta})\mathbf{I}_F^{-1}(\bar{\theta}_n)\mathbf{U}(\bar{\theta}_n)/n$, where $\bar{\theta}_n$ is the maximum likelihood estimator of $\theta$ and which is expected, for large sample sizes, to be distributed according to the chi-square distribution with one degree of freedom. The advantage of using $Q_R$ is the fact that $\theta$ has to be estimated only under $H_0$.

## 3 Application to the Dental Plaque Data

The data set analysed in the sequel is presented in Singer and Andrade (1997) where a ordinary regression analysis is considered. We reanalyse it by considering the null intercept measurement error model given in (5)-(8), with $x_{ij}$ being the true dental plaque index before toothbrushing with

toothbrush $i$ $(i = 1, 2)$ in childreen $j$ $(j = 1, \ldots, 26)$ and $X_{ij}(Y_{ij})$ the corresponding
observed index before (after) toothbrushing. To initiate the numerical procedure that will produce the maximum likelihood estimates, moment estimates are computed first. By using the expressions for the moment estimators derived in Secion 2, the following moment estimates are obtained: $\hat{\mu}_x = 1.7687$, $\hat{\beta}_1 = 0.1559$, $\hat{\beta}_2 = 0.4362$, $\hat{\sigma}_x^2 = 0.4077$, $\hat{\sigma}_u^2 = 0.1338$, $\hat{\sigma}_\delta^2 = 0.4173$, $\hat{\lambda}_1 = 0.2350$, $\hat{\lambda}_2 = 0.9161$. By considering the dependent model given in (5)-(8) with $\sigma_\delta^2 = 0$, we can solve the likelihood equations given to obtain the MLE of $\theta = (\beta_1, \beta_2, \mu_x, \sigma_x^2, \sigma_u^2, \lambda_1, \lambda_2)'$ which is given by $\hat{\theta} = (0.1471, 0.4540, 1.7589, 0.5379, 0.48122, 0.1024, 0.2676)'$. Using the Fisher information matrix derived in Aoki et al. (2000), the asymptotic variance of the maximum likelihood estimators are given by $(0.067, 0.2020, 2.9667, 4.0868, 1.5366, 0.1966, 1.6345)$. Using the estimated variances and covariances, it follows from (11) that $Q_R = 1.08649$ with the corresponding $p-value$ given by 0.298, indicating acceptance of $H_0 : \sigma_\delta^2 = 0$. Moreover, the hypothesis $H_0 : \beta_1 = \beta_2$ is rejected since, from (12), the Wald statistics $Q_w = 41.88$, which clearly indicates rejection of $H_0 : \beta_1 - \beta_2 = 0$. Clearly, $H_0$ is also rejected in favor of the one side hypothesis $H_0 : \beta_1 < \beta_2$, indicating that toothbrush one presents greater reduction of the dental plaque index.

## References

Aoki, R., Bolfarine, H. and Singer, J.M. (2000). Null intercept measurement error models.*Tecnical Report. IME-USP.*

Chan, L.K. and Mak, T.K. (1979). On the Maximun Likelihood Estimation of a Linear Structural Relationship when the Intercept is Known. *Journal of Multivariate Analyis*, **9**, 304-313.

Patefield, W.M. (1985). Information from the Maximized Likelihood Function, *Biometrics*, **72,** 3, 664-668.

Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: an Introduction with Applications.* London: Chapman & Hall.

Singer, J.M. and Andrade, D.F. (1997). Regression Models for the Analysis of Pretest, Posttest Data, *Biometrics*, **53,** 3, 729-735.

# Pattern mixture models for dropout in multi-spell multi-state labour market panel data

R. Crouchley[1] and G. Oskrochi[1]

[1] Centre for Applied Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF

**Abstract:** In a panel study each respondent is interviewed at successive waves. Panel studies are rarely complete because of dropout. Little proposed the use of pattern mixture models for repeated measures data. This approach can be used to test for the nature of the dropout process in panel data. Our version of the test is illustrated in a random effects competing risk model of labour market behaviour of individuals from the British Household Panel Survey.

## 1 Background

Panel studies are a rich source of data for the analysis of labour market behaviour. Little and Rubin (1987) introduced a taxonomy of missing data mechanisms which can be used with panel data; their taxonomy distinguishes between missing completely at random (MCAR), missing at random (MAR) and non-ignorable missing (NIM). In the analysis of labour market behaviour it is usually assumed that the stochastic process determining labour market behaviour and participation in the panel are independent (MCAR or MAR). If this assumption is correct then dropout from the panel before the end of the survey can be treated as independent right censoring. We propose a simple pattern mixture approach to detect the missing data mechanism in our multi-spell, multi-state, random effects competing risk model of labour market behaviour.

To illustrate the impact of dropout consider the behaviour of the unemployed. If those unemployed panel participants who have a relatively high probability of finding a job also have a higher probability of dropping out of the panel (for example, they may move to another area for work and thus be difficult to track) we will have non-ignorable missing data (NIM). In this situation any assumption of independent right censoring will create an underestimate of the hazard of becoming re-employed.

Little (1993) also classified the models that can detect the nature of the missing data mechanism into two categories: selection models and pattern-mixture models. The difference depends on how the joint distribution of

the labour market process and the missing data mechanism is factored. In this paper we concentrate on pattern mixture models.

## 2 Labour market behaviour

The first of seven annual British Household Panel Survey (BHPS) interviews was conducted during the autumn of 1991. Our analysis is based on the subsample of men and women aged 16 or over in 1991 who provided information on a set of labour market variables for the first and some subsequent waves. The between-wave data was obtained retrospectively. At wave 1, the sample contained 4,400 men and 5,081 women. By wave 7, the sample had been reduced to 3012 men and 3686 women.

### 2.1 Labour market model

Labour market behaviour is characterised as a three-state-origin and three-state-destination competing risk model. The states consist of employment ($e$) and unemployment ($un$) for those actively seeking work, and the state 'out of the labour market' ($out$) for those who are not actively seeking work. We have undertaken separate analyses of men and women, as they are likely to have different transition behaviours which cannot be captured by a simple indicator. For instance, women tend to have lower labour force attachment with interrupted careers because of child-rearing.
An individual's ($i$) work history can be considered as a series of consecutive time periods (calendar months). This is constructed from the presample periods and the sample periods. The first sample point (wave 1) typically interrupts an ongoing spell of labour market activity. We denote the length of this part of the event history as $U_i + 1$, and calendar time takes the values $t = -U_i, -U_i + 1, ..., 0$. The wave 1 and post-sample data periods are of length $T_i$, where $t = 1, 2..., T_i$, here calendar time starts with the wave 1 observation and ends at $T_i$ with dropout, or at the last sample point of the panel (wave 7). We suppose that an individual can leave labour market state $j$ for one of $k = 1, 2, 3$ destination states. The labels for the origin and destination states are the same so that we also have $j = 1, 2, 3$. The movement between states in a work history can be represented by a binary event history vector $\mathbf{y}_i^{U_i+T_i+1} = [\mathbf{y}_{i,-U_i}, \mathbf{y}_{i,-U_i+1}, ..., \mathbf{y}_{i,0}, \mathbf{y}_{i,1}, ..., \mathbf{y}_{i,T_i}]$, where $\mathbf{y}_{i,t} = [y_{ijkt}]$, such that

$$y_{ijkt} = \begin{cases} 1 & \text{if individual } i \text{ makes a } j \text{ to } k \text{ transition in time period } t \\ 0 & \text{otherwise.} \end{cases}$$

The probability of a $jk$ event for individual $i$ at time $t$ , which has lasted for duration $d$ is a discrete time (an extension of the Prentice & Gloeckler, 1976 model), random effect competing risk model where

$$h_{ijkt}(d) = 1 - \exp[-\exp(\gamma_{jk}\mathbf{I}_{jk}(d) + \beta_{jk}\mathbf{x}_{ijkt} + v_{ijk})],$$

where $\mathbf{I}_{jk}(d)$ is a vector of duration indicators with elements of the form $I(r < d \leq m)$ with $r$ and $m$ being the start and end points of an interval, $\mathbf{x}_{ijkt}$ is a vector of structural variables describing individual characteristics and contextual variables, $(\gamma_{jk}, \beta_{jk})$ are flow-specific vectors of structural parameters, $v_{ijk}$ is an individual-flow-specific random effect or incidental parameter which summarises the effects of the unmeasured, unobservable and excluded variables. The presence of unobserved explanatory variables is likely to be the norm in labour market data, see e.g. Flinn and Heckman (1982). In a work history, the origin of each labour market episode is given by the destination of the previous episode. This implies that we can write the joint likelihood of all transitions over the work history $(t = -U_i, -U_i + 1, ..., T_i)$ conditional on the initial state $(s_{0i})$ and the probabilities $(\mathbf{h}_i)$ as

$$L\left(\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{h}_i, s_{0i}\right)$$

$$= \prod_{t=-U_i}^{T_i} \prod_j \left[\prod_k \left[h_{ijkt}(d)^{y_{ijkt}} \left(1 - h_{ijkt}(d)\right)^{(1-\sum_l y_{ijlt})}\right]\right]^{\delta_{ijt}}$$

where $\delta_{ijt} = 1$ if a spell starts in state $j$ and equals zero otherwise. The evaluation of this likelihood requires that we deal with the relationship between $s_{0i}$ and the random effects $v_{ijk}$. This raises the problem of initial conditions.

## 2.2 Initial Conditions

The development of a process after the initial condition depends solely upon the current and future values of the covariates. We use the likelihood, conditional on the presample data from the interrupted spell, which takes the form

$$L\left(\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{y}_i^{U_i+1}, s_{0i}\right) = \frac{\int_{\mathbf{v}} \Pr\left(s_{0i} \mid \mathbf{v}\right) L\left(\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{h}_i, s_{0i}\right) DG\left(\mathbf{v}\right)}{\int_{\mathbf{v}} \Pr\left(s_{0i} \mid \mathbf{v}\right) L\left(\mathbf{y}_i^{U_i+1} \mid \mathbf{h}_i, s_{0i}\right) DG\left(\mathbf{v}\right)},$$

where $G(v)$ is the distribution over the unobserved covariates. With this likelihood, the impact of the presample (retrospective) information on the inference is reduced. For instance, in the absence of random effects, the presample contribution to the likelihood cancels and we get

$$\prod_{t=1}^{T_i} \prod_j \left[\prod_k \left[h_{ijkt}(d)^{y_{ijkt}(d)} (1 - h_{ijkt}(d))^{(1-\sum_l y_{ijlt}(d))}\right]\right]^{\delta_{ijt}}.$$

We use a multinomial logit model for the initial condition, i.e.

$$\Pr\left(s_{0i}=j \mid \mathbf{v}\right)=\frac{\exp\left(\beta_j \mathbf{x}_{is}+v_{ij}\right)}{1+\exp\left(\beta_j \mathbf{x}_{is}+v_{ij}\right)+\exp\left(\beta_k \mathbf{x}_{is}+v_{ik}\right)}.$$

With the random effects we use the one-factor decomposition of Heckman and Borjas (1980) to simplify the multivariate integral, i.e. $v_{ijk}=c_{jk}v_i$, $v_{ij}=c_j v_i$, where the $c_{jk}, c_j$ are the factor loading coefficients. We also assume that the random effects $(v_i)$ are $N(0,\sigma^2)$. Consequently the multidimensional integral simplifies and univariate Gaussian quadrature can be used.

## 3 Pattern Mixture Models

Park and Davies (1993) suggested stratifying the sample according to the missing data pattern, fitting the models of interest within each stratum and then testing homogeneity of the model parameter estimates across strata. If the model results are significantly different across strata, then the missing data mechanism should not be ignored. Let $M$ be the number of distinct missing data patterns in the data (in our case this is 7) and let $S_m$ be a set of observations with missing data pattern $m$ . For convenience let $S_0$ be the set of complete observations, i.e. without any missing data. Let $D_i$ be the missing data indicator for the $i^{th}$ subject, indexing missing data patterns. The pattern mixture models assume that the likelihood of $\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{y}_i^{U_i+1}, s_{0i}, D_i=m$,

$$L\left(\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{y}_i^{U_i+1}, s_{0i}, D_i=m\right)$$

$$=\frac{\int_{\mathbf{v}} \Pr^m\left(s_{0i} \mid \mathbf{v}\right) L^m\left(\mathbf{y}_i^{U_i+T_i+1} \mid \mathbf{h}_i, s_{0i}\right) DG^m\left(\mathbf{v}\right)}{\int_{\mathbf{v}} \Pr^m\left(s_{0i} \mid \mathbf{v}\right) L^m\left(\mathbf{y}_i^{U_i+1} \mid \mathbf{h}_i, s_{0i}\right) DG^m\left(\mathbf{v}\right)}.$$

We assume for each $m$ that $G^m=G$, and $\beta_j^m=\beta_j$, while the other parameters can vary with $m$. For subject $i$ having the $m^{th}$ missing data pattern we have the basic model

$$h_{ijkt}^m=1-\exp[-\exp(\gamma_{jk}\mathbf{I}_{jk}\left(d\right)+\beta_{jk}\mathbf{x}_{ijkt}+\theta_{jk}\mathbf{I}_{jk}\left(m\right)+v_{ijk})].$$

where $\mathbf{I}_{jk}\left(m\right)$ is an indicator variable for the $m^{th}$ missing data pattern. If $\theta_{jk} \neq \mathbf{0}$, the missing data mechanism is not MCAR.

## 4 Results

We specifically examined the main effect of $\mathbf{I}_{jk}(m)$ and its interactions with ethnicity, tenure and duration. We found that the effect of $\mathbf{I}_{jk}(m)$ was very significant for the *un* to *e* flows for both males and females (i.e. dropout is not MCAR). The interaction of $\mathbf{I}_{jk}(m)$ with ethnicity or tenure was not significant, but was for some of the duration indicators. The significance of $\mathbf{I}_{jk}(m)$ interactions with duration (the response variable) is particularly worrying. The MCAR assumption made by many social scientists who use the BHPS data for labour market research clearly does not hold. Our results challenge the validity of substantive conclusions drawn from analyses which do not allow for the potentially informative nature of the dropout mechanism.

### References

Flinn, C. & Heckman, J.J. (1982). Models for the analysis of labour force dynamics, in Basmann, R., & Rhodes, G. (eds.). *Advances in Econometrics*, Vol 2, JAI Press, Greenwich, Conn, 225-23.

Heckman, J.J. & Borjas, G.J. (1980). Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence, *Economica*, 47, 247-283.

Little, R. J. (1993). Pattern mixture models for multivariate incomplete data, *J. Amer. Stat. Assoc,* 88, 125-134.

Little, R. J. & Rubin, D. (1987). *Statistical analysis with missing data*, New York, Wiley.

Park, T. & Davies, C.S. (1993). A test of missing data mechanism for repeated categorical data, *Biometrics*, 49, 631-638.

Prentice, R.L. & Gloeckler, L.A. (1978). Regression analysis of grouped survival data with application to breast cancer data, *Biometrics*, 34, 57-67.

**Acknowledgements:** This research was carried out using ESRC research grant R000237522 awarded to Dr. R. Crouchley, though any errors are the sole responsibility of the authors.

# Location and Scale models with Type I Censored Data

Gabriela Damilano[1] and Pedro Puig[1]

[1] Unitat d'Estadística, Dep. de Matemàtiques, Universitat Autònoma de Barcelona, 08193 Cerdanyola (Barcelona), Spain.

**Abstract:** Some examples of application of Location and Scale models with Type I Censored Data are presented and the estimation of the parameters based on likelihood is analysed. In these examples the sample sizes are very small and the usual procedures for inference based on the asymptotic distribution of the statistics don't work properly. Higher-order asymptotic methods are developed and the performance is investigated by Monte Carlo experiments.

**Keywords:** Location and Scale Models; Type I Censored Data; Saddlepoint Approximation.

## 1 Introduction

Nowadays, there exist a social conscience that the use of animals in scientific experiments must be adequately controlled and reduced . This position is enforced by the ethical committees who authorize or deny these kind of experiments. As a consequence of this situation the statistician sometimes has to work with very small sample sizes.
The problem becomes grave when data is also censored. Inference with censored data has been studied before by many authors (see Cohen, 1991) but essentially they provide methods only valid for large samples. Many of the procedures used in statistical inference are based on the knowledge of the asymptotic distribution of the statistics involved. For instance, it is extensively used that the maximum likelihood estimator is asymptotically normal distributed or the likelihood ratio test statistic, under null hypothesis, follows a chi-squared distribution. However this information is not very useful when the sample sizes are very small. On the other hand, the exact distribution of these statistics is intractable in most real problems.
This lets to improve the classical asymptotic approximations in order to adapt them to smaller samples. A way to do it is to use the higher-order asymptotic methodology. In this communication we present Saddlepoint approximations in order to do inference with one and two samples for some location and scale models which have censored data of type I.

## 2 Location and Scale models and Censoring

Let be $X_1, X_2, ..., X_n$ independent and identically distributed continuous random variables belonging to a location and scale family, that is, its density function has the form

$$f(x;\mu,\sigma) = \frac{1}{\sigma}p(\frac{x-\mu}{\sigma}) \ , -\infty < x < \infty, \tag{1}$$

where $\mu \in \Re$ and $\sigma \in \Re^+$. Its distribution function can be expressed as $F(x;\mu,\sigma) = F((x-\mu)/\sigma)$ where $F(x) = \int_{-\infty}^{x} p(t)dt$.
Suppose that the i-th observation is a realization of the random variable $(Y_i, \varepsilon_i)$ with $Y_i = \min(X_i, c_i)$ where $c_1, c_2, ..., c_n$ are known constants and $\varepsilon_i = I_{[0,c_i]}(X_i)$ where $I_\Omega(.)$ denotes de indicator function of the set $\Omega$. This observational pattern is called type I censoring and is very important in survival analysis, reliability theory, quality control, etc.(see Lawless, 1982).

**Example 1:** The following values correspond to determinations of the level of glucose in the blood of 9 mice with an experimental induced diabetes:

592, 544, 466, 600+, 600+, 600+, 443, 524, 600+

The measures have been done with the "Glucometer Elite" of Bayer that only appreciates quantities below 600. Observe that there are 4 censored values denoted by 600+, that is, determinations of glucose above 600. For this example $c_1 = c_2 = ... = c_9 = 600$, and the distribution of the data can be assumed as a normal.

For these kind of problems the likelihood function is

$$l(Y;\mu,\sigma) = \sum_{i=1}^{n} \varepsilon_i \log(\frac{1}{\sigma}p(\frac{Y_i-\mu}{\sigma})) + (1-\varepsilon_i)\log(1-F(\frac{Y_i-\mu}{\sigma})) \tag{2}$$

Maximum likelihood estimators can be calculated by solving the likelihood equation, that is, $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma} = 0$. However if the likelihood equation has a solution it is not necessarily unique. The following theorem provides a sufficient condition of uniqueness:

### Theorem

Suppose $y_1, y_2, ..., y_n$ be independent variates of a scale and location family like in (1), with type I censored observations. Suppose that the function $-\log(p(y))$ has continuous and strictly positive second derivative. Then, if the likelihood equation has a solution this is unique and is the maximum likelihood estimate.

**Remark 1:** Burridge (1981) gets the same result but only for non censored data. See also Pace and Salvan (1997) for an easy and detailed exposition.

**Remark 2:** The theorem holds for the normal and extreme value distributions. The laster can be seen as the logarithm of a Weibull variate.

**Remark 3:** It can be showed that for the normal distribution, likelihood equation always has a solution. However it is not obvious and is false for the truncated normal distribution(see Castillo and Puig, 1999).

## 3 Saddlepoint Approximations

Let $\hat{\mu}, \hat{\sigma}$ be the maximum likelihood estimators of (1) and suppose that we are interested in determinate a confidence interval of level $(1-\alpha)$ for the location parameter $\mu$. For non censored data it can be used the pivotal quantity $(\mu - \hat{\mu})/\hat{\sigma}$ in order to calculate an exact interval (see Pace and Salvan, 1997). However for type I censored observations a pivotal quantity is unknown and we have to use approximate methods.
One of the methods is based on the likelihood ratio test statistic

$$W = 2(l(X; \hat{\mu}, \hat{\sigma}) - l(X; \mu, \hat{\sigma}_\mu)), \tag{3}$$

where the likelihood function is given in (2) and $\hat{\sigma}_\mu$ is the maximum likelihood estimator of the scale parameter when $\mu$ is known. The procedure consists in calculate the region determined by the inequality $W \leq \chi^2_{1,\alpha}$ with the right term equal to the $1-\alpha$ quantile of the $\chi^2_1$ distribution.
Monte Carlo studies show, see Table 3.1, that the performance of this procedure is not very good for small samples when data is normally distributed, although it is better than that based on the observed information matrix (OIM).
For this reason we can try to improve the convergence of the log-likelihood ratio test statistic expressed in (3) to its asymptotic distribution, that is, a $\chi^2_1$, which should be re-expressed as $Z = sign(\hat{\mu} - \mu)\sqrt{W}$ with asymptotic normal standard distribution.
We have used a transformation based on the Saddlepoint approximation, the $Z^*$ (see Barndorff-Nielsen, 1991), that is closer to the normal distribution. Our Monte Carlo simulations show that its performance is quite good, even for very small sample size.

Table 3.1:Estimated coverage levels of confidence intervals for $\mu$ based on OIM, W and Z*. Values (in %) based on 5000 trials for N(0,1)

| | Nominal Level | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 90 | | | 95 | | | 99 | | |
| N | OIM | W | Z* | OIM | W | Z* | OIM | W | Z* |
| 15 | 87.6 | 88.8 | 90.4 | 93.1 | 93.9 | 95.1 | 97.8 | 98.5 | 99.1 |
| 10 | 85.9 | 87.4 | 90.3 | 91.7 | 92.9 | 95.0 | 96.8 | 98.4 | 99.1 |
| 8 | 84.7 | 86.7 | 90.1 | 90.1 | 92.2 | 95.3 | 95.9 | 98.4 | 99.3 |
| 6 | 82.5 | 85.0 | 90.5 | 87.8 | 91.3 | 95.8 | 94.4 | 97.7 | 99.2 |

This transformation is of the form

$$Z^* = Z + \frac{1}{Z} \log\left(\frac{CU_p}{Z}\right), \tag{4}$$

where $C$ and $U_p$ must be computed for each distributional model as follows:

$$C = \frac{\tilde{l}_{\sigma\hat{\sigma}}}{\sqrt{\tilde{l}_{\sigma\sigma}\hat{l}_{\sigma\sigma}}} \qquad U_p = -\frac{\left[\tilde{l}_{\hat{\mu}} - \hat{l}_{\hat{\mu}} - \frac{\bar{l}_{\sigma\hat{\mu}}}{\bar{l}_{\sigma\hat{\sigma}}}\left(\tilde{l}_{\hat{\sigma}} - \hat{l}_{\hat{\sigma}}\right)\right]}{\sqrt{\frac{\hat{l}^2_{\sigma\mu}}{\hat{l}_{\sigma\sigma}} - \hat{l}_{\mu\mu}}} \tag{5}$$

where $\hat{l}$ and $\tilde{l}$ are the derivatives of (3) expressed as a function of $\mu$,$\hat{\mu}$,$\sigma$,$\hat{\sigma}$, evaluated respectively at $(\hat{\mu}, \hat{\sigma})$ and $(\mu, \hat{\sigma}_\mu)$. In the particular case of the normal distribution, $l\left(\mu, \sigma, \hat{\mu}, \hat{\sigma}\right)$ may be expressed as follows:

$$n \log \frac{1}{\sigma} + r \log \Phi\left(1 - \xi\right) - \frac{n}{2\sigma^2}\left[\left(\hat{\mu} - \mu\right)^2 + \hat{\sigma}^2 - \hat{\sigma}\Omega\left(h, -\hat{\xi}\right)\left(\hat{\mu} - 2\mu + x_0\right)\right]$$

where n and r are the number of observed and censored data, $\xi = \left(x_0 - \mu\right)/\sigma$, $\hat{\xi} = \left(x_0 - \hat{\mu}\right)/\hat{\sigma}$, $x_0$ is the censoring point, $\Omega\left(h, z\right) = \left(r/n\right)\phi\left(z\right)/\Phi\left(z\right)$ and $h = r/n$.

### 3.1 Confidence Intervals for the location parameter

Nowadays, there are statistical packages like SAS, Stata, EVIEWS, to mention some of them, that can be used to calculate the maximum likelihood estimates (MLE) for censored data for several distributional patterns. Assuming normality, we have developed a program in Mathematica, which code is presented below, that it calculates the confidence intervals for the mean based on W and Z*. It only needs to run the number of observed (n) and censored (r) values, the censoring point $(x_0)$ and the MLE.

```
n=5;r=4;x0=600;mu=582.815;si=93.994;si2=si^2;vcj=3.84;vcn=1.96;
Lv[m_,s_,me_,se_]:=-n*Log[s]+r*Log[(1-Erf[((x0-m)/s)/2^(1/2)])/2]-
  n/(2*s^2)*((me-m)^2+se^2+(-se*((2/Pi)^(1/2)*r)/(E^((me-x0)^2/
  (2*se^2))*(n + n*Erf[(me-x0)/(2^(1/2)*se)]))*(me+x0-2*m));

lse[m_,s_,me_,se_]=D[Lv[m,s,me,se],se]; ls[m_,s_]=D[Lv[m,s,mu,si],s];
lme[m_,s_,me_,se_]=D[Lv[m,s,me,se],me];
lsm[m_,s_]=D[Lv[m,s,mu,si],s,m];
ls1[m_,s_,me_,se_]=D[Lv[m,s,me,se],s,me];l2m[m_,s_]=D[Lv[m,s,mu,si],{m,2}];
ls2[m_,s_,me_,se_]=D[Lv[m,s,me,se],s,se];l2s[m_,s_]=D[Lv[m,s,mu,si],{s,2}];
sm[m_]:=Module[{a},a=FindRoot[ls[m,s]==0,{s,si}];s/.a];
w[m_]:=2*(Lv[mu,si,mu,si]-Lv[m,sm[m],mu,si]);
z[m_]:=Sqrt[2*(Lv[mu,si,mu,si]-Lv[m,sm[m],mu,si])];
dss[m_]:=l2s[m,sm[m]];ds2[m_]:=ls2[m,sm[m],mu,si];ds1[m_]:=ls1[m,sm[m],mu,si];
dme[m_]:=lme[m,sm[m],mu,si];dse[m_]:=lse[m,sm[m],mu,si];
cup[m_]:=(ds2[m]/Sqrt[dss[m]*l2s[mu,si]])*(-1/Sqrt[-(l2m[mu,si]-lsm[mu,si]^2/
 l2s[mu,si])])*(dme[m]-lme[mu,si,mu,si]-ds1[m]*(dse[m]-lse[mu,si,mu,si])/ds2[m]);
zs[m_]:=z[m]^2+2*Log[Abs[(cup[m])/z[m]]]+(1/z[m])^2*Log[Abs[(cup[m])/z[m]]]^2;

simu=1/Sqrt[-(l2m[mu,si]-lsm[mu,si]^2/l2s[mu,si])];
LNI=mu-simu*vcn;LNS=mu+simu*vcn;
LWS=Module[{a},a=FindRoot[w[m]==vcj,{m,{mu,LNS}}];m/.a]
LWI=Module[{a},a=FindRoot[w[m]==vcj,{m,{LNI,mu}}];m/.a]
LS=Module[{a},a=FindRoot[zs[m]==vcj,{m,{mu+0.1,LNS}}];m/.a]
LI=Module[{a},a=FindRoot[zs[m]==vcj,{m,{LNI,mu-0.1}}];m/.a]
```

**Example 2: a)** For the data in example 1, we have n=5, r=4, $x_0$=600, the MLE are $\hat{\mu}$=582.815 and $\hat{\sigma}$=93.994. With only this information the program gives us the 95% confidence intervals for the mean based in W (513.897,698.748) and Z* (503.315,758.613). Notice that the interval based on Saddlepoint Approximation is longer than that obtained with W. It happens because the last one doesn't have the wished level of coverage. This difference almost disappears when the sample size is large, as we can see in the next example.

**b)** Consider the data presented by Nelson and Schmee (1979), about the life spans of n=37 failures corresponding to N=96 electronic locomotive controls occurred prior to termination of the test at the fixed point of $x_0$=2.1303. For these data, Cohen(1991) obtains the MLE $\hat{\mu} = 2.224$, $\hat{\sigma} = 0.307$ and the approximate 95% confidence interval for $\mu$ based in the Fisher information matrix as (2.134,2.314). With our program we get (2.146,2.330) for W and (2.149,2.339) for Z*. The resulting intervals are very close each other. This example is also considered by Lawless (1982).

### 3.2 Two sample problems

One of the applications of the confidence interval calculations presented above is for comparing the means of two matched samples. This kind of situation have a special feature, because the sample of the differences could produce more than one censoring point.

**Example 3:** This data corresponds to the time of permanency (in seconds) over a Rotorod "Treadmill for mice 7600, Ugo Basile, Italy" for 8 mice. In this experiment, used to study the neurotoxicity of acrylamide, the animals are first induced by 20 mg/Kg of the substance producing a performance decrement, and later they receive a corrective treatment. The values are

Sample I: 10,3,29,1,3,1,3,3 Sample II: 88,115,120+,3,2,2, 4,120+.
Sample of differences (II-I): 78, 112, 91+, 2, -1, 1, 1, 117+.

The value 120+ indicates the censoring point because the mouse is removed from the rotorod after 120 seconds if it doesn't fall before. For the sample of the differences the censoring points are 91+ and 117+. In order to compare the means by using the saddlepoint approximation, we have to consider a new $l\,(\mu, \sigma, \hat{\mu}, \hat{\sigma})$ that now takes the form:

$$n \log \frac{1}{\sigma} + \sum_{j=1}^{k} r_j \log[1 - \Phi\,(\xi_j)] - \frac{n}{2\sigma^2}[(\hat{\mu} - \mu)^2 + \hat{\sigma}^2 - 2\hat{\sigma} \sum_{j=1}^{k} \hat{\Omega}_j\,(\hat{\mu} - \mu) - \hat{\sigma}^2 \sum_{j=1}^{k} \hat{\xi}_j \hat{\Omega}_j]$$

where k indicates the number of different patterns of censoring, $r_j$ the number of censored data in the jth pattern, and $\Omega_j$ , $\hat{\Omega}_j$ are the reduced expressions of the functions defined in section 3.1.
In order to perform this test the same program listed above could be used only changing the first block by,

```
n=6;r1=1;r2=1;x01=91;x02=117;mu=59.633;si=63.929;si2=si^2;
Omega[h_,x_]:=h*Exp[-x^2/2]/(Sqrt[2*Pi]*((1 + Erf[x/2^(1/2)])/2));
h1=r1/n;h2=r2/n;
Lv[m_,s_,me_,se_]:=-n*Log[s]+r1*Log[(1-Erf[((x01-m)/s)/2^(1/2)])/2]+
 r2*Log[(1-Erf[((x02-m)/s)/2^(1/2)])/2]-(n/(2*s^2))*((me-m)^2+se^2-
 se*(Omega[h1,(me-x01)/se]+Omega[h2,(me-x02)/se])*(me-m)-
 se*((me-x01)*Omega[h1,(me-x01)/se]+(me-x02)*Omega[h2,(me-x02)/se]));
```

and replacing the third block by the computation of zs[0] and w[0]. From these values the p-values can be computed by using the $\chi^2_1$ distribution. For our example that is w[0]=3.5978 with p-value=0.058 and zs[0]=1.619 with p-value=0.203. As it can be seen, with the W-method one could reject the equality of means, while with the Z* statistic clearly we couldn't.
This fact, is very important because if we don't use the appropriate methods of inference for small samples with censored data, we may arrive to erroneous conclusions. We have also analyzed the tests of hypothesis about the mean, for two independent normally distributed samples with censoring of type I, supposing that the variances are equal and different (problem of Behrens-Fisher). Once more, transformations like in (4) provide good tests for small sample sizes.

**Acknowledgements:** We thanks to Drs. F. Bosch and P. Otaegui of the Biochemistry and Molecular Biology Dep. of UAB, as well as to Dr. J. Guerrero of Animal Physiology Dep. of UAB, for provide us their data sets.

## References

Barndorff-Nielsen, O. (1991). Modified signed log likelihood ratio. *Biometrika*, **78**, 557-563.

Burridge, J. (1981). A note on maximum likelihood estimator regression models using grouped data. *Journal of the Royal Statistical Society, B* , **43**, p. 41-45.

Castillo, J. and Puig, P. (1999). Invariant Exponential Models Applied to Reliability Theory and Survival Analysis. *Journal of the American Statistical Association* , **94**, p. 522-528.

Cohen, C.A. (1991). *Truncated and Censored Samples. Theory and Applications.* Marcel Dekker. New York.

Lawless, J.F. (1982). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons. New York.

Nelson, W.B. and Schmee, J. (1979). Inference for (log) normal life distributions from small singly censored samples and blue's. *Technometrics*, **21**, p. 43-54. John Wiley & Sons. New York.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference.* World Scientific Press. Singapore.

# Approximating Nonlinear Models

Tamraparni Dasu[1], Theodore Johnson[2]

[1] tamr@research.att.com, AT&T Labs Research, A265, 180 Park Avenue, Florham Park, NJ 07932
[2] johnsont@research.att.com, AT&T Labs - Research, A111, 180 Park Avenue, Florham Park, NJ 07932

**Abstract:** Fitting nonlinear models to large data sets is computationally expensive even when the choice of models (e.g. log-linear) or distributional assumptions (exponential family) is clear. Computing the model parameters (e.g. maximum likelihood estimates) requires several iterations over all the data, using expensive algorithms such as Newton Raphson or EM. The problem becomes much harder when there is no prior knowledge to facilitate distributional assumptions or model selection. Ad hoc exploratory methods are an unsatisfactory option. To address this issue, we propose a new, widely applicable method of constructing nonlinear models through piecewise fitting of nonparametric models. We create special partitions of the covariate space and fit a nonparametric model within each class of the partition. Nonparametric models are free of distributional and model assumptions (wide applicability), inexpensive (require one or two passes over the data) and often can be computed from summaries such as sufficient statistics, reducing the data storage needs considerably. Furthermore, the nonparametric models can often be leveraged to identify a suitable a set of parametric models that can be built using a small sample of the data. We have used the last aspect with great success in modeling proprietary telecommunications data.
In this paper, we demonstrate the technique in the context of survival analysis. Using simulated data, we illustrate the nonparametric modeling of the effect of covariates on the survival function, while highlighting the abovementioned advantages of such an approach. We show that the technique is several magnitudes faster (with some loss of accuracy) than the parametric models even when the latter are applicable.

**Keywords:** Nonparametric Approximations; Large Datasets; Computational Speed; DataSphere Partition; Survival Analysis.

## 1 Introduction

Consider the case where we wish to estimate the effect of the vector of *covariates* $\vec{x}$ on the *response variable* $Y$, represented by $g(Y) = f(\vec{x})$, where $g$ and $f$ are functions. Parametric methods assume that $f(\vec{x})$ has a convenient reduced representation such as

$$g(Y) = f(\vec{x}, \vec{\beta})$$

where $g$ and $f$ are known functions and $\vec{\beta}$ is a vector of *parameters*. The problem is reduced to estimating just the parameters $\vec{\beta}$. Log-linear models are examples of parametric models. See McCullagh and Nelder (1989). If the assumptions are true, the fitted parametric models can be powerful, accurate and can be fitted using very little data. However, the assumptions can restrict the applicability of such models. In addition, distributional assumptions about the error residuals are required for statistical inference and hypothesis testing. Verifying the validity of assumptions can be expensive and time consuming. Furthermore, estimating the parameter vector $\vec{\beta}$ often involves several iterations over the data, requiring long computation times. If we have no prior knowledge of either $f$ or the distribution of $Y$, we cannot use parametric methods at all.

Exploratory analysis based on sampling is an option if the data are homogeneous or if we have some knowledge as to how to sample. Otherwise, sampling can be misleading and can exclude rare but interesting instances. Furthermore, when the dataset is massive, sampling itself can be expensive. Worse, the exploratory analysis might not reveal any structure that aligns with known parametric models. In fact, large data sets tend to be heterogeneous and complex. The inter-relationships can seldom be captured through a compact, closed form expression as posited by parametric models.

To address the abovementioned issues, we propose a widely applicable method of fitting a collection of piecewise nonparametric models over a partition of the covariate space. We start by choosing a *data-driven* partition of the covariate space that results in homogeneous regions. This is in contrast to *stratification*, which is a pre-defined partition based usually upon values of categorical attributes or an a priori discretization of numerical attributes into intervals. While any partition can be used, partitions that are easy to compute and whose size (number of classes) grows manageably with the number of attributes is desirable. Rectilinear partitions (defined by dividing each attribute into intervals as in a grid) grow exponentially in size with the number of attributes. A mere six attributes, each divided into just ten intervals will result in $10^6$ classes! Other methods that induce multidimensional partitions (clustering, classification) can be computationally expensive, besides being objective specific.

An ideal general purpose partitioning technique is the fast, scalable space partitioning scheme called *DataSphere (DS) Partitioning* based upon multivariate distance contours, proposed in our previous work (Dasu et al. (1997)). See Fig 1(a). In order to create a DS partition of a data set, we standardize each attribute separately using an appropriate center (e.g. mean) and scaling parameter (e.g. standard deviation). Next, we compute the distance (e.g. Euclidean) of every point from the multivariate center (e.g. multivariate mean or dimensionwise median) and divide the data into concentric *distance layers* containing roughly equal mass, much like the layers of an onion. The layer boundaries are quantiles of the distribution of

the distances from the center of the data points. The partition is made finer using *directional pyramids* (see Berchtold et al. (1998)), which identify the direction (attribute) in which the data point $x$ has the maximum deviation from the expected value. A data point $x = (x_1, x_2, \ldots, x_i, \ldots, x_d)$ lies in the pyramid corresponding to attribute $X_i^+$ if

1. $x_i$ is above average i.e.

$$x_i > \overline{x_i}$$

   and

2. the standardized deviation of $x_i$ from the mean is the largest among all the $d$ components of the data point $x$ i.e.

$$\mid \frac{x_i - \overline{x_i}}{\sigma_{x_i}} \mid > \mid \frac{x_j - \overline{x_j}}{\sigma_{x_j}} \mid \forall j \neq i.$$

where $\overline{x_i}$ and $\sigma_{x_i}$ are the sample mean and standard deviation respectively of the $i^{th}$ attribute $X_i$ of the data set. The $X_i^-$ pyramid is defined in a similar fashion. Note that there are two pyramids corresponding to every attribute, one for the points that lie above the average, the other for those that lie below average. In the 2-D example in Figure 1(a), there are four pyramids $X^+, X^-, Y^+, Y^-$ corresponding to the two attributes $X$ and $Y$. The DS partition can be created in **linear time with just one pass** over the data, with $2dl$ classes where $d$ is the number of attributes and $l$ is the number of distance layers. Note that the size of the partition scales linearly with the number of attributes. See Johnson et al. (1999) for details. The choice of the nonparametric model depends on the problem at hand. As mentioned in earlier, the models should be aggregable to enable the computation of a minimal partition. In this paper, we illustrate the technique in the context of survival analysis, where we wish to estimate the effect of covariates $\vec{x}$ on the time to failure $T$, using the *survival function*, $S(t)$=Probability$(T > t)$, represented by $S(t; \vec{x})$ to reflect the influence of the covariates. See Cox et al. (1984) for definitions. For example, how do blood pressure, weight, gender and age effect the time to death ($T$) due to cardiac arrest, measured from the first detection of coronary heart disease? We consider three instances using simple examples to facilitate a clear explanation.

### 1.1 Three Scenarios: Simulation Results

**(1) $S(t; \vec{x})$ has no known reduced representation:**
In this situation, parametric methods are not applicable. We use an empirical nonparametric estimator, the *product limit estimator* of the survival function $\hat{S}$ (see Cox et al. (1984)) to build estimates of $S(t; \vec{x})$ within each class of the partition of the covariate space.

We generated a data set of 200,000 observations with three covariates $(x, y, z)$ whose distance from the center is proportional to the response time. We created a DataSphere partition with 3 concentric shells based on distance from the center (the dimensionwise median), each shell containing roughly equal mass. Fig 1(b) below shows the survival curves estimated using the nonparametric product limit estimator for the data in each of the three shells. As we move from the innermost shell of the covariate space to to the outer most shell, the survival curves have longer tails indicating longer survival times for values of the covariates that are farther from the median.

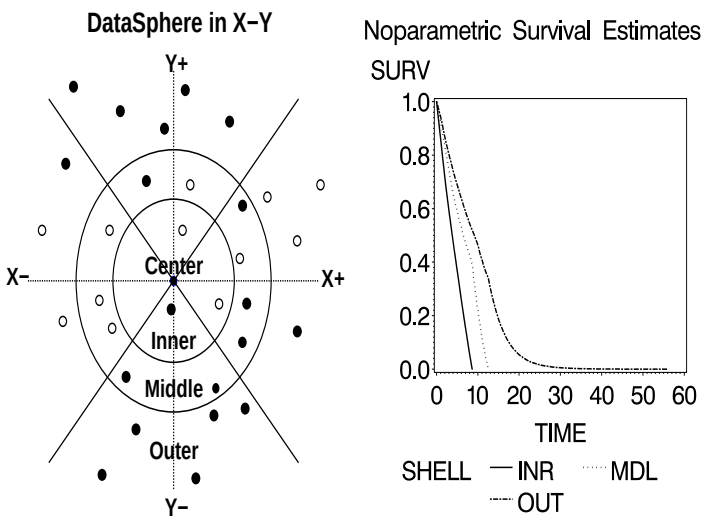



FIGURE 1. a) DataSphere b)Approximation.

**(2) $S(t; \vec{x})$ has a different parametric form in different unknown regions of the attribute space:**
We use the piecewise models to discover and define the regions where $S(t; \vec{x})$ has different forms. We created a data set of 80,000 data points with two covariates $x$ and $y$ and a time to failure response variable $T$, which has an exponential distribution with different parameters in three different **unknown** regions of the $(x, y)$ covariate space. We used a rectilinear partition of approximately 64 classes based on marginal quantiles of $x$ and $y$, and fit a product limit estimator within each class. Note that finding a minimal partition (identifying the smallest set of distinct groups) is a hard problem and that we do not need to identify the regions explicitly for the method we propose. The parametric model on the other hand, suffers in accuracy by being ignorant of the three groups. See Table 1.1 for error estimates based upon deviation from the true value.

**(3) $S(t;\vec{x})$ is completely parametric, known up to the parameters:** In this scenario, only the parameters are unknown and need to be estimated. We generated a dataset of 200,0000 data points with survival function of $T$ given by:

$$S(t;\vec{x}) = e^{-(e^{-(\beta_1 x+\beta_2 y+\beta_3 z)})t},$$

an exponential distribution with the rate parameter $e^{(\beta_1 x+\beta_2 y+\beta_3 z)}$, where $\vec{x} = (x, y, z)$ is the covariate vector and $(\beta_1, \beta_2, \beta_3)$ are the unknown parameters to be estimated.

The parametric models give accurate estimates but take a long time to run. We used SAS software for proportional hazards regression (SAS Technical Report P-229 (1992)) based on the Newton Raphson algorithm. To compute the approximate models, we used a DataSphere partition of the covariate based on the multivariate depth contours. The partition had 60 classes and we fit a product limit estimator within each class. It took just 7 seconds for the approximate models to run, while one iteration of the proportional hazards regression ran for over 34 hours. See Table 1.1. Therefore, even when parametric models are appropriate, it might be much quicker to fit approximate models for exploratory purposes. Sampling is effective (primarily because the data set is homogeneous) but the parametric model still takes a long time to run. The nonparametric models are good approximations in the classes which are close to the center (inner) of the data cloud and degenerate towards the outer classes that usually contain outliers. See Figure 2.

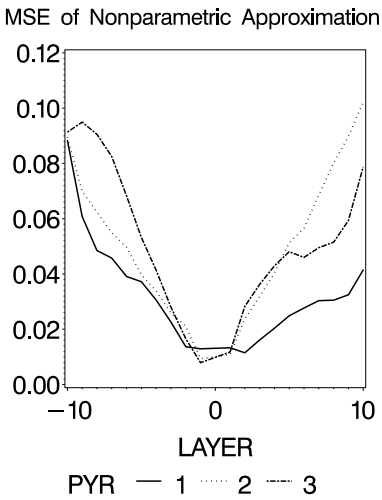



FIGURE 2. Mean Squared Error Measured from the True Value (Scenario 3).

Table 1.1

| SN* | N* | Parametric Model | | | Approximate Model | | |
|---|---|---|---|---|---|---|---|
| | | MSE* | MAE* | Time* | MSE* | MAE* | Time* |
| 2 | 80K | 0.0185 | 0.1026 | 5h57m | 0.0096 | 0.0326 | 1.99s |
| 3 | 200K | 0.0059 | 0.0594 | 34h37m | 0.0446 | 0.1579 | 6.61s |

*(SN=Scenario, N=Number of data points, MSE=Mean Squared Error, MAE=Mean Absolute Error, Time = Running time in hours(h), minutes(m) and seconds(s)).

## 2 Conclusions and Further Research

We have proposed a new, widely applicable method for constructing approximate nonparametric models to capture nonlinear relationships among attributes for large data sets. The method is free of distributional and model assumptions, computationally fast and inexpensive in terms of data storage. Note that the nonparametric estimates can be computed, compared and combined using aggregates (number of failures, censored events, initial risk set) without requiring the raw data. This aspect is important for determining a minimal partition. While there is a loss of information in the proposed approach, there is a tremendous gain in computation time and memory requirements, often making it the only feasible choice for fitting nonlinear models to large data sets.
Further research includes other approaches such as iterative scaling, adaptive triangulation and alternative space partitioning schemes.

### References

Berchtold, S., Bohm, C., and Kriegel, H. (1998). The Pyramid_Tree: Breaking the Curse of Dimensionality. In *ACM SIGMOD*.

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Dasu, T., and Johnson, T. (1997). An Efficient Method for Representing, Analyzing and Visualizing Massive, High Dimensional Data. *Computing Science and Statistics*, **29** 70-75.

Johnson, T., and Dasu, T. (1999). Scalable Data Space Partitioning in High Dimensions. In: *Proceedings of the Statistical Computing Section, American Statistical Association*. (http://www.research.att.com/info/tamr).

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

SAS Institute Inc., SAS Technical Report P-229 (1992). *SAS/STAT Software: Changes and Enhancements, Release 6.07*, Cary, NC: SAS Insitute Inc.

# Statistical Modelling for Matched Tables

Antoine de Falguerolles[1] and Michael Greenacre[2]

[1] Laboratoire de Statistique et Probabilités, UMR C5583, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex, France
[2] Facultat de Ciènces Econòmiques i Empresarials, Universitat Pompeu Fabra, Ramon Trias Fargas, 23-27, E-08005 Barcelona, Spain

**Abstract:** We compare different statistical analyses for pairs of matched two-way tables (same dimensions and same cross-classifying factors). It turns out that seemingly unrelated methods can be unified within the framework of generalised bilinear models. In passing, square table analyses are revisited.

**Keywords:** Generalised linear models, Generalised bilinear models, Contingency table, Correspondence analysis, Square table, Symmetry, Skew-symmetry.

## 1 Introduction

We consider the statistical modelling of a pair of matched two-way tables **A** and **B** where **A** and **B** have the same dimensions. A set of two two-way contingency tables cross-classified by the same two factors ($I$ and $J$) and stratified according to gender ($H$) typifies the situation that we have in mind.
Such data are often explored with descriptive methods driven by singular value decompositions (SVD), possibly generalised, of some *ad hoc* matrices. For a square table (a mobility table, a transition matrix, a sociomatrix ... ) SVD is separately applied to the symmetric and skew-symmetric parts derived from the Gower (Hermitian) decomposition of square matrices. In this line, a simple correspondence analysis (CA) algorithm and a special coding of the input provide in one stroke the reduced rank approximations for each part (Greenacre, 2000). But this extends also to matched tables.
All these descriptive approaches can be reformulated as bilinear models within the generalised linear modelling framework (McCullagh & Nelder, 1989). This follows from the consideration that CA of a given order can be defined as a reduced rank (or multiplicative) interaction model with a Gaussian distribution, a non-canonical link function and particular prior weights for the entries in the table. Clearly, other distributions and link functions are of interest and give rise to the flexible class of bilinear models (Falguerolles & Francis, 1992, 1995). Note that bilinear models lend themselves to graphical representations in the form of biplots (Gabriel, 1971) but this point of view will not be further pursued.

## 2 Examples

We consider three data sets which exemplify the situations outlined in the introduction: matched tables where one of the table is a standard of comparison, matched tables which have an equal status, and a square table. In all examples we analyse count data. This may appear somewhat restrictive but the method presented in this paper apply just as well to continuous data. In particular, variance stabilising transformations of count data produce continuous data and statistical analyses of the latter often supply adequate surrogate analyses for the former.

### 2.1 Two matched tables with unequal status

The first example is taken from James & Segal (1982). In an analysis of mortality incidence, the counts of death are compared to the counts of population at risk, each table being classified by age category and date of birth category. In this type of situation the matched tables of count data are not on an equal status: the standard is the population at risk. Factors age and date of birth are respectively denoted by $I$ and $J$ and their levels by $i$ and $j$.
It is quite natural to assume that the death counts $y_{ij}^{IJ}$ (table **A**) are observed values of independent binomial random variables $Y_{ij}^{IJ}$ with unknown probabilities $\pi_{ij}^{IJ}$, and fixed parameter $N_{ij}^{IJ}$ (table **B**). Thus $E[Y_{ij}^{IJ}] = N_{ij}^{IJ}\pi_{ij}^{IJ}$. Taking the logit link, two baseline models for the predictor ($\eta_{ij}^{IJ} = \text{logit}(\pi_{ij}^{IJ})$) are of interest, namely the independence and the saturated:

$$\begin{aligned}\eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J \\ \eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ}\end{aligned}$$

The interaction term $\beta_{ij}^{IJ}$ in the saturated model can be constrained in several ways: a cohort effect, a reduced rank multiplicative interaction or both. The latter is exactly what James and Segal (1982) have proposed and is an early example of generalised bilinear models.
Since the $\pi_{ij}^{IJ}$ are small, $\text{logit}(\pi_{ij}^{IJ}) \simeq \log(\pi_{ij}^{IJ})$ and a well known approximation implies that the death counts are observed values of independent Poisson distributed random variables with unknown parameter $\lambda_{ij}^{IJ} \simeq \pi_{ij}^{IJ} N_{ij}^{IJ}$. It then follows that $\log(\lambda_{ij}^{IJ}) \simeq \eta_{ij}^{IJ} + \log(N_{ij}^{IJ})$. In the same spirit, the $\sqrt{y_{ij}^{IJ}}$ can be taken as independent Gaussian observations with means $\mu_{ij}^{IJ} \simeq \sqrt{\pi_{ij}^{IJ} N_{ij}^{IJ}}$. Considering a non-canonical link function for the Gaussian distribution, it follows that $2\log(\mu_{ij}^{IJ}) \simeq \eta_{ij}^{IJ} + \log(N_{ij}^{IJ})$.
Interestingly, the role of $N_{ij}^{IJ}$ or $\log(N_{ij}^{IJ})$ in all three analyses supports the idea that **B** is taken as a standard in the analysis of the set of matched tables **A** and **B**.

### 2.2 Two matched tables with equal status

A second example is taken from van der Heijden & de Leeuw (1985) where a contingency table with data on suicide behaviour is analysed by a complementary application of log-linear modelling and CA. For each sex, the suicides counts are cross classified by age (18 levels) and method (10 levels). Factors sex, age and method are respectively denoted by H, I, J and their levels by h, i, j.
The suicide counts are considered as independent Poisson data with parameters $\lambda_{hij}^{HIJ}$. Assuming the canonical log link ($\log(\lambda_{hij}^{HIJ}) = \eta_{hij}^{HIJ}$), the following baseline hierarchical models are of interest, namely the independence, the all two-way interaction, the saturated:

$$\begin{array}{rcl}
\eta_{hij}^{HIJ} & = & \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J \\
\eta_{hij}^{HIJ} & = & \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \beta_{ij}^{IJ} \\
\eta_{hij}^{HIJ} & = & \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \beta_{ij}^{IJ} + \beta_{hij}^{HIJ}
\end{array}$$

A square root transformation of the counts gives rise to comparable models for Gaussian observations with unknown mean $\mu_{hij}^{HIJ} \simeq \sqrt{\lambda_{hij}^{HIJ}}$, variance equal to 1/4 and non-canonical link function.

### 2.3 A square table

The last example is taken from an article by Stigler (1994) on "citation patterns in the journals of statistics and probability" where his Table 4 gives a square table of cross-citations involving statistics journals. Rows correspond to citing journal (I), columns to cited journal (J) and the entries are the numbers of citations ($y_{ij}^{IJ}$).
The modelling of square tables adresses several substantive questions such as symmetry versus departure from symmetry, inclusion of the diagonal counts versus exclusion, marginal homogeneity versus marginal heterogeneity and so on. Departure from symmetry is usually investigated by fitting quasi-symmetry and interesting bilinear extensions (van der Heijden & Mooijaart, 1995). Among the baseline models are (quasi-) independence, quasi-symmetry, saturated:

$$\begin{array}{rcl}
\eta_{ij}^{IJ} & = & \beta^{\emptyset} + \beta_i^I + \beta_j^J \\
\eta_{ij}^{IJ} & = & \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ} \text{ (with } \beta_{ij}^{IJ} = \beta_{ji}^{IJ}) \\
\eta_{ij}^{IJ} & = & \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ}
\end{array}$$

It turns out that these models can be fitted as particular cases of matched tables by pairing the square table and its transpose. This is known in the literature as the three-dimensional representation of square tables (Bishop, Fienberg & Holland, 1975, page 289).

# 3 Initial analysis by SVD

## 3.1 Preprocessing the tables

The matched tables may have counts which are not directly comparable. A simple procedure for avoiding this discrepancy is to transform both table by iterative proportional fitting into biproportional tables having same marginal totals ($1/I$ and $1/J$). For continuous data, a comparable approach is to double centre the entries. (If needed, specific scale factors can be applied to each table.) Let $\mathbf{A}$ and $\mathbf{B}$ denote the resulting matched data matrices.

## 3.2 SVD analysis of A relative to B

A simple and obvious approach is to perform the SVD of $\mathbf{A} - \mathbf{B}$. In the case of count data, the SVD of $\mathbf{F}$ or of $\log(\mathbf{F})$ where $\mathbf{F}$ is obtained by elementwise division of $\mathbf{A}$ by $\mathbf{B}$ is more appropriate. Denoting by $d_k$, $u_k$ and $v_k$ the singular elements of $\log(\mathbf{F})$, the (saturated) reconstitution formula for $\mathbf{A}$ relatively to $\mathbf{B}$ gives:

$$[\log(\mathbf{A})]_{ij}^{IJ} = [\log(\mathbf{B})]_{ij}^{IJ} + \sum_{k=1}^{M} d_k u_{i,k} v_{j,k}.$$

## 3.3 SVD analysis of A and B

$\mathbf{A}$ and $\mathbf{B}$ have now an equal status, common part $(\mathbf{A} + \mathbf{B})/2$, and specific parts $(\mathbf{A} - \mathbf{B})/2 = -(\mathbf{B} - \mathbf{A})/2$. Each part may then be analysed by SVD. Interestingly, the two different SVDs can be performed in one stroke by considering the partitioned matrix:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}$$

With standard notation for the separate SVDs, the respective (saturated) reconstitution formulas give: $[(\mathbf{A} + \mathbf{B})/2]_{ij}^{IJ} = \sum_{k=1}^{M} d_k^1 u_{i,k}^1 v_{j,k}^1$ and $[(\mathbf{A} - \mathbf{B})/2]_{ij}^{IJ} = -[(\mathbf{B} - \mathbf{A})/2]_{ij}^{IJ} = \sum_{k=1}^{M} d_k^2 u_{i,k}^2 v_{j,k}^2$.

## 3.4 SVD of a square matrix

Setting $\mathbf{B} = \mathbf{A}'$ in the partitioned matrix above, a unique SVD gives the singular elements of the symmetric part $(\mathbf{A}+\mathbf{A}')/2$ and the skew symmetric part $(\mathbf{A} - \mathbf{A}')/2$. As expected, the singular elements for both parts have uncommon structures which preserve the symmetry and the skew symmetry (Greenacre, 2000):
$[(\mathbf{A} + \mathbf{A}')/2]_{ij}^{IJ} = \sum_{k=1}^{M} d_k^1 u_{i,k}^1 \ (-1)^{\epsilon_k} \ u_{j,k}^1$ (where $(-1)^{\epsilon_k}$ is either 1 or $-1$) and $[(\mathbf{A} - \mathbf{A}')/2]_{ij}^{IJ} = \sum_{k=1}^{[M/2]} d_k^2 (u_{i,2k-1}^2 u_{j,2k}^2 - u_{i,2k}^2 u_{j,2k-1}^2)$.

# 4 Bilinear modelling: the situation of equal status

The situation of matched tables where one table is a standard, can be taken care of by introducing an offset in the predictor. This is a straightforward procedure. Therefore, we restrict our attention to the situation where the two tables have an equal status.

## 4.1 Modelling A and B

We consider the baseline models for three-way tables and interleave them with relevant bilinear models taken from the SVD decompositions obtained in subsection 3.3. The spectrum now looks as follows:

$$\begin{aligned}
\eta_{hij}^{HIJ} &= \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} \\
\eta_{hij}^{HIJ} &= \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \sum_{k=1}^{m_1} \phi_k^1 \xi_{i,k}^{1I} \xi_{j,k}^{1J} \\
\eta_{hij}^{HIJ} &= \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \beta_{ij}^{IJ} \\
\eta_{hij}^{HIJ} &= \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \beta_{ij}^{IJ} + (-1)^h \sum_{k=1}^{m_2} \phi_k^2 \xi_{i,k}^{2I} \xi_{j,k}^{2J} \\
\eta_{hij}^{HIJ} &= \beta^{\emptyset} + \beta_h^H + \beta_i^I + \beta_j^J + \beta_{hi}^{HI} + \beta_{hj}^{HJ} + \beta_{ij}^{IJ} + \beta_{hij}^{HIJ}
\end{aligned}$$

where $m_1$ ($m_1 = 1, \ldots, M-1$) and $m_2$ ($m_2 = 1, \ldots, M-1$) are selected orders for the two different reduced rank interactions. Model selection, which includes the choice of order for the bilinear terms, can be based on some form of penalised likelihood (AIC, BIC ... )
Note that the bilinear terms in these models can be fitted by alternating generalised regressions (see Falguerolles & Francis, 1992, 1995). As expected, the choice of a Gaussian distribution with identity link returns the SVD estimates.

## 4.2 Modelling a square table

In this context, the models above become simplified since the third dimension ($H$) is only technical:

$$\begin{aligned}
\eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J \\
\eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J + \sum_{k=1}^{m_1} \phi_k^1 \xi_{i,k}^1 (-1)^{\epsilon_k} \xi_{j,k}^1 \\
\eta_{hij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ} \quad (\text{with } \beta_{ij}^{IJ} = \beta_{ji}^{IJ}) \\
\eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ} + \sum_{k=1}^{m_2} \phi_k^2 (\xi_{i,2k-1}^2 \xi_{j,2k}^2 - \xi_{i,2k}^2 \xi_{j,2k-1}^2) \\
\eta_{ij}^{IJ} &= \beta^{\emptyset} + \beta_i^I + \beta_j^J + \beta_{ij}^{IJ} \quad (\text{without } \beta_{ij}^{IJ} = \beta_{ji}^{IJ})
\end{aligned}$$

These include the bilinear extensions considered in van der Heijden & Mooijaart (1995). Again, the three dimensional trick has worked!

## References

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975) *Discrete multivariate analysis: theory and practice.* Cambridge, MA: MIT Press.

Falguerolles, A. de, & Francis, B. (1992). Algorithmic Approaches for Fitting Bilinear Models, (ed.) Y. Dodge and J. Whittaker, *COMPSTAT 92, Computational Statistics*, Vol. 1, 77-82, Physica-Verlag.

Falguerolles, A. de, & Francis, B. (1995). Fitting Bilinear Models in GLIM. *GLIM Newsletter*, **25**, 9-20.

Gabriel, R. K. (1971). The Biplot Graphical Display of Matrices with Applications to Principal Component Ananlysis, *Biometrika*, **58**, 453-467.

Greenacre, M. (2000). Correspondence Analysis of Square Asymmetric Matrices. *Applied Statistics*, accepted for publication.

James, I. R., & Segal, M. R. (1982). On a Method of Mortality Incorporating Age-Year Interaction, with Application to Prostate Cancer Mortality, *Biometrics*, **38**, 433-443.

McCullagh, P., & Nelder J. A. (1987). *Generalized Linear Models*, 2nd ed., London: Chapman and Hall.

Stigler, S. M. (1994). Citation Patterns in the Journals of Statistics and Probability. *Statistical Science*, **9**, 94-108.

van der Heijden, P.G.M & de Leeuw, J. (1985). Correspondence Analysis Used Complementary to Loglinear Analysis. *Psychometrika*, **50**, 429-447.

van der Heijden, P.G.M., & Mooijaart, A. (1995). Some New-Bilinear Models for the Analysis of Asymmetry in a Square Contingency table. *Sociological Methods and Research*, **24**, 7-29.

# Goodness of Fit Tests for Categorical Time Series

Konstantinos Fokianos [1]

[1] University of Cyprus, Department of Mathematics & Statistics, P.O. Box 20537, Nicosia 1678, Cyprus

**Abstract:** We study the power divergence family of goodness of fit tests in the context of regression models for categorical time series. We show that this family of test statistics provides a useful framework for checking the adequacy of the fit of those models.

**Keywords:** martingale, power divergence, partial likelihood.

## 1 Introduction

Regression models for categorical time series have been examined by many authors. Some references include Fahrmeir and Kaufmann (1987), Kaufmann (1987) and more recently Fokianos and Kedem (1998). A question that arises after fitting a regression model is that of the quality of the fit. In the context of generalized linear models, goodness of fit is examined either by a Pearson chi–square test or by the residual deviance (see, for example, Fahrmeir and Tutz (1994) ). However, in the special case of regression models for categorical time series this approximation can be poor due to the fact that data are sparse. Thus, some other techniques should be developed in order to take advantage of the fitting output for resolution of issues related to the fit of the model. The topic of goodness of fit of a regression model for a categorical time series has been addressed by some authors either by conducting a chi–square test or by inspection of the residuals. Our method is based on the so called power divergence family of tests which we describe in detail below.

## 2 Power Divergence Family

The power divergence family of goodness of fit tests has been introduced by Cressie and Read (1984) as a generalization of the well known Pearson's $X^2$ and likelihood ratio $G^2$ test statistics. The family of these test statistics is useful on examining the goodness of fit of a model for independent data. Let us be more specific. Denote by $\alpha_\lambda$ the deviation–or power divergence–between observed and expected counts, that is $\alpha_\lambda$ is a distance which is

given by

$$\alpha_\lambda(\text{observed}, \text{expected}) = \frac{2}{\lambda(\lambda+1)} \text{observed} \left[ \left( \frac{\text{observed}}{\text{expected}} \right)^\lambda - 1 \right].$$

Then the power divergence family of test statistics indexed by a parameter $\lambda \in R$, say $I(\lambda)$, is just the sum over all cells of these deviations. Namely,

$$I(\lambda) = \sum_{\text{cells}} \alpha_\lambda(\text{observed}, \text{expected}).$$

Recently Osius and Rojek (1992) showed that for independent multinomial data and under increasing–cells assumptions, the power divergence family of tests is asymptotically normally distributed. We extend this result to regression models for categorical time series in the next section.

## 3 Main Results

Suppose we observe a nonstationary categorical time series, say $\{\mathbf{Y}_s, s = 1, \ldots, T\}$. Let $m$ denote the possible number of categories for each observation. We assume that the $s-th$ observation is given by the vector $\mathbf{y}_s = (y_{s1}, \ldots, y_{sq})'$ of length $q$, with elements

$$y_{sj} = \begin{cases} 1, & \text{if the j-th category is observed at time } s \\ 0, & \text{otherwise} \end{cases}$$

for $s = 1, \ldots, T$ and $q = m-1$. In addition, we denote by $\mathbf{p}_s = (p_{s1}, \ldots p_{sq})'$ the vector of conditional probabilities given $\mathcal{F}_{s-1}$, that is $p_{sj} = P(y_{sj} = 1 \mid \mathcal{F}_{s-1})$, $j = 1, \ldots, q$, $s = 1, \ldots, T$. Here $\mathcal{F}_{s-1}$ stands for the whole information up to and including time $s$. Clearly, $y_{sm} = 1 - \sum_{j=1}^q y_{sj}$ and $p_{sm} = 1 - \sum_{j=1}^q p_{sj}$. Finally, we let $\mathbf{Z}_{s-1}$ to denote a $p \times q$ matrix that represents a covariate process. The latter may include past values of the process or/and any other auxiliary processes. Let

$$\mathbf{p}_s(\boldsymbol{\beta}) = h(\mathbf{Z}'_{s-1}\boldsymbol{\beta}).$$

Here $\boldsymbol{\beta}$ denotes a $p$ dimensional vector of time invariant unknown parameters and the function $h$ is the so called link function. Inference regarding the vector of unknown parameters is based on partial likelihood (see Fokianos and Kedem (1998)). It can be shown that the maximum partial likelihood estimator–denoted by $\hat{\boldsymbol{\beta}}$–is consistent and asymptotically normally distributed.
Consider now the following quantity

$$u_s(\boldsymbol{\beta}) \equiv \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^m y_{sj} \left[ \left( \frac{y_{sj}}{p_{sj}(\boldsymbol{\beta})} \right)^\lambda - 1 \right].$$

Observe that the expression $\sum_{s=1}^{T} u_s(\boldsymbol{\beta})$ is the analog of the power divergence statistic for independent data. Indeed, we compare the observed versus the expected counts raised to the power $\lambda$ with 1. Thus, if we denote by

$$e_s(\boldsymbol{\beta}) \equiv E[u_s(\boldsymbol{\beta}) \mid \mathcal{F}_{s-1}] = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} p_{sj}(\boldsymbol{\beta}) \left[ \left( \frac{1}{p_{sj}(\boldsymbol{\beta})} \right)^{\lambda} - 1 \right],$$

the conditional expectation of $u_s(\boldsymbol{\beta})$ given the past process, it is sensible to expect that under the hypothesis that the model is true, the difference of $\sum_{s=1}^{T} u_s(\hat{\boldsymbol{\beta}}) - \sum_{s=1}^{T} e_s(\hat{\boldsymbol{\beta}})$ varies around 0. It turns out that

$$I_T(.) = \sum_{s=1}^{T} [u_s(.) - e_s(.)]$$

evaluated at the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is approximated by a normal random variable with mean 0 and some variance, say $\xi_T(\boldsymbol{\beta})$. Thus, under some regularity conditions, we can show that the ratio

$$\frac{I_T(\hat{\boldsymbol{\beta}})}{\sqrt{\xi_T(\hat{\boldsymbol{\beta}})}} \to \mathcal{N} \tag{1}$$

in distribution as $T \to \infty$ where $\mathcal{N}$ is a standard normal random variable. Hence, a two–sided goodness of fit test can be based on (1).

## 4 Summary

We discussed the power divergence family of goodness of fit tests for categorical time series. We extended the definition to regression models for categorical time series and showed that it is asymptotically normally distributed.

## references

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440–464.

Fahrmeir, L. and Kaufmann H. (1987). Regression models for nonstationary categorical time series. *Journal of Time Series Analysis*, **8**, 147–160.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models.* New York:Springer-Verlag.

Fokianos, K. and Kedem, B. (1998). Prediction and classification of nonstationary categorical times series. *Journal of Multivariate Analysis*, **67**, 277-296.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *Annals of Statistics*, **15**, 79-98.

Osius, G. and Rojek, D. (1992). Normal goodness of fit tests for multinomial models with large degrees of freedom. *Journal of the American Statistical Association*, **87**, 1145–1152.

# Modelling Subject-specific Economic Behaviour by Random Coefficient Models

Jutta Gampe[1]

[1] Fachhochschule Osnabrück, Fachbereich Wirtschaft, Postfach 1940, 49009 Osnabrück, Germany

**Abstract:** For modelling the inter-individual variation in reaction patterns in an economic experiment a random coefficient model is proposed. The two-dimensional mixing distribution of the random coefficients is left unspecified and estimated by nonparametric maximum likelihood.

**Keywords:** Random Coefficient Model; NPML; Subject-specific Modelling.

## 1 Introduction

In economic theory conclusions are often derived from assumptions on individual behaviour, but data are commonly only available on some aggregated level. Experimental data for checking assumptions usually cannot be collected. One exception is the work of Fehr, Kirchsteiger & Riedl (1993), where they consider the question whether there is some notion of fairness present in economic actions. They conducted several experiments where all the participants could control the amount of money they could take home after the session by choosing their actions according to their individual preferences. Contrasting the observed behaviour with forecasts from economic theory is a valuable empirical check of assumptions.

## 2 Experiment and Data

In several sessions the following experiment was conducted: Two groups of participants were randomly assigned: Buyers and sellers. The groups were situated in different rooms to guarantee anonymous trading. In the first step buyers announced price offers for the good traded which could be accepted by any of the sellers. As soon as a price offer was accepted, the seller, in a second step, fixed the effort level at which he was willing to produce this good. This effort level simultaneously determined the gain of the seller as well as the gain of the buyer (by increased/reduced quality of the good).
Detailled specifications were as follows: Prices $p$ could be chosen from the range 30 to 120 (as multiples of 5), effort levels could be fixed in steps of

0.1 between 0.1 and 1.0. Effort levels $e$ were translated into monetary cost for the seller $c(e)$ by

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Effort level $e$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| Monetary cost $c(e)$ | 0 | 1 | 2 | 4 | 6 | 8 | 10 | 12 | 15 | 18 |

Total monetary gain $g_s$ for the seller was given by

$$g_s(p, e) = p - 26 - c(e)\,.$$

The gain $g_b$ for the buyer was given by

$$g_b(p, e) = (126 - p) \cdot e\,.$$

All these specifications were known to all participants.
Following classical economic theory of strictly money-maximizing agents, it could be expected that any seller, who agreed on a price offer, would fix his effort level at $e = 0.1$ giving zero cost. Buyers would anticipate that behaviour and therefore would strictly offer minimum prices.
During all the sessions there were 276 contracts observed, the distibution of prices offered is given in Figure 1:

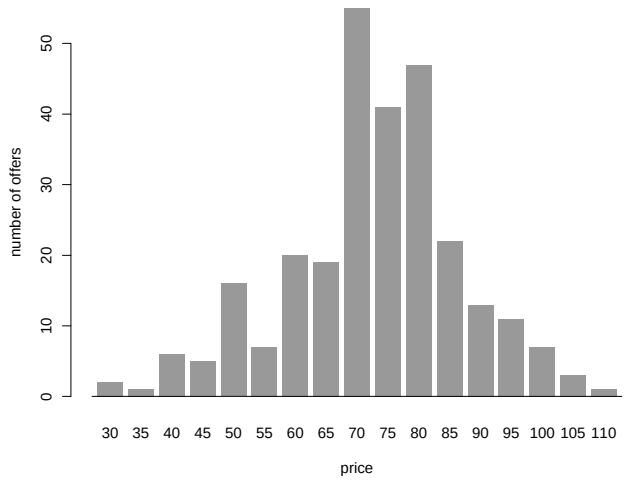



FIGURE 1. Offered prices during all sessions

As can be seen, price offers are significantly higher than the minimum value of 30, which is almost never observed. In this analysis we will focus on the reaction patterns of the sellers, that is, we want to describe how their effort choices depend on the offered prices. If no fairness ideas are present, we would expect that there is no increase of effort for higher prices. Figure 2

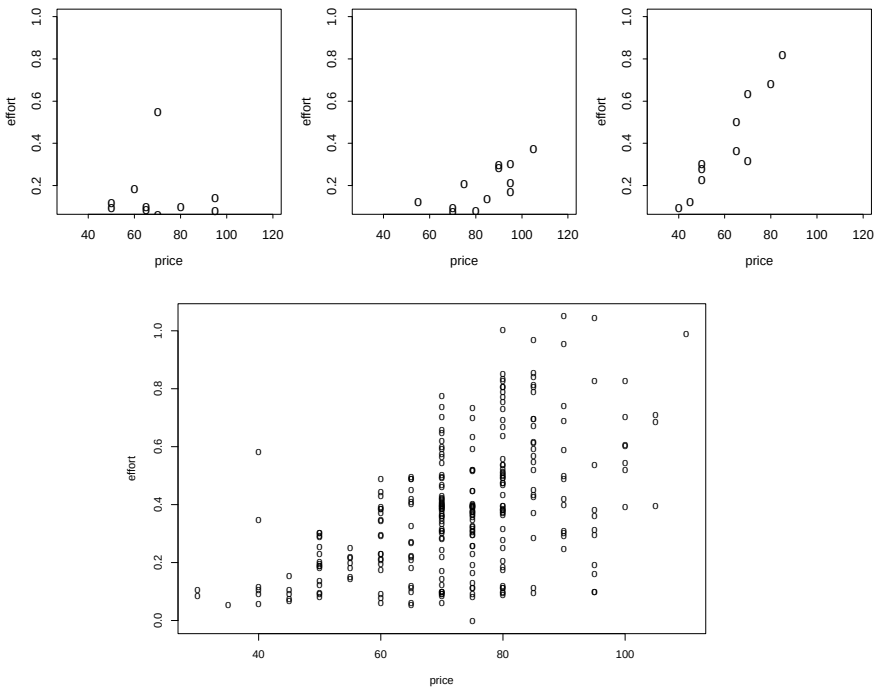



FIGURE 2. Price-effort relation: Three selected individuals (top), for all contracts (bottom)

shows three of the overall 34 individual reaction patterns as well the complete effort-price pattern. (Points are slightly jittered to show observations with identical coordinates.)

## 3 Random Coefficient Models

As can be seen, quite severe inter-individual variation is present, reflecting the obvious fact that economic behaviour is heterogeneous in the population. Ignoring subject-specific variability not only neglects an important feature of the data set but also renders conclusions for the mean behaviour incorrect.

Effort levels $e$ are in $[0.1, 1.0]$ and act as proportions on the buyers'gains. Furthermore, prices will have to exceed some limit until sellers choose efforts $> 0.1$. Therefore, effort will be transformed to the response as $y = logit(e)$ to linearize the relation.

We will focus on the following random coefficient model (Longford, 1993)

for $y_{ij}$ and prices $x_{ij} = p_{ij}$:

$$y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})\, x_{ij} + \varepsilon_{ij}\,, \tag{1}$$

where $i = 1, \ldots, s$ is the subject index of the seller, and $j = 1, \ldots, n_i$ is the index of the contract accepted by seller $i$. Intercepts and slopes vary independently and randomly across individuals, and are independent of the errors $\varepsilon_{ij} \sim N(0, \sigma^2)$. Conditionally on the random parameters $b_{0i}$ and $b_{1i}$ observations on individuals are independent.

## 4 NPML Estimation

In contrast to other applications where the mixing distribution of the random effects is often a nuisance feature, in this context the distribution of the random coefficients is also of interest — especially of the random slopes $b_{1i}$. A normality assumption for the random coefficients does not seem plausible as there are several money-maximizing persons present that should give rise to some mass at the lower end of the distribution. To avoid any parametric specification, the mixing distribution $\pi(b_0, b_1)$ is left completely unspecified and estimated by nonparametric maximum likelihood (NPML; Aitkin, 1999). The estimate is a discrete distribution on a finite number of masspoints $(z_K, u_k)$ and masses $\pi_k$, $k = 1, \ldots, K$.
Computation of the estimates is achieved by an appropriate EM-setup:
The log likelihood for model (1) is

$$l(\theta) = \sum_{i=1}^{s} \log \sum_{k=1}^{K} \pi_k\, f_{ik}\,, \tag{2}$$

where $f_{ik} = \prod_j f(y_{ij} \,|\, \theta,\ (z_k, u_k))$ , $f(.)$ the normal density and $\theta = (\beta, \sigma^2)$ the fixed model parameters. Finding the likelihood equations for $\beta$ gives

$$\frac{\partial\, l}{\partial\, \beta} = \sum_{i=1}^{s} \sum_{k=1}^{K} w_{ik}\, \frac{\partial log f_{ik}}{\partial \beta} \tag{3}$$

which is a weighted sum of ordinary likelihood equations. The weights

$$w_{ik} = \frac{\pi_k\, f_{ik}}{\sum_l \pi_l\, f_{il}}$$

can be interpreted as posterior probability observation $y_{ij}$ comes from the mixture component $k$. Calculating the weights in the E-step and then solving (3) in the M-step gives one cycle of the algorithm. The $\pi_k$ are obtained by averaging the weights $w_{ik}$ over the individuals $i$. Following the idea of Hinde & Wood (1987), the location of the masspoints can be estimated simultaneously with the other parameters by extending the regression model by a K-level factor with parameters $z_k$ and a factor-variable-interaction with parameters $u_k$. For $K$ components this means expanding the original data $(y_{ij}, x_{ij})$ to length $K \cdot \sum n_i$.

### 4.1 Computational Issues

This approach can easily be implemented in S-Plus by using the appropriate weights from the E-step in the fitting procedure. Convergence of the EM-algorithm is quite fast and stable, though convergence to local maxima was observed for these data. Variation of the starting values is one simple way of dealing with this problem. The number of masspoints $K$ can be obtained by starting with $K = 1$ and then increasing the number sequentially until the NPML estimate is attained. Fitting a model with a large number of masspoints and observing which locations coincide after convergence of the iterations can be used to verify the results. In case of convergence to local maxima this approach is, though necessary, quite cumbersome.

## References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics 55*, 117–128.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does fairness prevent from market clearing? An experimental investigation. *Quarterly Journal of Economics 108*, 437–460.

Hinde, J. P. and Wood, A. T. (1987). Binomial variance component models with a nonparametric assumption concerning random effects. In *Longitudinal Data Analysis*, R. Crouchley (ed.).

Longford, N. (1993). *Random Coefficent Models.* Oxford University Press.

# The Second Order Framework and the Modeling of Rare Events

M. Ivette Gomes[1]

[1] D.E.I.O. and C.E.A.U.L., Faculty of Science of Lisbon, Bloco C2, Campo Grande, 1749-016 Lisboa, Portugal, e-mail: ivette.gomes@fc.ul.pt

**Abstract:** The main objective of *Statistical Theory of Extremes* is the prediction of rare events, and its primary problem has been the estimation of the tail index $\gamma$, usually performed on the basis of the largest $k$ order statistics in the sample or the excesses over a high level $u$. The second order theory for extremes enhanced the importance of a second order parameter $\rho$, relevant in itself and for the semi-parametric estimation of other parameters of rare events. In this paper we shall describe the shape of the tails for two sets of financial data. The main objective is to perform the estimation of a heavy tail index through a previous estimation of the second order parameter, by means of an adaptive bootstrap estimator, an asymptotically unbiased generalization of the Hill estimator, and a Generalized Jackknife estimator.

**Keywords:** Statistical Theory of Extremes, Semi-Parametric Estimation, Resampling Methodologies.

## 1 Introduction and preliminaries.

In *Statistical Extreme Value Theory* our main interest is the prediction of *rare events*, and the primary functional of rare events is the *tail index* $\gamma$, which appears as the shape parameter of the limiting distribution function (d.f.) of the sequence of maximum values, $\{X_{n:n} = max(X_1, \ldots, X_n)\}_{n\geq 1}$, linearly normalized, associated to a random sample $(X_1, X_2, \ldots, X_n)$ from an underlying model $F(.)$. Indeed, that limiting d.f. is of the type of an *Extreme Value* (EV) d.f.,

$$G_\gamma(x) := \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right),\ 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0 \\ \exp(-\exp(-x)),\ x \in \Re & \text{if} \quad \gamma = 0 \end{cases} . \qquad (1)$$

Whenever there is such a non-degenerate limit we say that $F$ is in the *domain of attraction* of $G_\gamma$, and write $F \in D(G_\gamma)$. For heavy tails ($\gamma > 0$), usual in applications, $F \in D(G_\gamma)$ if and only if $U = F^{\leftarrow}(1-1/t) \in RV_\gamma$, i.e. $U(.)$ is a *regularly varying* function at infinity with *index* $\gamma$. This is the first order behaviour in *Extreme Value Theory*. The second order behaviour enhances the importance of a *second order parameter* $\rho$. Assuming that

there exists a function $A(t)$ of constant sign and going to 0 as $t \to \infty$, such that $U(tx)/U(t) - x^\gamma$ is of the same order of $A(t)$, we have

$$\lim_{t\to\infty} \frac{U(tx)/U(t) - x^\gamma}{A(t)} = x^\gamma \frac{x^\rho - 1}{\rho}, \tag{2}$$

for every $x > 0$. The limit in (2) must be of the stated form, and $|A(t)| \in RV_\rho$ (Geluk and de Haan, 1987). For heavy tails the most prominent estimator of $\gamma$ is the Hill estimator

$$\gamma_n^{(1)}(k) := \frac{1}{k}\sum_{i=1}^{k}[lnX_{n-i+1:n} - lnX_{n-k:n}] \tag{3}$$

(Hill, 1975), based on the $k+1$ largest observations, where $k = k_n$ needs to be an intermediate sequence, i.e., $k \to \infty$ and $k/n \to 0$. Then, and provided that (2) holds, we have the distributional representation (de Haan and Peng, 1998),

$$\gamma_n^{(1)}(k) \stackrel{\mathrm{d}}{=} \gamma + \frac{\gamma}{\sqrt{k}}P_n + \frac{1}{1-\rho}A(n/k) + o_p(A(n/k)), \tag{4}$$

where $P_n$ is asymptotically a standard Normal r.v. For general tails the most common estimator is the Moment estimator (Dekkers et al, 1989),

$$\gamma_n^{(M)}(k) := \gamma_n^{(1)}(k) + 1 - \frac{1}{2}\left\{1 - \frac{\left(\gamma_n^{(1)}(k)\right)^2}{\gamma_n^{(2)}(k)}\right\}^{-1}, \tag{5}$$

with

$$\gamma_n^{(2)}(k) := \frac{\sum_{i=1}^{k}[lnX_{n-i+1:n} - lnX_{n-k:n}]^2}{2\sum_{i=1}^{k}[lnX_{n-i+1:n} - lnX_{n-k:n}]}.$$

Under the same conditions as before, we have the validity of the distributional representation

$$\gamma_n^{(M)}(k) \stackrel{\mathrm{d}}{=} \gamma + \frac{\sigma_M}{\sqrt{k}}Z_n + b_M A(n/k) + o_p(A(n/k)), \tag{6}$$

where $Z_n$ is an asymptotically standard Normal r.v., and where

$$\begin{aligned} \sigma_M &= \sqrt{\gamma^2+1}, \\ b_M &= \tfrac{2-\rho+(\gamma-2)(1-\rho)}{\gamma(1-\rho)^2}. \end{aligned} \tag{7}$$

Both estimators, in (3) and in (5), may thus have a non-null asymptotic bias, dependent on $\rho$, and the "sample paths" of such estimators for different $k$ values are usually "disturbing" from a practical point of a view, claiming for alternative estimators of the tail index $\gamma$.

## 2 Estimation of the second order parameter

We shall consider here a first estimator of the second order parameter $\rho$ of the type suggested by Hall and Welsh(1985), and given by

$$\widehat{\rho}_n^{(1)} := -\left|\log\left|\frac{1/\gamma_n^{(1)}(t_1) - 1/\gamma_n^{(1)}(s)}{1/\gamma_n^{(1)}(t_2) - 1/\gamma_n^{(1)}(s)}\right| / \log\frac{t_1}{t_2}\right|, \tag{8}$$

for $s = [n^\sigma]$, $t_1 = [n^{\tau_1}]$, $t_2 = [n^{\tau_2}]$, with $\sigma = 0.5$, $\tau_1 = 0.9$, $\tau_2 = 0.95$. The second estimator considered is the one advanced by Drees and Kaufman (1998),

$$\widehat{\rho}_n^{(2)} := \log\frac{\max\limits_{1\le i\le[\lambda\overline{k}_n(r_n^\xi)]} \sqrt{i}\left|\gamma_n^{(1)}(i) - \gamma_n^{(1)}([\lambda\overline{k}_n(r_n^\xi)])\right|}{\max\limits_{1\le i\le[\overline{k}_n(r_n^\xi)]} \sqrt{i}\left|\gamma_n^{(1)}(i) - \gamma_n^{(1)}([\overline{k}_n(r_n^\xi)])\right|} / \log\lambda - \frac{1}{2}, \tag{9}$$

where $\xi = 0.7$, $\lambda = 0.6$, and

$$\overline{k}_n(r_n) = \min_{1\le k\le n-1} \max_{1\le i\le k} \sqrt{i}\left|\gamma_n^{(1)}(i) - \gamma_n^{(1)}(k)\right| > r_n.$$

All these estimators of $\rho$ are related to estimators of the tail index $\gamma$ and here, since the data analyzed come from populations with heavy tails, we have linked them with Hill's estimator, defined in (3). We shall also consider a third estimator based on the bootstrap technique introduced by Danielson et al (1998):

$$\widehat{\rho}_n^{(3)} := \frac{\ln\overline{k}_0^*(n_1)}{2\ln(\overline{k}_0^*(n_1)/n_1)}, \tag{10}$$

where $\overline{k}_0^*$ denotes the sample counterpart of $k_0^* = \arg\min_k MSE^*[T_n(k)]$, with $T_n(k)$ a suitable statistic with null mean value, related to our estimator, and with a distributional representation of the type of the one in (4) or in (6), being $n_1$ a suitable sub-sample size for the bootstrap samples, here taken equal to $n^{.95}$. A possible statistic will be presented in the following section; such a statistic is merely the difference of the semi-parametric estimator of the tail index, under consideration, computed at two different levels. All these estimators of the second order parameter $\rho$ are based on the regular variation properties of the $A(.)$ function. The multi-sample simulation carried out for different sample sizes $n$, and for different models, covering a reasonable region of $\rho$-values, $\rho \in (-\infty, 0]$, did not enable us to advance with an optimal choice among the class of estimators considered, and even claims for more suitable estimators, a topic which goes beyond the scope of this paper.

## 3 A bootstrap adaptive estimator

The estimation of $\rho$ has a high influence in the recently developed adaptive estimators of the tail index (Drees and Kaufman (1998), and references therein). We consider here adaptive Hill and Moment estimators $\gamma_n^{(j)}(\widehat{k}_0^{(j)}(n))$, where $\widehat{k}_0^{(j)}(n)$ is a consistent estimator of $k_0^{(j)}(n) := \arg\min_k MSE[\gamma_n^{(j)}(k)]$, $j = 1$ or $j = M$, obtained by means of Monte Carlo simulation, and bootstrap techniques. The type of bootstrap procedure we are going to use here was first applied by Danielson et al (1998), and relies on a suitable *auxiliary statistic*, with a null mean value, and with a *MSE* structure similar to the *MSE* structure of the original estimator, either (3) or (5). The auxiliary statistic we have used is

$$T_n^{(j)}(k) := \gamma_n^{(j)}(k/2) - \gamma_n^{(j)}(k), \ j = 1 \text{ and } j = M. \tag{11}$$

Let us denote by $T_n^{(j)*}(k)$ the bootstrap version of $T_n^{(j)}(k)$. Then, for a sample of size $n_1 < n$, here chosen equal to $n^{.95}$, obtain by Monte Carlo simulation, $k_0^{(j)*}(n_1) := \arg\min_k MSE[T_n^{(j)*}(k)]$, and $k_0^{(j)*}(n_1^2/n)$. Consider then, for $j = 1, M$,

$$\widehat{k}_0^{(j)}(n) := \left[ C_\rho \frac{[k_0^{(j)*}(n_1)]^2}{k_0^{(j)*}(n_1^2/n)} \right], \ C_\rho = (1-2^\rho)^{\frac{2}{1-2\rho}}, \ n_1 = n^{.95}, \tag{12}$$

where $\rho$ is going to be estimated by any of the estimators in section 2. For more details on this approach see Gomes (1999), and references therein.

## 4 A Generalized Hill estimator

Gomes et al (1999) introduced the Generalized Hill estimator

$$\gamma_n^{(\alpha)}(k) := \frac{k^{\alpha-2} \sum_{i=1}^k [\ln X_{n-i+1:n} - X_{n-k:n}]}{\Gamma(\alpha+1) \sum_{i=1}^k [\ln X_{n-i+1:n} - X_{n-k:n}]^{\alpha-1}}, \tag{13}$$

where $\Gamma(.)$ is the complete gamma function. The Hill estimator in (3) corresponds to the choice $\alpha = 1$ in (13), and the estimator $\gamma_n^{(\alpha)}(k)$ is asymptotically unbiased if we choose $\alpha$ such that

$$(1-\rho)^{\alpha-1}[1+\rho(\alpha-2)] = 1. \tag{14}$$

The estimators of $\rho$ described before may thus be used for the selection of the value of $\alpha$ to be considered in (13). The choice of the optimal sample fraction is then practically irrelevant due to the asymptotically unbiasedness of this tail index estimator. In the simulations and in the applications to real data we have used $k = n^{.95}$, which satisfies the requisites of consistency, providing an estimator with high efficiency relatively to the Hill estimator at the optimal level.

## 5 The Generalized Jackknife Hill estimator

The *Generalized Jackknife* statistic of Gray and Schucany (1972) is based on two different estimators of the same functional, with similar bias properties. As a particular case of the Jackknife theory, if we have two different biased consistent estimators $\gamma_n^{(1)}$ and $\gamma_n^{(2)}$ of the functional $\gamma(F(.))$, and if $E\left[\gamma_n^{(1)}\right] = \gamma + \varphi(\gamma)d_1(n)$; $E\left[\gamma_n^{(2)}\right] = \gamma + \varphi(\gamma)d_2(n)$, then, denoting by $q_n := \frac{BIAS[\gamma_n^{(1)}]}{BIAS[\gamma_n^{(2)}]} = \frac{d_1(n)}{d_2(n)}$, the *Generalized Jackknife* statistic associated to $(\gamma_n^{(1)}, \gamma_n^{(2)})$ is $\gamma_n^G\left(\gamma_n^{(1)}, \gamma_n^{(2)}\right) = \frac{\gamma_n^{(1)} - q_n\gamma_n^{(2)}}{1-q_n}$, which is an unbiased consistent estimator of $\gamma(F(.))$, provided $q_n \neq 1$, for every $n$.
In *Statistical Theory of Extremes*, whenever we are dealing with semi-parametric estimators of the tail index, or even other parameters of rare events, we have usually information about the asymptotic bias of those estimators, like the one given in (4) for the Hill estimator, or in (6) for the Moment estimator. We may thus choose estimators with similar asymptotic properties, and construct the associated *Generalized Jackknife* estimator. Here we shall consider the Generalized Jackknife estimator

$$\gamma_{n,\widehat{\rho}}^{G_j}(k) := \frac{\gamma_n^{(j)}(k) - 2^{-\widehat{\rho}}\gamma_n^{(j)}(k/2)}{1 - 2^{-\widehat{\rho}}}, \ j = 1 \text{ and } j = M, \tag{15}$$

where $2^{-\widehat{\rho}}$ is an estimator of $\frac{BIAS[\gamma_n^{(j)}(k)]}{BIAS[\gamma_n^{(j)}(k/2)]}$, being $\widehat{\rho}$ any of the estimators in section 2.

## 6 An application in the field of finance

We shall here describe the shape of the tails for monthly exchange rates of the US Dollar and of the Dutch Guilder published by *Banco de Portugal* (1984-1999), using the estimation of the second order parameter, followed by the estimation of the tail index, through the adaptive bootstrap Hill estimator, the asymptotically unbiased Generalized Moment estimator and the Generalized Jackknife Moment estimator based on affine combinations of Moments's estimators at two different levels. Assuming no *a priori* knowledge of the tails we began working with the Moment estimator in (5), which is highly negatively biased, but which exhibits the heavy tails of both sets of data. The sample paths of the Bootstrap Unbiased Moment estimator based on the statistic $T_n^{(M)}(k)$ and of the Generalized Jackknife Moment estimator $\gamma_{n,\widehat{\rho}}^{G_M}(k)$ in (15) were highly stable, providing an estimate of the tail index equal to 0.49 for the US Dollar data and equal to 0.54 for the Dutch Guilder data. The estimates of $\rho$, associated to the three methods explicited in 2, together with the associated estimates of $\alpha$ obtained through (14), were the following:

| | US Dollar | | Dutch Guilder | |
|---|---|---|---|---|
| | $\rho_n^{(j)}$ | $\alpha_n^{(j)}$ | $\rho_n^{(j)}$ | $\alpha_n^{(j)}$ |
| $j=1$ | -2.7483 | 2.2984 | -1.0789 | 2.6498 |
| $j=2$ | -2.3461 | 2.342 | -2.4428 | 2.3303 |
| $j=3$ | -0.8344 | 2.7952 | -0.8281 | 2.7999 |

The great discrepancy among the different values of $\rho$-estimates lead us to work the other way round: it seems better to choose first $\alpha$, by the drawing of sample paths of $\widehat{\gamma}_n^{(\alpha)}(k), 1 \leq k < n$, for different values of $\alpha$, together with an objective stability criterion. Here we have minimized the squared discrepancies to the median value, and have obtained $\alpha = 2.68$ for the US Dollar data, and $\alpha = 2.86$ for the Dutch Guilder data. If we use these values in equation (14) we obtain $\rho = -1.02$ and $\rho = -0.75$ for the US Dollar and for the Dutch Guilder, respectively, which suggests that the bootstrap estimates of $\rho$ are the ones with smaller bias among the $\rho$-estimators herewith considered. The estimates of $\gamma$ provided by the Generalized Hill estimator in (15) were then 0.39 and 0.48 for the US Dollar and for the Dutch Guilder data, respectively. The bootstrap adaptive Hill estimators of the tail index, again associated to the three estimators of $\rho$ considered, were for the US Dollar data $\gamma_1^{USD} = .43$, $\gamma_2^{USD} = .40$ and $\gamma_3^{USD} = .43$, and for the Dutch Guilder data $\gamma_1^{USD} = .48$, $\gamma_2^{USD} = .52$ and $\gamma_3^{USD} = .51$, which agree with the results obtained before, up to the associated standard errors.

## References


Dekkers, A.L.M., J.H.J. Einmahl and L. de Haan (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* 17, 1833-1855.

Drees, H. and Kaufman, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stoch. Proc. and Appl.* **75**, 149-172.

Geluk, J. and de Haan, L. (1987). *Regular Variation, Extensions and Tauberian Theorems.* CWI Tract 40, C. Math. Comp. Sc., Amsterdam.

Gomes, M.I. (1999). The Jackknife and the bootstrap methodologies in the estimation of parameters of rare events. *Revista de Estatística*, 5-23.

Haan, L. de and Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* **52**, 60-70.

Hall, P. and Welsh, A.H. (1985). Adaptive estimates of parameters of regular variation. *Ann. Statist.* **13**, 331-341.

Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

# Statistical Models for Conjoint Analysis

Mick Green[1]

[1] Centre for Applied Statistics, Lancaster University, Lancaster LA1 4YF, UK, e-mail: m.green@lancaster.ac.uk

**Abstract:**
Although the technique of Conjoint Analysis developed firstly from Psychometrics, recent research effort is largely centred on Market Research. The recognition of the potential commercial benefits of the technique for product development have lead to a headlong rush to devise estimation methods at the expense of careful consideration of statistical issues. Thus the majority of software implementations of the technique use inappropriate statistical assumptions with resultant deficiences in the performance of the analyses. This paper considers a proper statistical model for the case of ordinal categorical data and develops an estimation procedure based on Latent Class models. The results are compared to those obtained using the commonly used assumptions of commercial software.

**Keywords:** ordinal response, latent class models, proportional odds model

## 1 Introduction

The aim of conjoint analysis is to estimate the relative effects of the various attributes of an item that influence a person's preference for that item as against another item with different attributes. In a commercial context these attributes could be colour, some special feature or price. For other application areas, such as health care policy, these may be more abstract attributes. In all cases the methodology is basically the same, a person is presented with a collection of real or imaginary items each with differing attributes and is asked to compare them and assess their relative worth. In the form considered here this assessment is represented in terms of rating on a simple integer scale. From these data the analysis estimates the contribution (part-worth) of each attribute to the final evaluation of an item. In the case of a continuous attribute, such as price, this contribution may be modelled as function of the attribute value. A particular feature of conjoint analysis is that the population is represented as a collection of subgroups each with their own views on the relative worth of the different attributes. The majority of commercial software to conduct conjoint analysis uses a post hoc clustering algorithm applied to parameter estimates derived from a Normal regression model. This paper proposes a more appropriate statistical model, derives an estimation procedure for this model and shows how this can make a great difference to the results of the segmentation process.

## 2 A General Model

Observation $y_{ij}$ is a rate in the range $\{1, \ldots, R\}$ where 1 corresponds to "worst" and R to "best". The model for $y$ follows the model of McCullagh (1980), the proportional odds model:

$$P(y_{ij} = r \mid \theta, \beta, \gamma) = F(\theta_r + \eta_{ij}) - F(\theta_{r-1} + \eta_{ij})$$

where

$$\eta_{ij} = x_j^T \beta + z_i^T \gamma \text{ and } F(.) \text{ is a cdf}$$

$x_j$ are the item attribute covariates and $z_i$ are the individual covariates. The choices of $F(.)$ considered where those for the Logistic, Normal and Extreme Value distributions, leading to links logit, probit and complementary log-log.
For the segmentation model this model is extended by considering a latent class or mixture model of $K$ components such that

$$P(y_{ij} = r) = \sum_{k=1}^{K} \pi_k P(y_{ij} = r \mid \theta, \beta^{(k)}, \gamma)$$

That is, each segment has different parameters, $\beta$ , in the regression model on the item explanatory variables $\underline{x}$.

## 3 Model Fitting

The algorithm for model fitting combines the methods for fitting the proportional odds model with the EM algorithm for fitting latent class models. The method used for the proportional odds model was a simplified version of that of Wolfe (1996) with some modifications to allow recycling to increase efficiency. The EM algorithm for fitting latent class models has been described many times, in particular in Dempster et al (1977). As is common in mixture modelling there were problems with local maxima. A method similar to simulated annealing was developed to avoid this problem which worked effectively.

## 4 Implementation

The above algorithm was implemented as a set of GLIM macros. Representing polytomous data as a vector of zero/one indicators allows the use of the Poisson Likelihood in model fitting. The proportional odds model can

be seen as a partially non-linear model that can be accommodated using the OWN model facilities of GLIM4. The EM algorithm is then equivalent to iterative fitting of a weighted GLM with weights recalculated at each iteration. Operationalising the method this way gives the user flexibility in model specification.

## 5 Application

To illustrate the methodology we analyse a data set from a Market Research study.
The attributes of the items were Brand, Price and the presence/absence of a special feature. The relationship of worth to Price (P) was assumed to be

$$\exp(W) = P^{\beta_1} e^{-\beta_2 P}$$

This relationship allows a dual role of price, the negative price deterrent effect and a positive effect due to perceived quality. It was thought desirable to consider 5 different price values in order to allow for accurate assessment of this price model. Given that the study compared 4 brands it was not feasible to consider a complete design of the 40 combinations of attribute values. Even a half replicate design of 20 items was thought too complex a task to ask a person to complete. The design chosen had two blocks of 12 items each, so that a person was only presented with 12 items to compare, with random assignment to block. Each person was asked to assess the worth of each item on a scale of 1 to 7.
The analysis was firstly based on an assumption of the Normal distribution for the rate values. The fitting procedure used a simultaneous segmentation methodology similar to that described in section 3, so was an improvement on many of the commercially available methods. The analysis was then repeated using the proportional odds model and the results compared.
The following table shows the deviances of the fitted proportional odds model, using the probit link, for various numbers of segments, and the corresponding $\pi$ values.

| segments | Deviance | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\pi_4$ |
|---|---|---|---|---|---|
| 1 | 4610 | 1 | | | |
| 2 | 4509 | 0.573 | 0.427 | | |
| 3 | 4452 | 0.608 | 0.206 | 0.186 | |
| 4 | 4366 | 0.276 | 0.242 | 0.302 | 0.181 |

The logit link gave very similar results while the complementary log-log link gave a reduction in deviance of 45 and this model was used for the following analyses.
For ease of interpretation of the fitted model, predicted worth was converted to an expected value of the rating. Figure 1 shows the price profiles for

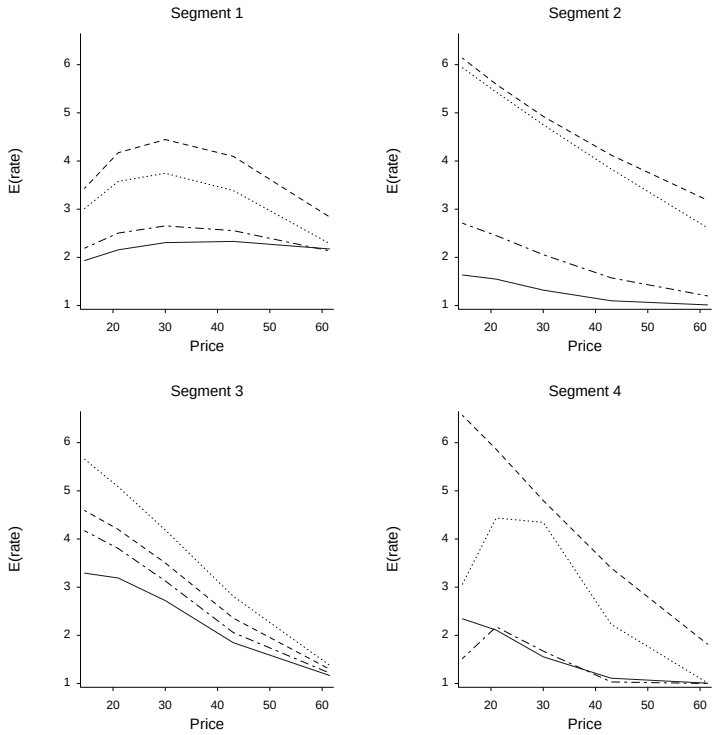



FIGURE 1.

each brand in the four fitted segments. These have sensible interpretations in terms of preference behaviour. For example segment 1 represents people who have a moderate brand preference but are not strongly influenced by price but manifest the positive price effect of perceived quality. The model for 5 segments showed a significant decrease in Deviance if we applied asymptotic results. However, it is well known that asymptotics do not apply in such cases so that it is not possible to devise an easy test for the best model. The model with 4 segments was chosen here on the basis that models with more segments generated price profiles that did not conform to what we would expect in terms of human behaviour.

## 6 Comparison of Analyses

The analysis above was compared to that based on an assumption of Normality for the distribution of the rates. One element that is easy to compare is the assignment of individuals to segments based on the maximum posterior probability. Comparing the 4 segment models, and equating segments

in the models to maximise the correspondence between them we still obtain less then 60% of individuals who are assigned to the same segments in the two models. By the same token, the price profiles are also quite different in the two models showing that we can obtain a very different conclusion when using an inappropriate probability model.

### 6.1 Validation

The ultimate test is in terms of effectiveness in prediction. The study included "hold-out cards", i.e. items that were excluded from the estimation procedure. These were be used in a validation of the model. Figure 2 shows a histogram of the predicted probability of the observed rate for the hold-out cards.

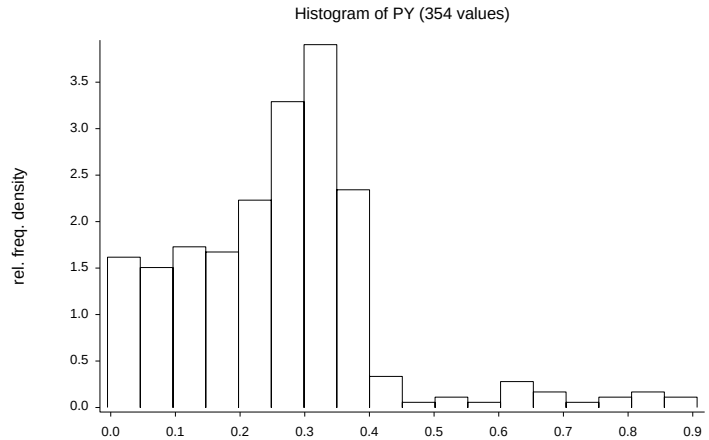



FIGURE 2.

While this looks poor, with a mean of 0.257, it is a considerable improvement over the non-segmented model with a mean probability of 0.205.
Alternatively, we can consider the predicted expected ratings. Figure 3 shows a graph of these against the observed ratings.
There is an indication that the model is not predicting the higher ratings as well as the lower ones, giving some concern to the market researcher.

### 6.2 Predicting Preference

A further task was included in the study, in which each person was asked to consider 4 particular items and choose the one they most preferred. Predicting preference as the item with highest worth does not give a useful validity test as all individuals assigned to segments 1 and 2 are predicted to prefer item 1 and all others to prefer item 4. A more effective method is

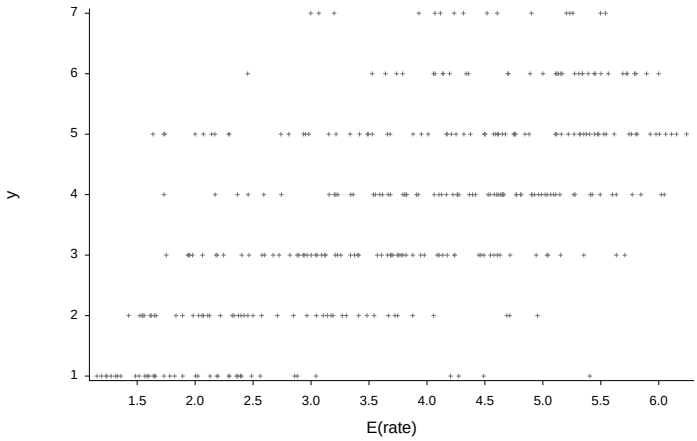



FIGURE 3.

to derive the probability of preference from the predicted worths. For the extreme value distribution this can be found from:

$$P(\text{prefer } j) = \frac{\exp(W_j)}{\exp(W_1) + \cdots + \exp(W_4)}$$

Using this for each segment gives the following expected frequencies.

expected frequency

| | | Y 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| segment | 1 | 11.9 | 6.5 | 7.9 | 2.8 |
| | 2 | 15.2 | 3.6 | 12.4 | 1.9 |
| | 3 | 3.9 | 18.7 | 1.3 | 22.2 |
| | 4 | 0.2 | 2.0 | 2.6 | 5.1 |

observed

| | | Y 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| segment | 1 | 19 | 5 | 4 | 1 |
| | 2 | 22 | 2 | 8 | 1 |
| | 3 | 10 | 17 | 2 | 17 |
| | 4 | 5 | 2 | 1 | 1 |

The observed frequencies are a very poor fit to these expected frequencies. However, visual comparison shows that the model is picking up some of the main features of individual preference. The broad features of segments 1,2 and 3 agree with observations, but there is some evidence that segment 4 is not a reliable predictor of behaviour.

**Acknowledgements:** Thanks to Ron Ventura who collected the data and assisted with some of the analysis.

## References

Dempster A.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *J.R.Statist.Soc B*, **51**, 127-138

McCullagh P. (1980) Regression models for ordinal data, *J.R.Statist.Soc B*, **42**, 109-142

Wolfe R. (1996) General Purpose Macros to Fit Models to an Ordinal Response, *GLIM Newsletter*, **26**, 20-27

# Causal Graphs and Unconfoundedness

Peter Kischka[1], Dietrich Eherler[1]

[1] Friedrich-Schiller-Universität Jena, Wirtschaftswissenschaftliche Fakultät, Lehrstuhl für Wirtschafts- und Sozialstatistik, Carl-Zeiß-Str. 3, D-07743 Jena, Tel.: 03641 - 94 33 00, Fax: 03641 - 94 33 02, e-mail: p.kischka@wiwi.uni-jena.de, d.eherler@wiwi.uni-jena.de

**Abstract:** This paper deals with several issues related to the ascertainment of unconfounded (causal) effects of a treatment variable $X$ on a response variable $Y$. Starting point of our consideration is a paper of J. Pearl (1998a) where the term stable unconfoundedness is introduced. We generalize this definition with respect to a set of variables $\mathbf{T}$. We characterize the definition and state, as Pearl, a necessary criterion for stable unconfoundedness in our more general situation that is operational for real world applications. Under more specific assumptions, we find a sufficient criterion for stable unconfoundedness.

**Keywords:** Causality, Unconfoundedness, Structral Equation Models, Graphical Bayesian Network Models

## 1 Introduction

When we examine the relationship between a treatment variable $X$ (e.g. interest / advertisement / medicine ) and a response variable $Y$ (e.g. investment / sales / well-being) our primary interest is to obtain an unbiased magnitude for the effect of the treatment on the response. Despite the knowledge of confounding this term has not yet been mathematically and applicably formalized. In a lately published working paper J. Pearl (1998a) introduces the term of stable unconfoundedness using so called manipulated distributions $\tilde{P}_x$ of $Y$, where $x$ varies on the range of $X$ (see section 2). We generalize his definition and results to a broader bunch of problems, when we know that common causes $\mathbf{T}$ exist. Nevertheless, the question remains, whether further not in $\mathbf{T}$ considered common causes exist, that still confound the adjusted effect

$$\tilde{P}_x(Y = y) = \sum_{\mathbf{t}} P(Y = y|\mathbf{T} = \mathbf{t}, X = x)P(\mathbf{T} = \mathbf{t}) \text{ for all } x, y$$

of $X$ on $Y$. This paper deals with this problem and tries to resolve it under weak assumptions.

## 2 Structural Equation Models

Let $\mathbf{V} = \{X_1, \ldots, X_n\}$ be defined by a recursive equation system as $X_i = f_i(\mathbf{PA}_i, U_i)$ $(1 \leq i \leq n)$ with measurable but elsewise unspecified functions $\mathbf{F} = \{f_1, \ldots, f_n\}$ where $\mathbf{PA}_i \subset \{X_1, \ldots, X_n\} \setminus \{X_i\}$ is the set of variables $X_i$ functionally depends on and with a random vector $\mathbf{U}= \{U_1, \ldots, U_n\}$ of mutually independent variables. We us the notation $M$=($\mathbf{U}$,$\mathbf{V}$,$\mathbf{F}$) for such a *structural equation model.* It is known, that the joint distribution of $\mathbf{V} = \{X_1, \ldots, X_n\}$ can be factorized to

$$P(x_1, \ldots, x_n) = \prod_{k=1}^{n} P(x_k|\mathbf{pa}_k)$$

in accordance to the variables $\mathbf{PA}_i$. The graphical depiction of the system $M$, constructed by drawing a directed arc from $X_j \in \mathbf{PA}_i$ to $X_i$, is called *causal graph* of $M$. This directed acyclic graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ is a minimal $I-map$, using Pearl's terminology and applying the d-separation criterion (Pearl, 1988). The $I-map$ property states that d-separation implies conditional independence. D-separation is a graphical criterion that takes the direction of edges into account.

**Definition 2.1**

Let $\mathbf{G}(\mathbf{V}, \mathbf{E})$ be a directed acyclic graph, $\mathbf{X}$,$\mathbf{Y}$,$\mathbf{Z}$ disjoint subsets of $\mathbf{V}$. Then $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ if on every path regardless of the direction from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$ there is a node $W$ with (i) $W$ has converging arrows and $W$ nor any descendent of $W$ is in $\mathbf{Z}$ or (ii) $W$ does not have converging arrows and $W$ is in $\mathbf{Z}$.

The assignment of distribution assumptions to $\mathbf{U}$ and the specification of the functions $f_i$ $(1 \leq i \leq n)$ is called *parameterization.* We say, that the joint distribution $P(\mathbf{V})$ induced by the parameterization and the causal graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ of $M$ are *faithful*, iff d-separation and (conditional) independence relations coincide (Spirtes et al., 1993). A *manipulation* $X_k \equiv x$ for some fixed $k$ $(1 \leq k \leq n)$ and $x \in \mathbb{R}$ is called the surgery in the equation system $M$ where we replace the $k$-th equation by the constant $X_k \equiv x$ and if $X_k \in \mathbf{PA}_i$ $(i \neq k)$ we set the corresponding realization in $\mathbf{pa}_i$ to the constant $x$ (Pearl, 1998b). The distribution of the manipulated system is given by

$$\tilde{P}_x(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n) = \frac{P(X_1 = x_1, \ldots, X_k = x, \ldots, X_n = x_n)}{P(X_k = x|\mathbf{PA}_k = \mathbf{pa}_k)}.$$

The marginal of $\tilde{P}_x(x_1, \ldots, x_{k-1}, x, x_{k+1}, \ldots, x_n)$ with respect to a variable $Y \in \mathbf{V} \setminus \{X_k\}$ is called the *effect if a manipulation* (Pearl, 1998b). To ascertain the effect of a manipulation we do not need to refrain to the whole distribution $\tilde{P}_x$. A simple formula is provided by the backdoor-criterion.

**Theorem 1** (Pearl, 1995)

Let $M$ be a recursive structural equation model and $\mathbf{G}$ be the corresponding causal graph. Further let be $\mathbf{T} \subset \{X_1,...,X_n\}$ and $X,Y \notin \mathbf{T}$. If

1. for all $X_j \in \mathbf{T}$: $X_j$ is not a descendent of $X$
2. $\mathbf{T}$ d-separates $X$ from $Y$ on every path (so called *backdoor-path*) between $X$ and $Y$, that has an arrow into $X$

then the effect of $X$ on $Y$ is given by:

$$\tilde{P}_x(Y = y) = \sum_{\mathbf{t}} P(\mathbf{T} = \mathbf{t}) P(Y = y | \mathbf{T} = \mathbf{t}, X = x) \ . \tag{1}$$

## 3 Unconfoundedness

Let a recursive equation system $M$ be given for the variables $X, Y, Z_1, \ldots, Z_m$ (Section 2). The variables $Z_1, \ldots, Z_m$ may be observable or not observable. The following definition of unconfoundedness is widely prevalent (e.g. Rothman, 1986; Schlesselman, 1982).

**Definition 3.1 Unconfoundedness** (Associational Criterion)
Let $\mathbf{T}$ be the set of variables of $M$ which are not affected by $X$. If

a) $X$ is independent of $Z_i$, $(Z_i \in \mathbf{T})$
*or*
b) $Y$ is independent of $Z_i$ given $X$ $\quad (Z_i \in \mathbf{T})$

then $X$ and $Y$ are said to be *unconfounded*. If a) and b) is not fulfilled for some $Z_i \in \mathbf{T}$, $X$ and $Y$ are said to be *confounded*.

Note firstly, that the definition of set $\mathbf{T}$ is not a property of the joint distribution of the variables $X, Y, Z_1, \ldots, Z_m$, but has to be ascertained from substantial knowledge. A variable $Z$ is not affected by a variable $X$, if in the causal graph $\mathbf{G}$ defined by $M$ there is no directed path from $X$ into $Z$. Secondly, if a) and b) are not fulfilled for some subset $\mathbf{W}$ of $\mathbf{T}$ then the effect of $X$ on $Y$ has to be computed by adjusting with $\mathbf{W}$. Instead of $P(Y = y|X = x)$ the formula $\sum_{\mathbf{w} \in \mathbf{W}} P(Y = y|X = x, \mathbf{W} = \mathbf{w}) P(\mathbf{W} = \mathbf{w})$ has to be used (e.g. Rosenbaum/Rubin, 1983) to obtain the effect of $X$ on $Y$. Pearl (1998a) gives an example that confoundedness in the sense of definition 3.1 can be present, however, an adjustment with the confounding variable leads to 'wrong' results. The following definition 3.2 generalizes definition 1 in Pearl (1998a).

**Definition 3.2 Causal, Stable Unconfoundedess**
$X$ and $Y$ are *causally unconfounded with respect to* $\mathbf{T} \subset \{Z_1, \ldots, Z_m\}$ in the equation system $M$, if the variables $Z_i \in \mathbf{T}$ are observable and the following equality holds:

$$\tilde{P}_x(Y = y) = \sum_{\mathbf{t}} P(Y = y | \mathbf{T} = \mathbf{t}, X = x) P(\mathbf{T} = \mathbf{t}).$$

$X$ and $Y$ are *stably unconfounded with respect to* $\mathbf{T} \subset \{Z_1, \ldots, Z_m\}$, if $X$ and $Y$ are causally unconfounded in every parameterization of the equation system (1).
A characterization for stable unconfoundedness in definition 3.2 states the following Theorem. The backdoor-criterion comes essentially to application.
**Theorem 2 Characterization of Stable Unconfoundedness**
Let $M = (\mathbf{U}, \mathbf{V}, \mathbf{F})$ be an equation system and $X, Y \in \mathbf{V}$, $\mathbf{T} \subset \mathbf{V} \setminus \{X, Y\}$. $X$ and $Y$ are stably unconfounded with respect to $\mathbf{T}$, if and only if $X$ und $Y$ are d-separated on every backdoor-path by the set $\mathbf{T}$ in the corresponding causal graph $\mathbf{G}$ to $M$.
This result is proved in Kischka/Eherler (1999). Unconfoundedness of $X$ and $Y$ is defined in definition 3.1 and in definition 3.2 with respect to the equation system $M$. The content of this notion therefore rests mainly upon the appropriateness and relevance of this system for the examined problem. Theorem 2 (Pearl, 1998a) states a criterion that rules out stable unconfoundedness even then when the equation system or respectively the corresponding graph are not completely specified. The following section generalizes this result.

# 4 Unconfoundedness in not Completely Specified Equation Systems

## 4.1 A Necessary Criterion

Starting point for the following considerations is a situation of a not completely specified recursive model $M$. We assume that the equation system $M$ does not necessarily include all variables nor all equations. If $X$ and $Y$ are unconfounded with respect to a set $\mathbf{T}$ in $M$ (Definition 3.2) they are not necessarily unconfounded in a $M$ embracing model $M'$.
In the following we assume that the unspecified model $M'$ corresponds to an acyclic causal graph. Let $V$ be a variable of $M'$ not contained in $M$ and let $M_V$, following Pearl's (1998a) notation, be a from $M$ resulting model with

- there is no directed path from $X$ to $V$ ($M_V 1$)
- if $V$ and $Y$ are dependent then there is a directed path from $V$ to $Y$ ($M_V 2$).

Pearl shows (Pearl, 1998a), that - in the case without adjustment - $X$ and $Y$ are not stably unconfounded in $M'$, if for a variable $V$ the conditions $M_V$ are fulfilled and the conditions a) and b) of definition 3.1 are not fulfilled. Theorem 3 generalizes this result by amending the set $\mathbf{T}$ of definition 3.2 into the conditions a) and b). Pearl's result follows with $\mathbf{T} = \emptyset$.
**Theorem 3 Necessity**
Assume that a model $M$ and an additional variable $V$ are given s. t. the conditions $M_V 1$ and $M_V 2$ hold. Assume that $X$ and $Y$ are stably unconfounded in $M$ with respect to $\mathbf{T}$. If

- $X$ is not independent of $V$ given $\mathbf{T}$
  and
- $Y$ is not independent of $V$ given $\{X\} \cup \mathbf{T}$

then $X$ and $Y$ are not stably unconfounded in $M'$ with respect to $\mathbf{T}$.
A proof of this theorem can be found in a discussion paper (Kischka/Eherler, 1999). Theorem 3 can be interpreted as follows. Given is a model $M$ to determine the effect of $X$ on $Y$. If there is a variable $V$ (outside of $M$) with the properties $M_V1$ and $M_V2$, we can test the independence of $V$ and $X$ given $\mathbf{T}$ as well as the independence of $V$ and $Y$ given $\{X\} \cup \mathbf{T}$. When both hypotheses are rejected we can infer that $X$ and $Y$ are not stably unconfounded in $M'$ with respect to $\mathbf{T}$. Conversely, Theorem 3 can be stated as a necessary criterion for stable unconfoundedness in the following way: If $X$ and $Y$ are stably unconfounded with respect to $\mathbf{T}$ and given that $M_V1$ and $M_V2$ hold then $V$ is independent of $X$ given $\mathbf{T}$ or $Y$ is independent of $V$ given $\{X\} \cup \mathbf{T}$.

### 4.2 A Sufficient Criterion

Theorem 3 states a necessary criterion for stable unconfoundedness. Of greater interest is, of course, to have a criterion for stable unconfoundedness in $M'$. Pearl shows (1998a, Example 2) that certain parameter constellations are possible to infer the independence of $V$ and $Y$ given $X$ from data, but that the equality of conditional and manipulated distribution might still not hold. To avoid such incidental independence relations, we have to assume that the to $M'$ corresponding graph $\mathbf{G}$ and the distribution $P$ induced by $M'$ are faithful.
**Theorem 4 Sufficiency**
Let an equation system $M$ be given. The variables $X$ and $Y$ are assumed to be stably unconfounded with respect to a set of variables $\mathbf{T}$ in $M$. Let $V$ be another variable in $M'$ with

a) $V$ is independent of $Y$ given $\{X\} \cup \mathbf{T}$
b) there is a directed path from $V$ to $X$
c) $V$ is *not* independent of $X$ given the set $\mathbf{T}$.

Then $X$ and $Y$ are stably unconfounded with respect to $\mathbf{T}$ in $M'$, if the graph $\mathbf{G}$ corresponding to $M'$ is faithful.
A proof of this result is given in Kischka/Eherler (1999). In the case we find a variable $V$ with the properties a),b),c) we can compute an unconfounded effect of $X$ on $Y$, given that the faithfulness assumption is not violated. Fortunately this happens only in very rare cases, i.e. almost surely it does not occur (Spirtes et al., 1993; Meek, 1995).

### 4.3 A Sufficient Criterion with $\mathbf{T} = \emptyset$

Let's assume we consider a pair of variables $X$ and $Y$ and the task is to determine the effect of $X$ on $Y$. We do not know a priori whether there exist confounding variables $\mathbf{T}$, thus we assume $\mathbf{T}=\emptyset$. In this case, we can express Theorem 4 as follows.

**Corollary**

Let an equation system $M$ be given. Let $V$ be another variable in $M'$ with

a) $V$ is independent of $Y$ given $X$
b) there is a directed path from $V$ to $X$

Then $X$ and $Y$ are stably unconfounded with respect to the empty set in $M'$, if the graph $\mathbf{G}$ corresponding to $M'$ is faithful.

Note that the existence of a directed path from $V$ to $X$ implies that $V$ and $X$ are not independent.

**Example**

Suppose we perform a survey on a population of graduated business administration students. We are interested in the effect of the mark in mathematics $X$ at the time of high-school graduation on the results in the final examination at university $Y$. We are not sure, whether there are other variables, like intelligence, that have influence on both, $X$ and $Y$. Another measurable variable $V$ satisfying the condition b) of the Corollary above is the type of high-school a student attended. Now we can perform a test whether the variable type of high-school $V$ is independent of the results in the final exam $Y$ given the mark in maths at high-school graduation $X$ and find out, assuming faithfulness, whether $X$ and $Y$ are stably unconfounded with respect to the empty set or whether there are common ancestors we have not yet considered.

## 5 Conclusion

This paper features results that can help researchers in all kind of practical applications to deal with the hard to handle problem of confounding. Based on graphical (Bayesian) models, which can be learned from data (e.g. Spirtes et al., 1993; Glymour/Cooper, 1999), testable criteria are given, that render ascertained causal effects either as confounded or as stably unconfounded. The task of the researcher is to find an additional observable variable that meets the required conditions. In the case of success, unbiased predictions of the effects of treatment on response variables can be computed.

### References

GLYMOUR, C., COOPER, G.F. (1999). Computation, Causation and Discovery, MIT Press.

KISCHKA, P., EHERLER, D. (1999). Causal Graphs and Unconfoundedness, Diskussionspapier, Reihe B, Nr. 99/05, Wirtschaftswissenschaftliche Fakultät, Friedrich-Schiller-Universität Jena, Jena, (to appear in: Allgemeines Statistisches Archiv).

MEEK, C. (1995). *Strong Completeness and Faithfulness in Bayesian Networks*, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence, 403-410, Morgan Kaufman, San Francisco.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufman, San Mateo.

PEARL, J. (1995). *Causal Diagramms for Empirical Research*, Biometrica, Vol. 82, No. 4, 669-709.

PEARL, J. (1998a). *Why there is no statistical test for confounding, why many think there is, an why they are almost right*, UCLA, Cognitive Systems Laboratory, R-256.

PEARL, J. (1998b). *Graphs, Causality and Structural Equation Models*, Sociological Methods and Research 27.

ROSENBAUM, P.; RUBIN, D. (1983). *The Central Role of Propensity Score in Observational Studies for Causal Effects*, Biometrica, 70,41-55.

ROTHMAN, K.J. (1986). *Modern Epidemiology*, Brown Little, First Edition.

SCHLESSELMAN, J.J. (1982). *Case-Controll-Studies: Design Conduct Analysis*, Oxford University Press, Oxford.

SPIRTES, P., GLYMOUR, C., SCHEINES, R. (1993). *Causation, Prediction and Search*, Lecture Notes in Statistics 81, New York.

# A Hierarchical Bayesian Model for Space-Time Variation of Disease Risk

Corrado Lagazio[1], Emanuela Dreassi[2], Annibale Biggeri[2]

[1] Dept. of Statistical Science, University of Udine, Via Treppo 18, 33100 Udine, Italy, e-mail: lagazio@dss.uniud.it

[2] Dept. of Statistics "G. Parenti", University of Florence, Viale Morgagni 59, 50134 Florence, Italy, e-mail: dreassi@ds.unifi.it

**Abstract:** In this paper we propose a hierarchical Bayesian model to study the variation in space and time of disease risk. We represent spatial effects following the usual Bayesian specification, while we adopt the birth cohort (instead of the commonly used death period) as the main time scale. The model includes also a space-time interaction term to take into account for structured inseparable space-time variability. The model is applied to male lung cancer mortality data in Tuscany, 1971-1994. While a period analysis points out a general increase of mortality, the cohort model shows that for the younger cohort there is a general and substantial decrease of the relative risk. Moreover, the cohort showing the maximum risk level varies over municipalities, showing a strong north-west/south-east gradient.

**Keywords:** Cohort Effects, Hierarchical Bayesian Model, Space-Time Analysis.

## 1 Introduction

In the past space-time variation of disease risk has been frequently studied simply by describing the difference between two risk maps, estimated separately for two periods (see the important work of Mason *et al.*, 1979). A hierarchical Bayesian model for the geographical analysis of disease risk including time dimension has been proposed by Bernardinelli *et al.* (1995). In that formulation time variation for each area is defined as spatially depending from neighbouring areas. For a critical evaluation of this choice see Knorr-Held (2000), in which are discussed the possible choices for a certain degree of dependence between space and time effects. Knorr-Held and Besag (1998) have improved the model considering two different time scales, age and calendar time. This solution is interesting because it considers, in a indirect way (age-period interaction), the birth cohort dimension. In epidemiology, time variation of disease risk is usually studied considering birth cohort as the main time scale (see the example on tuberculosis in Massachusetts 1870-1930, Lilienfeld and Lilienfeld, 1980). This choice is motivated by several biological reasons, that make birth cohort preferable

to calendar period. In this paper we propose a Bayesian space-time model to analyse the risk of disease in a defined region that explicitly considers the birth cohort time dimension.

## 2 Materials and methods

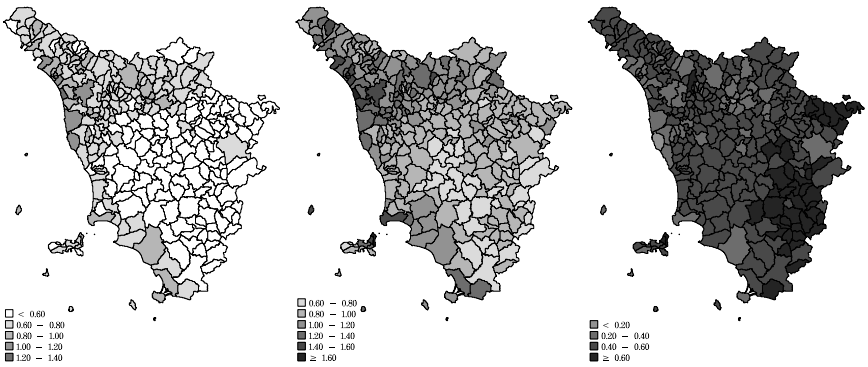
Male lung cancer death certificates (ISTAT) were considered for the 287 municipalities of Tuscany Region from 1971 to 1994. Deaths and corresponding populations for each municipality were tabulated for 18 age classes (0-4,...,85+) and five periods of five years (the first excluded) 1971-74,...,1990-94. The age-specific reference rates for the whole region were estimated using a simple age-cohort model (Clayton and Schifflers, 1987). The expected deaths for each age class and calendar time in each municipality were calculated by applying the age-specific rates to the area population. Observed and expected deaths were then determined for the seven birth cohorts 1900-05,...,1930-35, that are the most complete ones. We have considered a total of 33197 deaths on a mean annual population in the region of about 3,5 million people.
We defined a Poisson distribution with mean value $\lambda_{it}E_{it}$ ($E_{it}$ expected cases, $\lambda_{it}$ relative risk in the $i$-th area for the $t$-th cohort) for the number of observed deaths $O_{it}$ ($i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$) and we specified a log-linear model for the relative risk:

$$\log(\lambda_{it}) = v_i + u_i + \theta_t + \psi_{it} \tag{1}$$

where $v_i$ and $u_i$ represent respectively unstructured and structured spatial variability terms (see Besag, York and Mollié, 1991), $\theta_t$ is the effect of the $t$-th cohort and $\psi_{it}$ is an area specific cohort effect. The prior distributions for $v_i$, $i = 1, 2, \ldots, N$, $u_i$, $i = 1, 2, \ldots, N$ and $\theta_t$, $i = 1, 2, \ldots, T$ are multivariate normal distributions with mean zero and precision matrix respectively $\tau_v \mathbf{K}_v$, $\tau_u \mathbf{K}_u$ and $\tau_\theta \mathbf{K}_\theta$. The structure matrices (Clayton, 1996) $\mathbf{K}_v$, $\mathbf{K}_u$ and $\mathbf{K}_\theta$ were specified following the different nature of effects (for $u_i$ and $\theta_t$ a conditional autoregressive model was used, see Besag, York and Mollié, 1991, Bernardinelli *et al.*, 1995, Clayton, 1996).
Different specifications are possible for the interaction term, depending on the assumptions about their dependence structure. In our model, we assumed that interaction terms are structured both in space and time. The prior distribution is multivariate normal with mean zero and precision matrix $\tau_\psi \mathbf{K}_\psi$. The structure matrix $\mathbf{K}_\psi$ was defined as the Kroneker product between the matrices $\mathbf{K}_u$ and $\mathbf{K}_\theta$ (Clayton, 1996 and Knorr-Held, 2000). For the hyperparameters $\tau_v$, $\tau_u$, $\tau_\theta$ and $\tau_\psi$ non informative gamma prior distributions were used.

## 3 Results

Figure 1a and 1b show the lung cancer risk map for Tuscany in the periods 1971-74 and 1990-94. The maps points out a gradient of risk, from south-east to north-west. Figure 1c is the map of risk variation between the two calendar periods obtained by applying a simple two-period model with spatial interactions (Bernardinelli *et al.*, 1995). The map shows a general increase of lung cancer risk, stronger in the south-eastern part of the region.




(a) (b) (c)

FIGURE 1. a) Relative risk 1971-74; b) Relative Risk 1990-94; c) Risk variation between 1971-74 and 1990-94.

Figure 2 reports the pattern of risk by birth cohort for the whole study area. We can see that the mortality risk for lung cancer begins to decrease from the cohort of those born in 1930-35, whose exposure to smoking begun presumably in the middle '50. The epidemic reached its maximum for the cohort born from 1920 to 1930.

Figure 3a and 3b show the spatial distribution of relative risk for birth cohorts 1900-05 and 1920-25 estimated using model 1. The two maps have been chosen because they correspond to the cohorts with the lowest and highest risk over the study area: in both maps the risk is characterised by a strong south-east to north-west gradient, as already evidenced by the period analysis. In figure 3c is showed the risk difference between the birth cohorts 1925-30 and 1920-25 resulting from the model is shown. The areas with highest mortality (northern part of the region) show a decrease of risk, while the areas in the south-eastern part of the region behave in an opposite way.

We have estimated also a model without spatial-cohort interaction. This model may be compared with the one containing the interaction term using the Deviance Information Criterion (DIC) (see Spiegelhalter, Best and

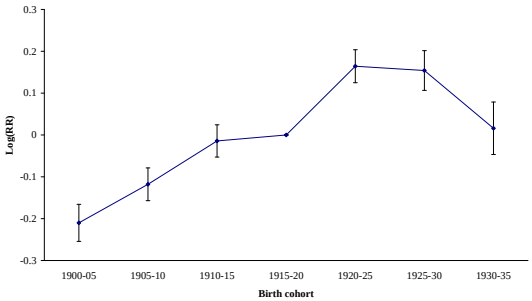
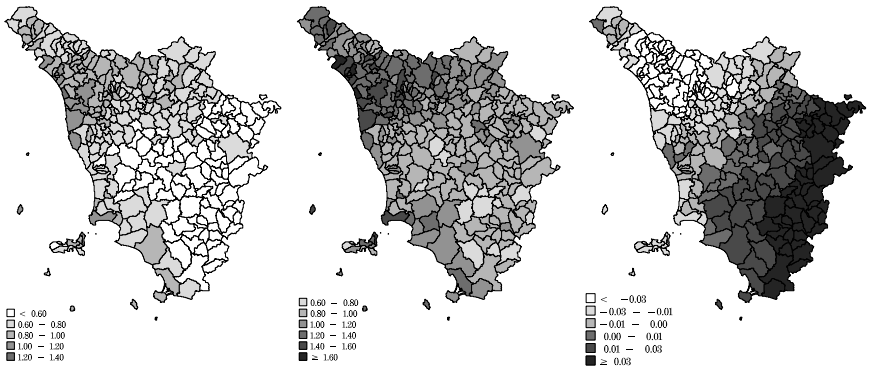



FIGURE 2. Cohort Relative Risk.




(a) (b) (c)

FIGURE 3. a) Lung cancer relative risk, birth cohort 1900-05; b) birth cohort 1920-25. ; c) risk variation between the birth cohorts 1925-30 and 1920-25.

Carlin, 1998), a generalisation of the Akaike's Information Criterion in Bayesian framework. This is defined as DIC= $D(\bar{\theta}) + 2(\bar{D} - D(\bar{\theta}))$, where $\bar{D}$ is the posterior expectation of the deviance and summarises the fit of the model, and $\bar{D} - D(\bar{\theta})$ is the expected deviance minus the deviance evaluated at the posterior expectations that represents the effective number of parameters and measures the complexity of the model. DIC value for the spatial-cohort model without interaction is 2577.18, considering the interaction is 2470.48. That confirms the presence of interaction between spatial and cohort effects and so the preference for this model.
In table 1 we reported the deviances of different age-period-cohort models (see Clayton and Shifflers, 1987) for each Tuscany Province. As expected,

the best fitting model for Tuscany as whole was the fully parametrized age-period-cohort model (APC).

TABLE 1. Deviances and degrees of freedom of age (A), age-drift (AD), age-period (AP), age-cohort (AC) and age-period-cohort (APC) models for each Tuscany Province and for the whole Tuscany.

| Province | A (72) | AD (71) | AP (68) | AC (51) | APC(48) |
|---|---|---|---|---|---|
| Arezzo | 191.58 | 100.78 | 93.04 | 28.22 | 24.79 |
| Firenze | 337.76 | 245.31 | 176.25 | 64.88 | 31.68 |
| Grosseto | 155.84 | 101.02 | 84.19 | 41.96 | 32.73 |
| Livorno | 107.91 | 94.94 | 90.89 | 24.60 | 23.91 |
| Lucca | 231.73 | 178.25 | 148.56 | 47.34 | 36.07 |
| Massa | 117.59 | 86.56 | 75.20 | 56.66 | 48.89 |
| Pisa | 241.07 | 147.18 | 128.27 | 41.66 | 35.27 |
| Pistoia | 137.82 | 81.57 | 62.89 | 30.70 | 21.24 |
| Siena | 205.02 | 108.10 | 80.83 | 70.55 | 51.35 |
| whole Tuscany | 1212.39 | 702.49 | 525.64 | 124.58 | 46.49 |

Considering the simpler age-period (AP) or age-cohort (AC) models the full model can be interpreted as evidence of an interaction between either age$*$cohort or age$*$period. Therefore when we combine data from each Province different cohort or period patterns by Province could result in a significant interaction by age.
We analyzed the data for each Province separately to evaluate age, period, cohort models in homogeneous populations. The findings reported in table 1 show that the age-cohort model (AC) has the best fitting. They are also in agreement with the existence of strong spatial interactions between time (by cohort) and space (by Municipality).

## 4 Conclusions

In analyzing time trends it is of utmost importance to identify the relevant time axis for the phenomenon under study. Since observations are taken from the Lexis diagram (age by calendar period) and subjects move diagonally by year of birth (cohort) we face at least three time dimensions. The usual approach to disease mapping collapse the age dimension by means of modelling standardized occurence ratios (SMR or SIR).
We proposed a general approach to fit age-cohort models in the context of space-time modelling of disease patterns.
The Tuscany example shows how apparent period effects can be explained by birth cohort trends which have different peaks in different geographically defined populations. The time trend observed could be due to an increase in cigarette consumption after the second world war, when also the most backward and isolated areas adopted smoking habits similar to those of the

most developed part of the region. They could reflect as well the effect of stopping cigarette consumption from the seventies in the most developped areas of the Region.
More interesting, while the analysis based on calendar periods points out an overall increase of risk, the birth cohort model evidences a decrease of relative risk for people born after 1930. Moreover, the highest level of the epidemic curve shows an evident spatial pattern, with a strong north-west/south-east gradient.

**Acknowledgements:** The research was partially supported by MURST (COFIN 1999, Lagazio) and by Department Funds 1998/99 (Biggeri).

## References

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M. (1995). Bayesian analysis of space-time variation in disease risk, *Statistics in Medicine*, **14**, 2433-2443.

Mason, T.J., McKay, F.W., Hoover, R., Blot, W.J., Fraumeni, J.F. (1975). *Atlas of cancer mortality for U. S. Counties: 1950-1969*, USGPO, Washington.

Clayton, D., Schifflers, E. (1987). Models for temporal variation in cancer rates. I: age-period and age-cohort models. II: age-period-cohort models, *Statistics in Medicine*, **6**, 449-481.

Clayton, D. (1996). Generalized linear mixed models, in: *Markov Chain Monte Carlo in practice*, Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (Eds.), Chapman & Hall, 275-301.

Knorr-Held, L., Besag, J. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine*, 17, 2045-2060.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine*, to appear.

Lilienfeld, A.M., Lilienfeld, D.E. (1980). *Foundations of Epidemiology*, Oxford University Press.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P.(1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report 98-009, Division of Biostatistics, University of Minnesota.

# On Ill-Conditioned GEEs and Toward Unified Biased Estimation

Brian D. Marx

[1] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803 USA (bmarx@lsu.edu)

**Keywords:** Longitudinal data, PLS, PCA, Quasi-likelihood, Repeated measures, Signal regression

## 1 Introduction

I revisit the generalized estimating equations (GEEs) (Liang and Zeger, 1986) which consider the multivariate setting generated from correlated or clustered response variables that can arise either from longitudinal studies or by sampling within several clusters of units. In the spirit of the GLM, non-normal response variables are allowed. Regressors are available and adjustments are made for GLM parameter estimation through a modified score function to account for the presence of correlation. Unlike the GLM, full likelihood parameter estimation is not usually feasible. GEEs utilize quasi-likelihood that depends on only the mean and covariance structure of the response variable (Wedderburn, 1974). In view of these GLM extensions, there also has been an increased awareness and understanding of how collinear data problems extend from standard regression particularly into the GLM through an ill-conditioned Fisher information matrix (Lesaffre and Marx, 1993). Despite the popularity of ML parameter estimation in the GLM, ill-conditioned information can be responsible for lack of convergence, large estimated coefficient variances, poor prediction in certain regions, as well as deflating power for hypotheses concerning model assessment. Little or no work has been done to my knowledge focusing on the detrimental features of a (nearly) singular *working* Fisher information matrix and alternative estimation within the GEE framework. Additionally, I aim to show that many asymptotically biased estimators are members of a broader class of shrinkage estimators for the GLM, and these can be transplanted into the GEE framework. I divide these alternative estimators into two main groups: generalized fractional principal component estimators (GFPCE) and penalized quasi-likelihood estimators (PQLE). GFPC estimation is accomplished by taking a general weighting of the principal component variables, whereas PQL estimation in many

cases broadens ridge type estimation with a variety of clever penalizations. I link these two groups together.

## 2 Marginal Models: Background and Notation

Marginal models are used in situations where the primary research interest is to analyze the *marginal* mean of the response given the explanatory variables. The association between the responses is often of secondary interest or even perhaps a nuisance. Consider $m_i$ repeated measures on unit $i$, $i = 1, \cdots, n$. Let $y_i = (y_{i1}, \cdots, y_{im_i})'$ be the vector of (exponential family) responses and $X_i = (x_{i1}, \cdots, x_{im_i})'$ be the corresponding $m_i \times (p+1)$ matrix of regressors (including an intercept term). The specifications for marginal models in a non-normal setting are presented in FAHRMEIR and TUTZ (1994, sec. 3.5.2) and can be outlined as: (i) The marginal means $\mu_{ij} = E(y_{ij}|x_{ij}) = h(\eta_{ij})$, where $h(\cdot)$ is the (monotone and twice differentiable) inverse link function, $\eta_{ij} = x'_{ij}\beta$ and $\beta$ is the unknown coefficient vector; (ii) The marginal variance is function of $\mu_{ij}$, $\mathrm{var}(y_{ij}|x_{ij}) = \sigma^2(\mu_{ij}) = \phi v(\mu_{ij})/\omega_{ij}$, where $\phi$ and $v(\mu_{ij})$ represents the scale parameter and variance function, respectively, determined by the specific exponential family member; and (iii) To account for within unit dependence, the covariance between $y_{ij}$ and $y_{ij'}$ is a function of the marginal means and possibly an additional association parameters $\theta$. For a known function $\zeta$, $\mathrm{cov}(y_{ij}, y_{i'j'}) = \zeta(\mu_{ij}, \mu_{i'j'}, \theta)$ for $i = i'$, and uncorrelated for $i \neq i'$. Thus for unit $i$, a $m_i \times m_i$ *working* covariance matrix is defined $\mathrm{cov}(y_i) = \Sigma_i(\beta, \theta)$. It is convenient to express $\Sigma$ $(m_i \times m_i)$ in terms of the correlation matrix $R$, i.e. $\Sigma_i = A_i^{1/2} R(\theta) A_i^{1/2}$ with $A_i = \mathrm{diag}(\sigma_{i1}^2, \cdots, \sigma_{im_i}^2)$. Several choices of $R$ are common: uncorrelated repeated observations (independence), fully unspecified, exchangable, or auto-regressive.

## 3 Score Functions and Quasi-likelihood

The generalized estimating equations for $\beta$ can be expressed (for fixed $\theta$)

$$s_\beta(\beta, \theta) = \sum_{i=1}^{n} X_i' D_i(\beta) \Sigma_i^{-1}(\beta, \theta)(y_i - h(\eta_i)). \tag{1}$$

The matrices $X_i$ and $D_i = \mathrm{diag}(h'(\eta_{ij}))$ are of dimension $m_i \times (p+1)$ and $m_i \times m_i$, respectively. In general, some alternating estimation between $(\theta, \phi)$ and $\beta$ iterations is needed in (1). Given current estimates, say $(\hat{\theta}, \hat{\phi})$, (1) is set to zero and solved by $\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + (\hat{F}^{(t)})^{-1} \hat{s}^{(t)}$, where the estimated *working* Fisher matrix, $\hat{F}^{(t)} = \sum_{i=1}^{n} X_i' D_i(\hat{\beta}^{(t)}) \Sigma_i^{-1}(\hat{\beta}^{(t)}, \hat{\theta}, \hat{\phi}) D_i(\hat{\beta}^{(t)}) X_i$, and $\hat{s}^{(t)} = s(\hat{\beta}^{(t)}, \hat{\theta}, \hat{\phi})$. Under regularity conditions with $\hat{\theta}$ fixed, consistent estimates of $\beta$ are ensured, and asymptotically $\hat{\beta} \sim N(\beta, F^{-1} V F^{-1})$ for

the mentioned $R(\theta)$. The matrices $F = \sum_{i=1}^n X_i' D_i \Sigma_i^{-1} D_i X_i = X'\Omega X$, $V = \sum_{i=1}^n X_i' D_i \Sigma_i^{-1} \mathrm{cov}(y_i) \Sigma_i^{-1} D_i X_i$, $\Omega =$ block diagonal$(\Omega_i)$, and $\Omega_i = D_i \Sigma_i^{-1} D_i$. In practice, consistent estimation for $\beta$ is achieved by using $\hat{V}$ for $V$: substituting converged $\hat{\beta}$ into $D_i$, $(\hat{\theta}, \hat{\phi})$ into $\Sigma_i$, and using $(y_i - h(\hat{\eta}_i))(y_i - h(\hat{\eta}_i))'$ for $\mathrm{cov}(y_i)$. Thus the estimated covariance matrix, often referred to as the *sandwich matrix*, is

$$\hat{H} = \hat{F}^{-1} \sum_{i=1}^n X_i' \hat{D}_i \hat{\Sigma}_i^{-1} (y_i - h(\hat{\eta}_i))(y_i - h(\hat{\eta}_i))' \hat{\Sigma}_i^{-1} \hat{D}_i X_i \hat{F}^{-1}. \tag{2}$$

If the working covariance structure is correctly specified, then estimation is asymptotically efficient. It can be useful to re-express the iterative solution as

$$\hat{\beta}^{(t+1)} = (\hat{F}^{(t)})^{-1} X' \hat{\Omega}^{(t)} \hat{y}^{\star(t)}, \tag{3}$$

where $y^\star = \eta + D^{-1}(y - h(\eta))$ is the adjusted dependent vector and $D = \mathrm{diag}(D_i)$.

## 4 A Unifying GEE Template

It will be useful to work with the principal components for each observation, $Z$, where the $((i,j),k)$th element of $Z$ is the score of the $k$th principal component for the $(i,j)$th observation. Define the principal components $Z = X_{-1}M$, where $X_{-1}$ $(N \times p)$ contains regressors, and $M$ is the $p \times p$ orthogonal matrix whose $k$th column is the $k$th eigenvector of the information matrix $F_{-1}$ constructed without the intercept, $k = 1, \cdots, p$. Hence $M'F_{-1}M = \mathrm{diag}(\lambda_k) = \Lambda$, where $\lambda_k$ are the corresponding eigenvalues. It is clear that near-singular $\hat{F}$ can have a dramatic impact on the stability, and even existence, of the sandwich covariance matrix of $\hat{\beta}$ in (2), as well as the iterative algorithm in (3). I recommend to standardize such that estimation of the intercept coefficient is uncorrelated with estimation of the other coefficients. In general this can be achieved by centering and scaling $X_{-1}$ (without intercept) using a weighted mean and sum of squares, respectively. The $N = \sum_{i=1}^n m_i$ vector of weights can be computed (iteratively) by defining the ($i$th, $j$th) weight as the $j$th column total $\tau_{ij}$ of $\Omega_i$. To introduce a class of estimators, we generalize the model further through a matrix $\Gamma$. Let

$$\eta_{ij}^* = \beta_0 + z_{ij}' \Gamma \alpha = \beta_0 + z_{ij}' \alpha^*, \tag{4}$$

where $\alpha^* = \Gamma\alpha$. The matrix $\Gamma = \mathrm{diag}(\gamma_k)$ is a diagonal weight matrix with $\gamma_k$ usually contained in the closed unit interval. For given $\Gamma$, GEE strategies may be applied to estimate $\alpha^*$ in (4). One potentially expensive candidate iterative scheme can be defined as

$$\tilde{\eta}^{*(t+1)} = \tilde{\beta}_0^{(t)} \mathbf{1}_N + \tilde{Z}^{(t)} \Gamma \tilde{\Lambda}^{-1(t)} \tilde{Z}^{(t)\prime} \tilde{\Omega}^{(t)} \tilde{D}^{(t)} \tilde{y}^{\star(t)}, \tag{5}$$

where $\tilde{e}^* = y - h(\tilde{\eta}^*)$ and the entries of the adjusted dependent variable are $\tilde{y}^\star = \tilde{\eta}^* + \text{diag}(h'(\tilde{\eta}^*))^{-1}\tilde{e}^*$. Equation (5) is particularly taxing because new eigen-decomposition is required at each iteration through $\tilde{\Lambda}$ and $\tilde{Z} = X_{-1}\tilde{M}$. Some relief is possible if the converged GEE estimated parameter estimates and working Fisher information matrix both exist from (3) and (2) respectively. Define the matrices $\hat{M}$ and $\hat{\Lambda}$ to be the respective eigenvector matrix and diagonal matrix of eigenvalues of the converged $\hat{F} = X'_{-1}\hat{\Omega}X_{-1}$. The intercept $\hat{\beta}_0$ is the weighted $(\hat{\tau}_{ij})$ mean of the converged adjusted dependent variable $\hat{y}^\star$. Whenever possible I recommend substitution of $\hat{\beta}_0$, $\hat{Z}$, $\hat{\Lambda}$, $\hat{\Omega}$, and $\hat{D}$ into (5):

$$\tilde{\eta}^{*(t+1)} = \hat{\beta}_0 \mathbf{1}_N + \hat{Z}\hat{\Gamma}\hat{\Lambda}^{-1}\hat{Z}'\hat{\Omega}\hat{D}\tilde{y}^{\star(t)}. \tag{6}$$

A conversion can be made using $\tilde{\beta} = \hat{M}\tilde{\alpha}^*$, with $\text{var}(\tilde{\beta}) = M'\Gamma MHM\Gamma M'$. With proper choices of $\Gamma$: GEE, (fractional) PC, Stein, Sclove estimation, among others, are united. Choices for the number of components or shrinkage parameters can be made on cross-validation, information criterion, or minimization of estimation or prediction oriented criterion.

## 5 Penalized Quasi-likelihood Approaches

Consider modifying (1) such that penalized quasi-likelihood estimators can be generally expressed as

$$\tilde{\beta}_\kappa^{PQL} = \mathcal{Z}_\beta\{s(\beta,\theta) - \kappa\frac{\partial P(\beta)}{\partial \beta}\},$$

where $\mathcal{Z}\{\cdot\}$ is the zero solution, $P(\beta)$ is a penalty function for the coefficient vector and $\kappa$ is the non-negative regularization parameter. For example, there exist experimental situations when the set of regressors have some ordering, and it is reasonable to assume that adjacent coefficients cannot differ too much from each other. Eilers and Marx (1996) imposed a penalization scheme to B-spline coefficients in a variety of smoothing applications using penalized likelihood (P-splines). Such notions of penalized estimation can be extended into the GEE setting. The penalty matrix is constructed using the $d$th order difference operator $\Delta^d$, $d = 0, 1, \cdots, p-1$. Define $\Delta^1(\beta_k) = \beta_k - \beta_{k-1}$. Higher order differences can be found by induction and in general can be expressed as $P^d\beta$, where $P^d$ is the $(p-d)\times p$ banded matrix constructed by taking $d$ row differences of $I_p$. Estimation involves penalizing the score in (1) for fixed $(\hat{\theta},\hat{\phi})$,$\tilde{\beta}^D = \mathcal{Z}_\beta\{s(\beta,\theta) - \kappa P^{d'}P^d\beta\}$, where $\kappa \geq 0$. The PQL-solution results in modifying (1) as

$$\tilde{\beta}^{D(t+1)} = (\tilde{F}^{(t)} + \kappa P^{d'}P^d)^{-1}X'\tilde{\Omega}^{(t)}\tilde{y}^{\star(t)}. \tag{7}$$

Various penalized likelihood GLM estimators are extended to the GEE framework through PQL, namely: P-spline, ridge, garrote (Breiman, 1995), lasso (Tibshirani, 1996), bridge (FU, 1998), among others.

## 6 High Dimensional and Functional Regressors

I propose (iterative) GEE extensions to partial least squares (PLS) and signal regression. Consider situations when modern technology generates regressors (spectra, log-periodograms, time series, etc.). Such information comes in the form of hundreds or thousands of discrete digitizations of some signal, often resulting in $p \gg N$. JAMES and HASTIE (1999) provided an interesting subject-specific mixed modeling approach using principal component techniques for sparse functional data. Consider rewriting the marginal mean as

$$\mu = h(\beta_0 + X_{-1}\beta) = h(\eta). \tag{8}$$

This problem is highly ill-conditioned, and the only hope to get a sensible result is by constraining $\beta$ in some way. MARX and EILERS (1999) proposed a (GLM) P-spline modeling strategy, as a competitor to (IR)PLS, that forces $\beta$ to be smooth. The dimension of the *signal* coefficient vector is reduced initially by projecting it onto a B-basis, $B$ (of smooth functions): $\beta_{p\times 1} = B_{p\times q}\delta_{q\times 1}$, where $q < \min(N, p)$. Notice that this approach takes advantage of the spatial or temporal information along the signal and has an attractive linear nature of $\beta$, where $\delta$ is a relatively low dimensional vector of B-spline coefficients.

P-splines take one step further: use a moderate number of equally spaced B-spline knots (say 10 to 40) and further increase smoothness by imposing a difference penalty on $\delta$ (as in Section 6.1). Notice that (8) can be rewritten as

$$\mu = h(\beta_0 + X_{-1}B\delta) = h(\beta_0 + U\delta),$$

where we can define a new full rank regression matrix $U_{N\times q} = X_{-1}B$. Now re-express the PQL solution as

$$\tilde{\delta}^D = \mathcal{Z}_\delta\{s(\delta,\theta) - \kappa P^{d'}P^d\delta\},$$

where

$$s(\delta,\theta) = U'\Omega(\delta,\theta)D^{-1}(\delta)(y - h(U\delta)), \tag{9}$$

and $\kappa \geq 0$. The information matrix is $F_U = U'\Omega U$ (with $M_U' F_U M_U = \Lambda_U$), and $\tilde{\delta}^D$ can be estimated with an algorithm similar to (7). The penalized solutions are given upon convergence as $\tilde{\beta} = B\tilde{\delta}_\kappa$. Suggestions for optimization of $\kappa$ are given in Marx and Eilers using cross-validation and information criterion. Arguments can link this setting to GFPC estimation.

Simple data management tricks can make fitting such GEE models trivial in existing (GLM) GEE software and macros that allow a variety of correlation structures, by the *unit* variable. Consider constructing a $p \times q$ B-spline matrix using a modest number of equally spaced knots along the

indexing domain (e.g. frequency) of the signal. Thus $U$ is accessible. Instead of passing the $y$ response and the signal matrix $X_{-1}$ into the GEE fitting algorithm, use the augmented matrix $U_{\text{aug}} = \text{rbind}(U_{+1}, (0, \kappa P^d))$ and the augmented adjusted dependent variable $y^{\star}_{\text{aug}} = \text{rbind}(y^{\star}, 0_{q-d})$, where $U_{+1}$ includes the (unpenalized) intercept. Thus the software can automatically provide a penalized estimate of $\delta_\kappa$, and $\tilde{\eta} = U_{+1}\tilde{\delta}_\kappa$. Notice also that $H_{\tilde{\beta}} = BH_{\tilde{\delta}}B'$.

## Acknowledgments

I would like to thank Ludwig Fahrmeir, Lynn LaMotte and Rob Tibshirani for valuable discussions.

## References

Breiman, L. (1995). Better Subset Selection Using the Non-negative Garrote. *Technometrics*, **37**, 373-384.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing Using B-Splines and Penalized Likelihood (with comments and rejoinder). *Statistical Science* **11**(2): 89-121.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Fu, W.J. (1998). Penalized Regressions: The Bridge versus the Lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416.

James, G. and Hastie, T. (1999). Principal Component Models for Sparse Functional Data. Technical Report, Department of Statistics, Stanford University.

Lesaffre, E. and Marx, B.D. (1993). Collinearity in Generalized Linear Regression. *Communications in Statistics: Theory and Methods* **22**(7): 1933-1952.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal Data Analysis using Generalized Linear Models. *Biometrika* **73**: 13-22.

Marx, B.D. (1996). Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression. *Technometrics*, **38**, 374-381.

Marx, B.D. and Eilers, P.H.C. (1999). Generalized Linear Regression on Sampled Signals and Curves: A P-Spline Approach. *Technometrics* **41**: 1-13.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, B*, **58**, 267-288.

Wedderburn, R.W.M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika* **61**, 439-447.

# Strategies to Fit Pattern-Mixture Models

Geert Molenberghs[1], Herbert Thijs[1] Geert Verbeke[2], Bart Michiels[3] and Desmond Curran[4]

[1] Biostatistics, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, Building D, B-3590 Diepenbeek, Belgium; E-mail: geert.molenberghs@luc.ac.be
[2] Katholieke Universiteit Leuven, Belgium
[3] Janssen Research Foundation, Beerse, Belgium
[4] Icon Clinical Research, Dublin, Ireland

**Abstract:** Whereas most models for incomplete longitudinal data are formulated within the selection model framework, pattern-mixture models have gained considerable interest in recent years (Little 1993, 1994). In this paper, we outline several strategies to fit pattern-mixture models, including the so-called identifying-restrictions strategy. Multiple imputation is used to apply this strategy to a realistic settings, such as quality-of-life data from a longitudinal study on metastatic breast cancer patients.

**Keywords:** Delta Method; Linear Mixed Model; Longitudinal Data; Missing Data; Multiple Imputation; Selection Model.

## 1 Introduction

Most methods for incomplete longitudinal data are formulated within the selection modeling frame (Little and Rubin 1987) as opposed to pattern-mixture modeling (PMM; Little 1993, 1994). A selection model factors the joint distribution of the measurement and response mechanisms into the marginal measurement distribution and the response distribution, conditional on the measurements. This is intuitively appealing since the marginal measurement distribution would be of interest also with complete data. Further, Little and Rubin's taxonomy is most easily developed in the selection setting. However, it is often argued that, especially in the context of non-random missingness models, selection models, although identifiable, should be approached with caution (Glynn, Laird and Rubin 1986). Therefore, pattern-mixture models have gained renewed interest in recent years (Little 1993, 1994, Hogan and Laird 1997). Several authors have contrasted selection models and pattern-mixture models. This is done to either (1) answer the same scientific question, such as marginal treatment effect or time evolution, based on these two rather different modeling strategies, or (2) to gain additional insight by supplementing the selection model results with those from a pattern-mixture approach.

An important issue is that pattern-mixture models are by construction under-identified. Little solves this problem through the use of identifying restrictions: inestimable parameters of the incomplete patterns are set equal to (functions of) the parameters describing the distribution of the completers. Identifying restrictions are not the only way to overcome under-identification and we will discuss alternative approaches as well.

## 2 Models for Incomplete Longitudinal Data

We restrict attention to short series of repeated measures, for which the following longitudinal model is adequate:

$$Y_i = X_i\beta + \varepsilon_i \tag{1}$$

where $Y$ is the $n_i$ dimensional response vector for subject $i$ with components $Y_{ij}$, $1 \leq i \leq N$, $N$ is the number of subjects, $X_i$ is a $(n_i \times p)$ dimensional matrix of known covariates, $\beta$ is the $p$ dimensional vector containing the fixed effects, and $\varepsilon_i \sim N(0, \Sigma)$ is the vector of correlated error terms.

We additionally define, for each subject $i$, $D_i$ to be the dropout indicator, one higher than the occasion of the last obtained measurement. We split the vector $Y_i$ into observed ($Y_i^o$) and missing ($Y_i^m$) components respectively. The taxonomy of Little and Rubin (1987), is based on the second factor of the selection model factorization:

$$f(y_i, d_i|\theta, \psi) = f(y_i|\theta)f(d_i|y_i, \psi), \tag{2}$$

with obvious notation. In contrast, the pattern-mixture family is based upon:

$$f(y_i, d_i|\theta, \psi) = f(y_i|d_i, \theta)f(d_i|\psi). \tag{3}$$

The measurement model has to reflect dependence on dropout. Thus, the parameters in (1) become pattern-dependent: $\beta(d_i)$ and $\Sigma(d_i)$. The dependence of parameters on dropout can be done in several ways.

Assume a complete measurement sequence is of length $n$. Recall that the classical taxonomy considers the structure of $f(d|y)$. The missing data are MAR if a subject's missingness mechanism depends on its observed outcomes only, $f(d = t{+}1|y_1, \ldots, y_n) = f(d = t{+}1|y_1, \ldots, y_t)$, for $t = 1, \ldots, n$. Little's (1993) complete case misising value (CCMV) restrictions are given by ($t \geq 2$, $j < t$):

$$f(y_t|y_1, \ldots, y_{t-1}, d = j + 1) = f(y_t|y_1, \ldots, y_{t-1}, d = n + 1),$$

whereas ACMV (available case missing values) is the condition that

$$f(y_t|y_1, \ldots, y_{t-1}, d = j + 1) = f(y_t|y_1, \ldots, y_{t-1}, d > t). \tag{4}$$

Molenberghs, Michiels, Kenward, and Diggle (1998) have shown that, for longitudinal data with dropouts, MAR $\Longleftrightarrow$ ACMV.

## 3 Pattern-Mixture Models and Sensitivity Analysis

The key area where sensitivity analysis should be focused is on the unidentified components of the model and the way(s) in which this is handled.

- **Strategy 1.** Little (1993, 1994) advocated the use of identifiying restrictions, such as CCMV and NCMV.
- **Strategy 2.** Trends can be restricted to functional forms supported by the information available within a pattern. For example, a linear or quadratic time trend is easily extrapolated beyond the last obtained measurement. In order to fit such models, one simply has to carry out a model building exercise within each of the patterns separately.
- **Strategy 3.** One can let the parameters vary across patterns in a controlled parametric way. Thus, rather than estimating a separate time trend within each pattern, one could for example asssume that the time evolution within a pattern is unstructured, but parallel across patterns. This is effectuated by treating pattern as a covariate. The available data can be used to assess whether such simplifications are supported within the time ranges for which there is information.

## 4 Identifying Restriction Strategies

We will briefly sketch the strategy. Several points which require further specification will be discussed in subsequent sections. (1) Fit a model to the pattern-specific identifiable densities: $f_t(y_1, \ldots, y_t)$. This results in a parameter estimate, $\hat{\gamma}_t$. (2) Select an identification method of choice. (3) Using this identification method, determine the conditional distributions of the unobserved outcomes, given the observed ones:

$$f_t(y_{t+1}, \ldots, y_T | y_1, \ldots, y_t). \tag{5}$$

(4) Draw multiple imputations for the unobserved components, given the observed outcomes and the correct pattern-specific density (5). (5) Analyze the multiply-imputed sets of data using the method of choice. This can be another pattern-mixture model, but also a selection model or any other desired model. Conduct classical multiple-imputation inference.
For ease of understanding, consider the special case of 3 measurements. In this case, there are only three patterns and identification takes the following form:

$$\begin{aligned} f_3(y_1, y_2, y_3) &= f_3(y_1, y_2, y_3), && (6)\\ f_2(y_1, y_2, y_3) &= f_2(y_1, y_2) f_3(y_3 | y_1, y_2), && (7)\\ f_1(y_1, y_2, y_3) &= f_1(y_1) \left[\omega f_2(y_2|y_1) + (1-\omega) f_3(y_2|y_1)\right] \\ &\quad \times f_3(y_3|y_1, y_2). && (8) \end{aligned}$$

Since $f_3(y_1, y_2, y_3)$ is completely identifiable from the data, and for $f_2(y_1, y_2, y_3)$ there is only one possible identification; the only place where a choice has to be made in pattern 1. Setting $\omega = 1$ corresponds to NCMV, while $\omega = 0$ implies CCMV. ACMV corresponds to

$$\omega = \frac{\alpha_2 f_2(y_1)}{\alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}. \tag{9}$$

The conditional density $f_1(y_2|y_1)$ in (8) can be rewritten as

$$f_1(y_2|y_1) = \frac{\alpha_2 f_2(y_1, y_2) + \alpha_3 f_3(y_1, y_2)}{\alpha_2 f_2(y_1) + \alpha_3 f_3(y_1)}.$$

## 5 The Vorozole Study

This study was an open-label, multicenter, parallel group design conducted at 67 North American centers. Patients were randomized to either vorozole taken once daily) or megestrol acetate. The patient population consisted of postmenopausal patients with histologically confirmed estrogen-receptor positive metastatic breast carcinoma. All 452 randomized patients were followed until disease progression or death. Full details of this study are reported in Goss *et al* (1999). This paper focuses on overall quality of life, measured by the total Functional Living Index: Cancer. Precisely, a higher FLIC score is the more desirable outcome.

In order to concisely illustrate the methodology described in this chapter, we will apply it to the vorozole study, restricted to those subjects with 1, 2, and 3 follow up measurements, respectively. Thus, 190 subjects are included into the analysis, with subsample sizes 35, 86, and 69, respectively. The corresponding pattern probabilities are

$$\widehat{\pi} = (0.184, 0.453, 0.363)'. \tag{10}$$

These figures, apart from giving a feel for the relative importance of the various patterns, will be needed to calculate marginal effects (such as the marginal treatment effect) from pattern-mixture model parameters.

We will apply each of the three strategies, presented in Section 4. In order to apply the identifying restriction Strategy 1, one first needs to fit a model to the observed data. We will opt for a simple model, with parameters specific to each pattern. Such a model can be seen as belonging to the second modeling strategy. Next, restrictions are applied. Finally, in the third strategy, *pattern* is included as a covariate. An initial model is simplified using $F$ tests. A graphical impression of these models is given in Figure 1. These findings suggest, again, that a more careful reflection on the extrapolation method is required. This is very well possible in a pattern-mixture context, but then the first strategy, rather than the second and third strategies, has to be used.

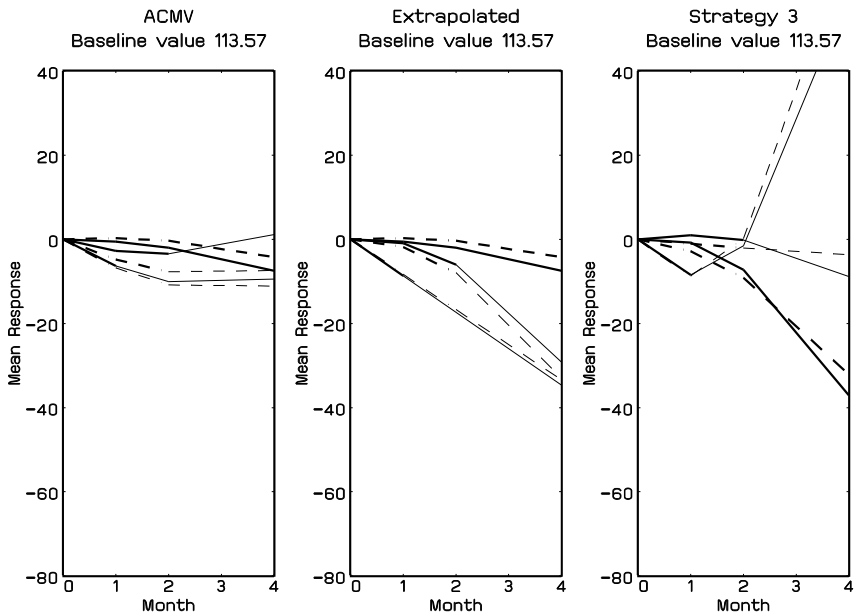



FIGURE 1. *Vorozole Study. For average level of baseline value, extrapolation based on initial model (strategy 2), ACMV, and strategy 3 is shown. The bold portion of the curves runs from baseline until the last obtained measurement, and the extrapolated piece is shown in thin type. The dashed line refers to megestrol acetate; the solid line is the Vorozole arm.*

## 6 Concluding Remarks

In this paper, we have illustrated three distinct strategies to fit pattern-mixture models. In this way, we have brought together several existing practices. Little has proposed identifying restrictions, which we here formalized using the connection with MAR and multiple imputation.

By contrasting these strategies on a single set of data, one obtains a range of conclusions rather than a single one, which provides insight into the sensitivity to the assumptions made. Especially with the identifying restrictions, one has to be very explicit about the assumptions and moreover this approach offers the possibility to consider several forms of restrictions. Special attention should go to the ACMV restrictions, since they are the MAR counterpart within the pattern-mixture context.

In addition, a comparison between the selection and pattern-mixture modeling approaches is useful to obtain additional insight into the data and/or to assess sensitivity.

The identifying restrictions strategy provides further opportunity for sen-

sitivity analysis. Indeed, since CCMV and NCMV are extremes, it is very natural to consider the idea of *ranges* in the allowable space of $\omega_s$. Clearly, any $\omega_s$ which consists of non-negative elements that sum to one is allowable, but also the idea of extrapolation could be useful, where negative components are allowed, given they provide valid conditional densities.

**Acknowledgements:** We gratefully acknowledge support from *Fonds Wetenschappelijk Onderzoek-Vlaanderen* Research Project G.0002.98 "Sensitivity Analysis for Incomplete and Coarse Data", and from *Vlaams Instituut voor de Bevordering van het Wetenschappelijk-Technologisch Onderzoek in Industrie.*

## References

Glynn, R.J., Laird, N.M., and Rubin, D.B (1986) Selection Modelling versus mixture modeling with nonignorable nonresponse. In *Drawing Inferences from Self-Selected Samples*, Ed. H. Wainer, pp. 115–142. New York: Springer Verlag.

Goss, P.E., Winer, E.P., Tannock, I.F., and Schwartz, L.H. (1999) Breast cancer: randomized phase III trial comparing the new potent and selective third-generation aromatase inhibitor vorozole with megestrol acetate in postmenopausal advanced breast cancer patients. *Journal of Clinical Oncology*, **17**, 52–63.

Hogan, J.W. and Laird, N.M. (1997) Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine*, **16**, 239–258.

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R.J.A. (1994) A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

Little, R.J.A. and Rubin, D.B. (1987) *Statistical Analysis with Missing Data.* New York: Wiley.

Molenberghs, G., Michiels, B., Kenward, M.G., and Diggle, P.J. (1998) Missing data mechanisms and pattern-mixture models. *Statistica Neerlandica*, **52**, 153–161.

# Identification of Vector AR and ARMA models with recursive structural errors using Conditional Independence Graphs

Marco Reale[1] and Granville Tunnicliffe Wilson[2]

[1] Mathematics and Statistics Dept., University of Canterbury, Private Bag 4800 Christchurch, New Zealand.
[2] Mathematics and Statistics Dept., Lancaster University, Lancaster LA1 4YF, UK.

**Abstract:** In canonical vector autoregressions, which permit dependence only on past values, the errors generally show contemporaneous correlation. By contrast structural vector autoregressions allow contemporaneous series dependence and assume errors with no contemporaneous correlation. Such models having a recursive structure can be described by a directed acyclic graph. We show how the identification of these models may be assisted by examination of the conditional independence graph of contemporaneous and lagged variables, and extend this to structural vector autoregressive moving average models. The structural modelling of the canonical errors alone is a useful initial step which we use to demonstrate the potential of the approach. The models are useful for indicating how shocks pass through the set of contemporaneous values, besides identifying a more meaningful and parsimonious relationship between the series.

**Keywords:** VARMA Models; Graphical Modelling; Granger Causality.

## 1 Introduction

The canonical $p$th order vector autoregressive model, VAR($p$), of a stationary, $m$ dimensional time series $x_t = (x_{t,1}, x_{t,2}, \ldots, x_{t,m})'$ is of the form:

$$x_t = c + \Phi_1 x_{t-1} + \Phi_2 x_{t-2} + \cdots + \Phi_p x_{t-p} + e_t \tag{1}$$

where $c$ allows for a non-zero mean of $x_t$ and $e_t$ is multivariate white noise with general covariance matrix $V$. Our working assumption is that the series is Gaussian but our methods should be applicable under wider conditions, such as $e_t$ being I.I.D., presented for example in Anderson (1971). This model is attractive because its estimation from a sample $x_1, x_2, \ldots, x_n$, by least squares applied separately to each component of $x_t$, is straightforward. For large sample length $n$ it is also fully efficient provided there are no subset constraints on these separate regressions, (Judge *et al* 1985), the properties of the estimates given by the regression are reliable, and the estimate of $V$ is independent of the estimates of $\Phi_k$.

There are various approaches to multiple time series modeling which seek either to transform models such as (1) to a form which includes contemporaneous relationships among the variables, or to identify directly such a form, see for example Box and Tiao (1977) and Tiao and Tsay (1989). Our aim in this paper is similar; our approach is to consider the structural autoregressive model of the same form as (1) but with the addition of contemporaneous dependence through a matrix coefficient $\Phi_0$:

$$\Phi_0 x_t = d + \Phi_1^* x_{t-1} + \Phi_2^* x_{t-2} + \cdots + \Phi_p^* x_{t-p} + a_t. \tag{2}$$

In our present context the equivalence between (1) and (2) is given by $\Phi_i^* = \Phi_0 \Phi_i$ and $\Phi_0 e_t = a_t$. A requirement of (2) is that the variance matrix $D$ of $a_t$ is diagonal. We require a further condition on $\Phi_0$, that it represent a recursive (causal) dependence of each component of $x_t$ on other contemporaneous components. This is equivalent to the existence of a reordering of the elements of $x_t$ such that $\Phi_0$ is triangular with unit diagonal. Each possible ordering of $x_t$ therefore gives a potentially distinct form of (2), but these are all statistically equivalent, corresponding to factorizations of

$$V^{-1} = \Phi_0' D^{-1} \Phi_0. \tag{3}$$

This contrasts with the unique form of (1), which is the attractive feature of that model from a time-series modeling viewpoint.

The value of (2) therefore lies in the possibility that there is one particular form which, as a consequence of its representing a true simple mechanism, is more parsimonious in its parameterization than either (1) or the other forms of (2). This would be reflected in the ability to exclude many of the elements of $\Phi_0$ and $\Phi_i^*$ from the model without penalizing the fit in comparison with the saturated forms of either (1) or (2). Identification of such a model may then provide added insight into the true mechanisms which generate the data.

That is what we seek to achieve by the method described in this paper. The relationship (3) shows that some, though not necessarily complete, information on the structure of $\Phi_0$ is available from the variance matrix of the innovations.

## 2 Recursive structure and partial correlations

Neglecting, for the present treatment, any effects of time series model estimation, we suppose that we have observations on the vector Gaussian white noise innovations process $e_t$ with the usual sample covariance matrix $\hat{V}$. We wish to determine *from the data* the form of possible sparse structural matrices $\Phi_0$ which are compatible with $\hat{V}$. There may be no unique such form without imposing further constraint using insight from the modeling context. Swanson and Granger (1997) consider an almost identical problem,

but focus more on testing for the constraints which are implied by a particular structural form of $\Phi_0$ which has commonly occurred in practice. Their tests are expressed in terms of partial autocorrelations which, as they remark, are not directional and would therefore appear less appropriate for recursive (causal) models. We also use pair-wise partial autocorrelations, but conditioning on all remaining variables (i.e. components of $e_t$) rather just one other variable at a time. This is because such partial correlations are used to construct the conditional independence graph (CIG) of the variables, following procedures presented for example in Whittaker (1990). As Swanson and Granger also remark, the structural form of dependence between the variables is naturally expressed by (and is equivalent to) a directed acyclic graph (DAG), in which nodes representing variables are linked with arrows indicating the direction of any causal dependence. A DAG implies a single CIG for the variables, but the possible DAGs which might explain a particular CIG may be several or none. The point is that, subject to sampling variability, the CIG is a constructible quantity and a useful one for expressing the data determined constraints on permissible DAG interpretations.

The CIG consists of nodes representing the variables, two nodes being *without* a link if and only if they are independent conditional upon *all* the remaining variables. In a Gaussian context this conditional independence is indicated by a zero partial autocorrelation:

$$\rho\left(e_{it}, e_{jt} | \{e_{kt}, k \neq i, j\}\right) = 0. \tag{4}$$

In the wider linear least squares context, defining linear partial autocorrelations as the same function of linear unconditional correlations as in the Gaussian context, (4) still usefully indicates lack of linear predictability of one variable by the other given the inclusion of all remaining variables. The link with Granger causality is quite evident. The set of all such partial correlations required to construct the CIG is conveniently calculated as

$$\rho\left(e_{it}, e_{jt} | \{e_{kt}, k \neq i, j\}\right) = -W_{ij}/\sqrt{(W_{ii}W_{jj})} \tag{5}$$

where $W = V^{-1}$. The sample values are obtained by substituting the sample value $\hat{V}$ of $V$.

The significance levels are obtained by using the relationship between a regression $t$ value and the sample partial correlation $\hat{\rho}$ given by $\hat{\rho} = t/\sqrt{(t^2 + \nu)}$ (see Greene, 1993,p 180). Here $\nu$ is the residual degrees of freedom in the regression of one of the variables in the partial autocorrelation, upon all the other variables. The $t$ value is that attached, in this regression, to the other variable in the partial autocorrelation. This is a relationship deriving from the linear algebra of least squares, and is not reliant upon statistical assumptions. Standard assumptions *are* needed to support the usual distribution of $t$ under the null hypothesis that the true value of the relevant variable is zero, which is equivalent to $\rho = 0$. There are of course statistical

pitfalls in applying the test simultaneously to all sample partial autocorrelations. Our attitude is similar to that advocated by Box and Jenkins (1976) for the identification, for example, of autoregressive models using time series partial autocorrelations. We use these values to suggest possible models; after fitting these we apply more formal tests and diagnostic checks to converge on an acceptable model.

Central to the interpretation of a CIG is the separation theorem. The CIG is *constructed* by *pairwise* separation of variables which are independent conditional on the remainder. The separation theorem states that if two *blocks* of variables are separated, i.e. there is no link between any member of the first block and any member of the second, then the two blocks are completely independent conditional on the remaining variables (see for example Whittaker (1990, pp 64-67)).

This brings us to the next step which is to determine what DAG structures can explain this CIG, and to estimate them. This is part of a much wider problem of the search for causal structure, covered for example by Spirtes *et al* (1993). The procedure to determine the CIG implied by a given DAG has become known as *moralization*, following Lauritzen and Spiegelhalter (1988). A node B in a DAG is a *parent* of a node A if there is an arrow *from* B directly linking *to* A. Moralization is the construction of a CIG by linking (marrying), for each node of a DAG, all of its parents. The original links are retained with their directional arrows deleted. In the Gaussian context it can be seen from (3) that moralization corresponds to the creation of non-zero entries in $V^{-1}$, which characterizes the CIG, from the non-zero entries in $\Phi_0$, which characterizes the DAG.

## 3 Identifying structural autoregressions

It will be usual that the order $p$ of a canonical autoregression will have been, at least tentatively, identified for the series. Structural autoregressive model identification then proceeds by construction of the CIG using the data matrix $X$ consisting of the collection of contemporaneous and lagged data vectors $(x_{p+1-k,i}, \ldots, x_{n-k,i})'$ for each series $i = 1, \ldots, m$ and each lag $k = 0, \ldots, p$. Assuming that the time series or data vectors have been mean corrected we use the covariance matrix estimate $\hat{V} = X'X/(n-p)$. A CIG is constructed from $\hat{V}$ in a similar manner as before, but with two differences:

1. the significance levels used are $z/\sqrt{(z^2+\nu)} \approx z/\sqrt{n-p}$, where $z$ is a critical value of the standard normal distribution.

2. we retain only those links which are significant and are either *between* contemporaneous variables or *attach to* contemporaneous variables from lagged variables.

These differences arise from the time series context, where the usual properties of regression estimation hold only in large samples and for regression on lagged values. See for example Anderson (1971, p 211). The main consequence is that we can assume only a large sample Normal distribution for the (so-called) $t$ values in the autoregression. Also, these properties do not hold for a time series regression equation which includes both future and past regressors, because the errors are not then in general uncorrelated. A sample partial autocorrelation between $x_{t-h,i}$ and $x_{t-k,j}$ for some $0 < h, k < p$ can correspond only to a $t$ value in the regression of one of these, say $x_{t-h,i}$ on all the other values, including at least one past and one future value. The $t$ value for $x_{t-k,j}$ and therefore the sample partial correlation between $x_{t-h,i}$ and $x_{t-k,j}$ will not have the required properties. In summary, the significance levels specified in 1 can only be applied to the links specified in 2. These are however the only links we consider for selection of a structural autoregression which, viewed as a DAG, only contains such links. For a stationary VAR(p) model, the subgraph of the CIG that consists of just the links specified in this way will be unchanged if the maximum lag used in its construction is greater than the true order $p$, but there may be some loss of efficiency in the statistical inference.

Efficient estimation of the selected model is done by separate regressions of each contemporaneous variable on those causal variables indicated in the DAG. The overall model is again assesed by a deviance $n' \sum \log \hat{\sigma}_r^2$, where $n' = n - p$ is here the length of data vectors used in the regression, and $\hat{\sigma}_r^2$ are the MLEs of the residual variances from these regressions (not the bias corrected mean square estimates).

Progress can only be made if the CIG is relatively sparse; no discrimination of structural models can be made if it saturated. Much depends on noting the absence of links which would be present if certain contemporanous directions were not avoided. One must be wary though of building up long chains of logic based upon the statistical evidence in the CIG. Likelihood based comparisons with a saturated model will indicate which models are plausible and may discriminate clearly between some competing models. Checks should be applied to confirm that the residuals are orthogonal innovations, and used as a possible guide to model improvement.

This methodology may be extended to VARMA models making use of a recursive estimation procedure.

The models obtained are useful for indicating how shocks pass through the set of contemporaneous values, besides identifying a more meaningful and parsimonious relationship between the series.

## References

Anderson, T.W. (1971). *The Statistical Analysis of Time Series.* New York: Wiley.

Box, G.E.P., and Jenkins, G.M. (1976). *Time Series Analysis, Forecasting and Control*. Oakland: Holden Day.

Box, G.E.P., and Tiao, G.C. (1977). A canonical analysis of multiple time series. *Biometrika*, **64**.

Greene, W.H. (1993). *Econometric Analysis*, Englewood Cliffs: Prentice-Hall.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lütkepohl, H., and Lee, T.C. (1985). *The Theory and Practice of Econometrics*. New York: Wiley.

Lauritzen, S.L., and Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *Journal of the Royal Statistical Society Series B*, **50**.

Spirtes, P., Glymour, C., and Scheines, R (1993). *Causation, Prediction and Search*. New York: Springer-Verlag.

Swanson, N.R., and Granger, C.W.J. (1997), Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, **92**.

Tiao, G.C., and Tsay, R.S. (1989). Model specification in multivariate time series. *Journal of the Royal Statistical Society Series B*, **51**.

Whittaker, J.C. (1990). *Graphical Models in Applied Multivariate Statistics*. Chichester: Wiley.

# A Recursive Solution for Continuous-Time Linear Estimation Problems

Ruiz-Molina, J.C., Navarro-Moreno, J. and Fernández, R.M.[1]

[1] Departamento de Estadística e I.O.,
Universidad de Jaén,
23071 Jaén, Spain.
{jcruiz,jnavarro,rmfernan}@ujaen.es,
Speaker: Fernández, R.M

**Abstract:** A recursive solution is provided to the problem of linear least-squares estimation of a continuous stochastic process that is based on approximate Karhunen-Loève expansions. The approach is applicable to the problem of estimating any linear operation of the signal process (such as mean-square derivatives or integrals) and it applies, without restrictions, to three frequently encountered problems: fixed interval smoothing, filtering and prediction.

**Keywords:** Approximate Karhunen-Loève Expansions, Linear Least-Squares Estimation Problems.

## 1 Introduction

The problem of linear least-squares estimation of a signal process based upon noisy observations has been a topic widely studied in the statistical literature due to its great practical and theoretical interest in a number of fields of knowledge: Engineering, Physics, Economy, etc. This problem is basically that of solving an integral equation, called Wiener-Hopf equation, which is a Fredholm equation of the second kind whose solution is the impulse-response of the optimum estimate. Usually two approaches have been applied to solve this integral equation:

*i)* By introducing structural assumptions on the covariance functions, such as stationarity, state-space structure (leading to the Kalman-Bucy filtering) or more general structures (see, e.g., Kailath et al. (1983)). The major drawback of the above methods of solution is that a number of physical phenomena do not satisfy these assumptions.

*ii)* By using methods of approximation for integral equations. This approach is closely connected with series representations for stochastic processes (see, e.g., Gardner and Frank (1971), Fortmann and Anderson (1973), Gardner (1974), Navarro-Moreno et al. (2000)). The most

general solution is given in Gardner (1974) where it is shown how any mean-square equivalent series representation can be employed to solve a very broad class of linear least-squares estimation problems. However, it is well-known that the Karhunen-Loève (KL) expansion is the optimal series representation under truncation in least mean-square sense and so, the solutions of Gardner and Frank (1971) for estimating a linear operation (in quadratic mean) of the signal process and that of Fortmann and Anderson (1973) for estimating the signal itself are those that optimally synthesize the available information. The main disadvantage of such solutions is that their implementation need to solve a homogeneous Fredholm integral equation, for which there is no general method of solution.

This paper addresses the solution of the problem of linear least-squares estimation from this second perspective. Specifically, we use approximate KL expansions (Gutiérrez et al. (1992) and Ruiz-Molina et al. (1999)) in the method of solution suggested by Gardner (1974) to provide an alternative solution. Approximate KL expansions are obtained by using the Rayleigh-Ritz (RR) method (Baker (1977)) to solve numerically the homogeneous Fredholm integral equation involved. The proposed solution takes the form of a suboptimum estimate that converges to the desired optimum estimate as the length of the series representation goes to infinity. Then, to obtain a recursive computational method useful in practical applications, such an estimate is expressed as a function of the solution to a stochastic differential equation which is similar to a Kalman-Bucy filter.

## 2 Linear Least-Squares Estimation

We consider the problem of estimating a signal process $\{x(t) : t \in [0,T]\}$, which is a real second-order stochastic process defined on the probability space $(\Omega, \mathcal{B}, P)$, with zero mean, and continuous in quadratic mean. We assume that the signal process cannot be observed directly, and that the observation process is of the type

$$y(s) = \int_0^s x(\tau)\, d\tau + w(s) \qquad 0 \le s \le t \le T$$

where $w(s)$ is a Wiener-Lévy process with parameter $r$ and uncorrelated with $x(s)$. Based on the observations $\{y(s) : s \in [0,t]\}$ we want to find the linear minimum variance estimator $\hat{z}(t)$ of a linear operation (in quadratic mean) of $x(t)$: $\{z(\tau) : \tau \in [0,T]\}$. According to the projection theorem of Hilbert spaces, this element $\hat{z}(t)$ exists, is unique and is the orthogonal projection of $x(t)$ onto $H(y,t)$ (the subspace of $L_2(\Omega)$ generated by the observation process $\{y(s) : s \in [0,t]\}$). Therefore, $\hat{z}(t)$, as an element of

$H(y,t)$, may be expressed in the following way:

$$\hat{z}(t) = \int_0^t h(t,s)\, dy(s)$$

where the impulse response $h(t,\cdot) \in L_2[0,t]$. It is well known that a necessary and sufficient condition on the impulse response function for the optimum estimate is that it satisfies the following integral equation

$$\int_0^t h(t,s) R_x(\tau,s)\, ds + rh(t,\tau) = R_{xz}(\tau,t) \tag{1}$$

$\forall t \in [0,T]$ and $\forall \tau \in [0,t]$, where $R_x$ is the autocorrelation function of $x$ and $R_{xz}$ is the cross correlation function for $x$ and $z$. The construction of the optimum estimate is thus complete. From the computational point of view, however, a more expedient solution procedure is necessary. For this purpose we follow the second approach pointed out in the preceding section.
In general, there is no known closed-form solution to (1) and therefore, we have to seek a sequence $\{h_N\}$ of closed-form approximate solutions that monotonically converge to $h$ as $N \to \infty$. Gardner (1974) obtains such approximations by truncating a mean-square equivalent infinite-series representation for $x(t)$. The best finite series representation of $x(t)$ is obviously a truncated KL expansion due to its optimality properties (giving rise to the solutions of Gardner and Frank (1971) and Fortmann and Anderson (1973)). Unfortunately, these approaches do not appear to be practical because the computation of eigenfunctions and eigenvalues is generally difficult. In the following section it is shown that the approximate KL expansions can be used to provide an alternative solution.

## 3 Suboptimum Estimate

Denote by $H(x)$ the subspace of $L_2(\Omega)$ spanned by the random variables of $\{x(t) : t \in [0,T]\}$. Let $\|\cdot\|_H$ be the norm defined in $H(x)$ and $\|\cdot\|_\infty$ the supremum norm on $[0,T] \times [0,T]$. The basic result to develop our solution is the following one.

**Theorem 3.1 (Ruiz-Molina et al. (1999))** *If $\frac{\partial^{2m}}{\partial t^m \partial s^m} R_x(t,s)$ exists and is continuous on $[0,T] \times [0,T]$, then*

1. $\left\| \frac{\partial^{2m}}{\partial t^m \partial s^m} R_x(t,s) - \sum_{i=1}^{N} \tilde{\lambda}_i \hat{\phi}_i^{(m)}(t) \hat{\phi}_i^{(m)}(s) \right\|_\infty \xrightarrow{N \uparrow \infty} 0$

   *where $\hat{\phi}_i(t) = \frac{1}{\lambda_i} \int_0^T R_x(t,s) \tilde{\phi}_i(s)\, ds$, $\{\tilde{\lambda}_i\}_i$ and $\{\tilde{\phi}_i(\cdot)\}_i$ are the RR approximate eigenvalues and eigenfunctions of $R_x$, respectively.*

2. $\left\| x^{(m)}(t) - \tilde{x}_N^{(m)}(t) \right\|_H \xrightarrow{N\uparrow\infty} 0$ *uniformly in* $t \in [0,T]$
*where* $\tilde{x}_N^{(m)}(t) = \sum_{i=1}^N \tilde{b}_i \hat{\phi}_i^{(m)}(t)$ *and* $\tilde{b}_i = \int_0^T \tilde{\phi}_i(t) x(t)\, dt$.

**Definition 3.1** *We define the suboptimum filter of* $z(t)$ *in the following way:*

$$\hat{z}_N(t) = \int_0^t \tilde{h}_N(t,s)\, dy(s) \tag{2}$$

*where* $\tilde{h}_N$ *is the solution of the integral equation:*

$$\int_0^t \tilde{h}_N(t,s) R_{\tilde{x}_N}(\tau,s)\, ds + r\tilde{h}_N(t,\tau) = R_{\tilde{x}_N z}(\tau,t) \qquad 0 \le \tau \le t \le T$$

*and* $R_{\tilde{x}_N}(t,\tau) = E\big[\tilde{x}_N(t)\tilde{x}_N(\tau)\big] = \sum_{i=1}^N \tilde{\lambda}_i \hat{\phi}_i(t)\hat{\phi}_i(\tau) = \hat{\mathbf{\Phi}}_N'(t)\tilde{\mathbf{\Lambda}}_N \hat{\mathbf{\Phi}}_N(\tau)$, *where* $\hat{\mathbf{\Phi}}_N(\cdot)$ *is the* $N$*-vector with ith entry* $\hat{\phi}_i(\cdot)$ *and* $\tilde{\mathbf{\Lambda}}_N$ *is a diagonal matrix with ith diagonal entry* $\tilde{\lambda}_i$.

**Theorem 3.2 (Navarro-Moreno et al. (2000))**

1. $\left\| \hat{z}(t) - \hat{z}_N(t) \right\|_H \xrightarrow{N\uparrow\infty} 0 \qquad \forall t \in [0,T]$
2. $P_N(t) = E\big[z(t) - \hat{z}_N(t)\big]^2 \xrightarrow{N\uparrow\infty} P(t) = E\big[z(t) - \hat{z}(t)\big]^2 \qquad \forall t \in [0,T]$
3. *If* $\{z(t) : t \in [0,T]\}$ *is a quadratic mean continuous stochastic process, then the convergences in 1 and 2 are uniform in* $t \in [0,T]$.

The central problem is to derive an efficient computational algorithm for the suboptimum estimate $\hat{z}_N(t)$ of $\hat{z}(t)$, $t \in [0,T]$. The following result gives a Kalman-Bucy-like formulation of the above estimate.

**Theorem 3.3 (Navarro-Moreno et al. (2000))** *The suboptimum filter* (2) *may be expressed as*

$$\hat{z}_N(t) = \hat{\mathbf{\Upsilon}}_N'(t)\hat{\boldsymbol{b}}_N(t)$$

*where* $\hat{\mathbf{\Upsilon}}_N(\cdot)$ *is the* $N$*-vector with ith entry*

$$\hat{v}_i(\cdot) = \frac{1}{\tilde{\lambda}_i} \int_0^T R_{xz}(s,\cdot)\tilde{\phi}_i(s)\, ds$$

*and the* $N$*-vector* $\hat{\boldsymbol{b}}_N(t)$ *obeys the differential equation*

$$\dot{\hat{\boldsymbol{b}}}_N(t) = r^{-1}\hat{\boldsymbol{B}}_N(t)\hat{\mathbf{\Phi}}_N(t)\left[y(t) - \hat{\mathbf{\Phi}}_N'(t)\hat{\boldsymbol{b}}_N(t)\right]$$

*with* $\hat{\boldsymbol{b}}_N(0) = \mathbf{0}$ *and* $\hat{\boldsymbol{B}}_N(t) = \left[\tilde{\mathbf{\Lambda}}_N^{-1} + r^{-1}\int_0^t \hat{\mathbf{\Phi}}_N(s)\hat{\mathbf{\Phi}}_N'(s)\, ds\right]^{-1}$.

A number of specific estimation problems may be obtained as special cases of the preceding theorem.

**Corollary 3.1**

1. *If* $z(t) = x(t)$, *then* $\hat{z}_N(t) = \hat{\mathbf{\Phi}}_N'(t)\hat{\boldsymbol{b}}_N(t)$.

2. *If* $z(t) = x^{(m)}(t)$, *then* $\hat{z}_N(t) = \hat{\mathbf{\Phi}}_N^{(m)'}(t)\hat{\boldsymbol{b}}_N(t)$.

3. *If* $z(t) = \int_0^t x(s)\,ds$, $\forall t \in [0,T]$, *then* $\hat{z}_N(t) = \hat{\mathbf{\Psi}}_N'(t)\hat{\boldsymbol{b}}_N(t)$, *where* $\hat{\mathbf{\Psi}}_N(t)$ *is the* $N$*-vector with ith entry* $\int_0^t \hat{\phi}_i(s)\,ds$.

4. *If a noncausal problem is considered, then the suboptimum estimate is* $\hat{z}_N(t|T) = \int_0^T \tilde{h}_N(t,\tau)\,dy(\tau)$ *where the approximate impulse response function is*

$$\tilde{h}_N(t,\tau) = r^{-1}\hat{\mathbf{\Upsilon}}_N'(t)\hat{\boldsymbol{B}}_N(T)\hat{\mathbf{\Phi}}_N(\tau) \qquad 0 \le t,\tau \le T$$

## 4 Example

The implementation of the proposed estimate is illustrated considering a specific example where the signal is the Wiener process and the processes to be estimated are the Wiener process itself and its mean-square integral. For this purpose, a trajectory of 100 data for the Wiener process is simulated and then it is corrupted by white noise with unit variance. The resultant trajectory becomes the noisy observations of the signal, from which we have to filter the desired processes. We use a continuous-discrete Kalman-Bucy filter to obtain the optimum estimations of the processes and they are compared to the estimations obtained from the suboptimum filter. Figures 1 and 2 show the optimum and suboptimum estimations for the Wiener process and its integral, respectively. The results show the great accuracy achieved with the proposed suboptimum filter.

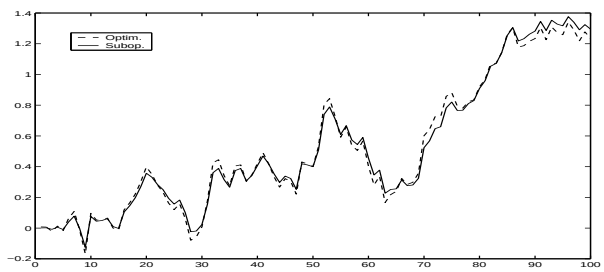



FIGURE 1. Optimum and Suboptimum Estimations for the Wiener Process.

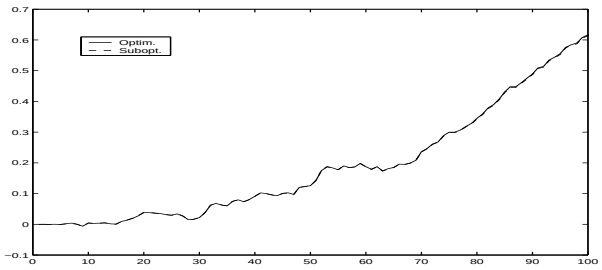



FIGURE 2. Optimum and Suboptimum Estimations for the Integral of the Wiener Process.

## References

Baker, C.T.H. (1977). *The Numerical Treatment of Integral Equations.* Oxford: Oxford University Press.

Fortmann, T.E. and Anderson, B.D.O. (1973). On the Approximation of Optimal Realizable Linear Filters Using a Karhunen-Loève Expansion. *IEEE, Trans. Information Theory*, **IT-19**.

Gardner, W.A. (1974). A Simple Solution to Smoothing, Filtering, and Prediction Problems Using Series Representations. *IEEE, Trans. Information Theory*, **IT-20**.

Gardner, W.A. and Franks, L.E. (1971). An Alternative Approach to Linear Least Squares Estimation of Continuous Random Processes. In: *5th Annu. Princeton Conf. Inform. Sciences and Syst.*, 267-275.

Gutiérrez, R., Ruiz-Molina, J.C. and Valderrama, M.J. (1992). On the Numerical Expansion of a Second Order Stochastic Process. *Appl. Stochastic Models Data Anal.*, **8**.

Kailath, T., Ljung, L. and Morf, M. (1983). Recursive Input-Output and State-Space Solutions for Continuous-Time Linear Estimation Problems. *IEEE, Trans. Automat. Contr.*, **AC-28**.

Navarro-Moreno, J., Ruiz-Molina, J.C. and Valderrama, M.J. (2000). A Solution to Linear Estimation Problems Using Approximate Karhunen-Loève Expansions. *IEEE, Trans. Inform. Theory*, to be published.

Ruiz-Molina, J.C., Navarro, J. and Valderrama, M.J (1999). Differentiation of the Modified Approximative Karhunen-Loève Expansion of a Stochastic Process. *Statist. Prob. Lett.*, **42**.

# A capture-recapture method that takes observed and unobserved heterogeneity into account

Elena Stanghellini[1] and Peter G.M. van der Heijden[2]

[1] Universitá di Perugia, Italy, stanghel@stat.unipg.it
[2] Utrecht University, the Netherlands, p.vanderheijden@fss.uu.nl

**Abstract:** In epidemiology the capture-recapture methodology is often used to estimate the size of an unknown population from a number of lists. One of the assumptions underlying this methodology is the homogeneity of capture probabilities of individuals. This assumption is often violated, and when it has been violated this leads to interactions between lists.
Two approaches exist for taking possible heterogeneity into account. The first is to incorporate observed heterogeneity into the model by letting, for example, capture probabilities be a function covariates. A more recent approach concentrates on unobserved heterogeneity alone, and models this by using latent variables. An example is the latent class model.
We extend the model by letting the parameters of the latent class model be functions of covariates. Unobserved heterogeneity is taken into account by the usage of the latent class model. Observed heterogeneity is taken into account in two ways. First, in each class of the latent class model the probabilities to fall in each of the lists are not homogeneous anymore because they differ over individuals as a function of their covariates. Second, membership of latent classes is also not homogeneous anymore but a function of covariates.
We illustrate the usefulness of this model with an example where we estimate the prevalence of diabetes in Italy.

**Keywords:** Capture-recapture, latent variables, latent class.

## 1 Introduction

As summarized by the International Working Group for Disease Monitoring and Forcasting (1995, further abbreviated as the IWG), the key assumptions underlying the most simple capture-recapture model for estimating the size of an unknown population from two lists are 1. the population is closed, 2. individuals can be matched if they are on more than one list, 3. for each list, each individual has the same chance of being included on the list, and 4. the two lists are independent. We will refer to the third assumption as the homogeneity assumption. Our paper will deal with assumptions 3 and 4, and assumes throughout that assumptions 1 and 2 are fullfilled.

Let $n_{11}$ be the number of individuals being on both lists, and $n_{10}$ and $n_{01}$ the number of individuals being only on list 1 or list 2 respectively. Under assumptions 1 to 4 it can be proven that $\hat{n}_{00} = n_{01}n_{10}/n_{11}$ gives the ML estimate of the number of individuals being on neither of the lists.
If there are $k$ lists, the independence assumption can be replaced by a much less demanding assumption. The observed frequencies can be coded into a $k$-way array having $2^k$ cells, with one entry zero by definition. This array can be analyzed by a hierarchical loglinear model where assumption 4 is replaced by assumption 4a: the k-factor interaction is zero (see for details and alternatives IWG, 1995).
It can be shown that violation of the homogeneity assumption will lead to dependence between lists. If $k = 2$, this is problematic because of assumption 4, and when $k \geq 3$ it is difficult to distinguish dependence of lists from this apparent dependence (heterogeneity) (see IWG 1995).
Two approaches exist to tackle the violation of the homogeneity assumption. The first approach uses only observed heterogeneity, and the second uses only unobserved heterogeneity. Bishop et al. (1975) suggest that, if one suspects heterogeneity, the individuals should be stratified into subgroups in which the homogeneity assumption is fulfilled. We refer to this as using a categorical covariate. Later Alho (1990) proposed a model that takes continuous covariates into account when the number of lists is two: for each list the probability to be in it is related to covariates by means of a logistic regression model.
The second approach is to assume a latent variable, and given the level of the latent variable the lists are independent. Thus the latent variable takes unobserved heterogeneity into account. One latent variable used is the Rasch model (see, for example, Darroch et al., 1993), and the other one is the latent class model (see, for example, Agresti, 1994, and for a recent extension Biggeri et al., in press).
In this paper we will present a model that takes both observed as well as unobserved heterogeneity into account. This model is an extension of the latent class model, that has close links to simultaneous latent class analysis as proposed by Clogg and Goodman (1994) and the concommittant variable latent class model as discussed by van der Heijden et al., (1996).

## 2 The latent class model

To keep the notation simple, we assume that there are four lists denoted by $Y_k, k = 1, \ldots, 4$, and we denote being on list $k$ by the index $j_k$, where $j_k = 1$ indicates being on list $k$ and $j_k = 0$ indicates not being on list $k$. We denote a latent variable $L$ with levels indexed by $z (z = 1, \ldots, Z)$. Let $\pi_{j_1 j_2 j_3 j_4}$ be the probability of status $j_1$ for list 1, to status $j_4$ for list 4. The

latent class model is defined as

$$\pi_{j_1 j_2 j_3 j_4} = \sum_{z}^{Z} \pi_z \pi_{j_1|z} \pi_{j_2|z} \pi_{j_3|z} \pi_{j_4|z} \tag{1}$$

where $\pi_z$ is the probability to be in class $z$, and $\pi_{j_k|z}$ is the probability to be having status $j_k$ in list $k$ given that one is in class $z$. Model (1) aims to take into account unobserved heterogeneity (assumed list dependence) as well as list dependence. Unobserved heterogeneity is taken into account by assuming that there are $Z$ groups of individuals, and in each group the individuals have homogeneous probabilities to fall on lists. In this instance $Z$ has to be 2 for the model to be identified. This model has been proposed in the area of capture-recapture models by Agresti (1994).

There is an extension by Biggeri et al. (in press) of (1) that is easily presented if we write (1) as a loglinear model with a latent variables. We do this by denoting the variables of the highest margins that define the hierarchical loglinear models, $[LY_1][LY_2][LY_3][LY_4]$, showing that there are direct relations between each of the lists and the latent variable, but the lists are independent given the latent variable. Biggeri et al. (in press) generalizes this model to take local dependencies between pairs of lists into account. An example of such a model is $[LY_1][LY_2][LY_3][LY_4][Y_1Y_3]$, i.e., this model assumes local dependence between lists 1 and 3.

Models like these can be estimated by the EM algorithm. Consider model $[LY_1][LY_2][LY_3][LY_4]$. In each E-step we find the conditional expectation of the probabilities $\pi_{z|j_1 j_2 j_3 j_4}$, given the current best parameter estimates. In the M-step we find new parameter estimates by fitting loglinear model $[LY_1][LY_2][LY_3][LY_4]$ to $Z \times (2^k - 1)$ probabilities, because the entries in the cells where $j_1, j_2, j_3, j_4 = 0000$, are zero by definition. After convergence, the $Z$ probabilities of not being in any of the lists are derived from the parameter estimates, and these provide us with an estimate of the number of individuals on neither of the lists.

### 2.1 The latent class model with categorical covariates

Assume now that there is one categorical covariate $C$. This allows to generalize the model $[LY_1][LY_2][LY_3][LY_4]$ to $[CLY_1][CLY_2][CLY_3][CLY_4]$ (for ordinary contingency tables models like these are also known as simultaneous latent class models, Clogg and Goodman, 1984). This model states that both the class sizes as well as the conditional probabilities may be different for different levels of $C$. Thus the model now accounts for observed heterogeneity. The model still assumes that the lists are independent given the latent variable and the covariate, but this assumption can be dropped by incorporating local dependencies between lists as discussed above (although identifiability of the model then has to be checked). Imposing restrictions to this model, and generalizations to more than one covariate, are straightforward.

### 2.2 The latent class model with continuous covariates

Assume now that the covariate is continuous. In this situation we go back to the model formulation given in (1). In this situation, that, for reasons of space, we do not discuss here in detail, we let the class size $\pi_z$ be related to the covariate by means of a multinomial logistic regression model. We can also do this for the conditional probabilities. For ordinary contingency tables models like these are also known as concommittant variable latent class models (van der Heijden, Dessens and Bockenholt, 1996).

## 3 Example

As an example we discuss the prevalence of diabetes in a town of Northern Italy (see Bruno et al. 1994; see also Fienberg et al., 1999). There are four sources, namely the diabetic clinic and/or family physicians data source ($S_1$), the hospital discharges data source ($S_2$), the insulin and oral hypoglycerin data source ($S_3$), and the reagent strips and insulin syringes data source ($S_4$). There is a categorical covariate $C$, namely treatment, having three levels, namely 1. Diet, 2. Hypoglycemic agents and 3. Insulin.
We will apply the methodology of Section 2.1. We will denote the latent variable by $C$. We start with a model with two latent classes. Thus we fit models to a table of 3 (levels of $C$), times 2 (levels of latent variable) times ($2^4$ - 1) cells. The sample size is 2,047.
We will denote the models fitted as loglinear models for the latent variable. The model we start with is $[CLS_1][CLS_2][CLS_3]$ $[CLS_4]$. In this model the unobserved heterogeneity is incorporated by the latent variable $L$: the probabilities to be in a source are different over the levels of the latent variable. The observed heterogeneity is incorporated by the fact that the probabilities to be in a source differ over the treatments. The four sources are independent conditional on the latent variable $L$ and the treatment $C$, so the model assumes that the covariate and the latent variable are able to explain the observed dependence between the sources. However, the deviance found is 40.2, for 15 df, so this model does not fit. In searching for better models we find that the model $[CLS_1][CLS_2][CLS_3][CLS_4]$ $[S_1S_3][S_1S_4]$ leads to a much better fitting model: by adding the two local dependencies the deviance becomes 19.5 for 13 df (an alternative is $[CLS_1][CLS_2][CLS_3][CLS_4]$ $[S_1S_2][S_1S_3]$, with a fit of 19.7, 13 df.). We can then omit one additional three factor interaction, namely either between $C, L$ and $S_2$, resulting in a deviance of 19.7 for 15 df (we will call this model 1), or between $C, L$ and $S_4$ resulting also in a fit of 19.7 and 15 df (we will call this model 2). Deleting both three factor interactions leads to a model with a significantly worse fit: deviance is 25.5, for 17 df.
Model 1 leads to an estimated total of diabetes patients of 2,755, and model 2 to 2,743. These estimates are similar to those reported by Fienberg et al. (1999). Both models are very similar, and therefore we only discuss model

1 in more detail. Notice that the loglinear model is Markov equivalent to the chain graph depicted in Figure 1 (Lauritzen, 1999, p.60). Also notice that this model (just like model 2) is hierarchical but not graphical, as the interaction between $C, L$ and $S_1$ is excluded (in model 2 the interaction between $C$, $L$ and $S_4$ is excluded). In Table 1 we find the estimated

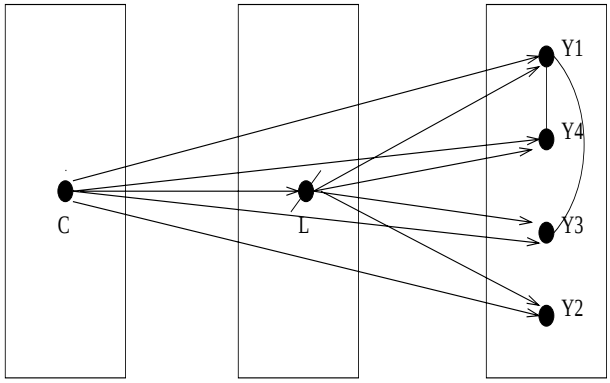



FIGURE 1. Graph of model 1

TABLE 1. Conditional probabilities for model 1.

| | | $S_1$ | | $S_2$ | | $S_3$ | | $S_4$ | |
|---|---|---|---|---|---|---|---|---|---|
| $C$ | $L$ | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0.31 | 0.69 | 0.31 | 0.69 | 0.78 | 0.22 | 0.15 | 0.85 |
| | 2 | 0.46 | 0.54 | 0.06 | 0.94 | 0.03 | 0.97 | 0.00 | 1.00 |
| 2 | 1 | 0.87 | 0.13 | 0.40 | 0.60 | 0.67 | 0.33 | 0.14 | 0.86 |
| | 2 | 0.60 | 0.40 | 0.09 | 0.91 | 0.40 | 0.60 | 0.02 | 0.98 |
| 3 | 1 | 0.87 | 0.13 | 0.62 | 0.38 | 0.77 | 0.23 | 0.55 | 0.45 |
| | 2 | 0.84 | 0.16 | 0.19 | 0.81 | 0.45 | 0.55 | 0.04 | 0.96 |

conditional probabilities to be in each of the sources. Overall, in latent class 1 the probabilities to be in a source is higher than in latent class 2 (for example, in treatment 2, in class 1 the probability to be in sources $S_1$ to $S_4$ is .87, .40, .67 and .14, but in latent class 2 .60, .09, .40 and .02.), with an exception for treatment 1 and source 1 (.31 and .46). The estimated latent class size 1 is, for treatment 1 to 3, .06, .11 and .70. This shows that in treatment 3 .70 of the individuals have higher probabilities to be in the sources, but for treatments 1 and 2 this is much lower. We notice that these (conditional) probabilities perfectly describe the fitted data under model $[CLS_1][CLS_2]$ $[CLS_3][CLS_4]$, but that model 1 has two additional parameters for the direct effects between sources 1 and 4 and 1

and 3. For sources 1 and 4 the direct effect is negative (loglinear parameter estimate is -1.1), and for 1 and 3 it is positive (1.2).
A few last remarks. The results in this section are preliminary. First, more models have to be studied. Second, the identifiability of the model discussed still has to be investigated and a confidence interval for the point estimate has to be constructed.

## References


Alho, J.M. (1990) Logistic regression in capture-recapture models. *Biometrics*, *46*, 623-635.

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics*, *50*, 494-500.

Biggeri, A., E. Stanghellini, F. Merletti, M. Marchi (in press). Latent class models for varying catchability and correlation among sources in capture-recapture estimation of the size of a human population. *Statistica Applicata*.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete multivariate analysis*. Cambridge: M.I.T.Press

Bruno, G., A. Biggeri, R.E. LaPorte, D. McCarty, F. Merletti and G. Pagano. (1994). Appplication of capture-recapture to count diabetes? *Diabetes Care*, *17*, 548-556.

Clogg, C.C. and Goodman, L.A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, *79*, 762-771.

Darroch, J.N., S.E. Fienberg, G.F.V. Glonek, B.W. Junker. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, *88*, 1137-1148.

Fienberg, S.E., M.S. Johnson and B.W. Junker (1999) Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. R. Statis. Soc. A, 162*, 383-405.

International Working Group for Disease Monitoring and Forecasting (1995). Capture-recapture and multiple-record systems estimation 1: history and theoretical development. *American Journal of Epidemiology*, *142*, 1047-1058.

Lauritzen, S.L . (1999). *Graphical models*. Oxford University Press.

van der Heijden, P.G.M., Dessens, J. and Bockenholt, U. (1996). Estimating the concommittant variable latent class model with the EM algorithm. *Journal of Educational and Behavioral Statistics*, *21*, 215-229.

# Variable selection in discriminant analysis: measuring the influence of individual cases

S.J. Steel[1] and N. Louw[1]

[1] Department of Statistics and Actuarial Science, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa, e-mail: sjst@fharga.sun.ac.za

**Abstract:** In this paper the influence of individual data cases in discriminant analysis, specifically in the case where an initial variable selection step precedes the analysis, is considered. Existing influence measures proposed by Fung (1995) in a non-selection context are extended to a selection context, and a new, more informative measure of selection influence is proposed. Application of these measures is illustrated by means of an example.

**Keywords:** classification, variable selection, influence

## 1 Introduction

Two of the important problems faced by a user of discriminant analysis are: (i) How can a subset of the available feature variables (i.e. variables which are used to classify an entity into one of the relevant groups) be selected to form a discriminant rule with desirable properties? (ii) How can the influence of an individual data case in the analysis be measured? In this paper we show how measures of influence which have been proposed in the literature in a non-selection context can be adapted for application in a selection context, and we propose a new more informative selection influence measure.

Identifying a subset of the available feature variables serves two purposes: firstly, the variables which best separate the groups are identified; secondly, the discriminant rule based on the selected variables frequently performs better in future classification than the rule based on all the feature variables. Many different variable selection techniques have been proposed in the literature (see McLachlan, 1992, and Le Roux, Steel and Louw, 1997, and the references therein). Few papers have been published dealing with measures which quantify the influence of individual data cases in a discriminant analysis. Contributions include Fung (1995) and the references therein. There are, however, no papers dealing with discriminant analysis influence measures in a selection context. For the multiple linear regression problem, Léger and Altman (1993) discuss influence measures taking variable selection into account. In §2 we use an approach similar to that

of Léger and Altman (1993) to adapt Fung's influence measures to the selection case.
The influence measures proposed by Fung (1995) are based on estimates of the expected squared change in the discriminant score if the $k$-th observation is omitted from the training data. These measures do not, however, give an indication of whether omitting an observation with a large value of this measure improves the fitted model. We therefore propose a new measure of influence which addresses this shortcoming. An example illustrating application of the influence measures is discussed in §3.

## 2 Selection influence measures

Consider a random vector $\mathbf{X}$, consisting of measurements on p feature variables. The entity corresponding to $\mathbf{X}$ belongs to one of two groups, $\Pi_1$ and $\Pi_2$. We assume that $E(\mathbf{X}) = \mu_i$ and $Cov(\mathbf{X}) = \Sigma$ in $\Pi_i$. Let $\mathcal{D} = \{(\mathbf{x}_j, y_j),\ j = 1, ...n\}$ be the training data, where $y_j$ is a group membership indicator variable. In practice, $\mu_1, \mu_2$ and $\Sigma$ are unknown, and are estimated from the available training data. The sample means $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ estimate $\mu_1$ and $\mu_2$ respectively, with the pooled covariance matrix, $\mathbf{S}$, acting as an estimate of $\Sigma$. Fisher's sample linear discriminant rule assigns $\mathbf{X}$ to $\Pi_1$ if

$$(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}^{-1}\mathbf{X} > \left[(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)' \mathbf{S}^{-1} (\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2)\right]/2, \tag{1}$$

and to $\Pi_2$ otherwise. Let $t_1(\mathbf{X}, \mathcal{D})$ be the estimated posterior probability of belonging to $\Pi_1$, and let $t_2(\mathbf{X}, \mathcal{D}) = 1 - t_1(\mathbf{X}, \mathcal{D})$. The estimated log-odds are

$$l(\mathbf{X}, \mathcal{D}) = \log\{t_1(\mathbf{X}, \mathcal{D})/t_2(\mathbf{X}, \mathcal{D})\}. \tag{2}$$

Fung (1995) proposes

$$\mathcal{F}(k) = E\left[l(\mathbf{X}, \mathcal{D}) - l\left(\mathbf{X}, \mathcal{D}_{(k)}\right)\right]^2 \tag{3}$$

as a non-selection measure of the influence of case $k$ in the analysis. Here $\mathcal{D}_{(k)}$ denotes the training data with case $k$ omitted. Fung (1995) shows that the expectation can be evaluated parametrically or non-parametrically (giving nearly identical results), and he states that a large value of $\mathcal{F}(k)$ implies that case $k$ is influential in the analysis.
Suppose now a variable selection technique is applied to $\mathcal{D}$, to identify a data-dependent subset $V(\mathcal{D})$ of the available variables, $\{1, ..., p\}$. The estimated log-odds, $l\left(\mathbf{X}^{V(\mathcal{D})}, \mathcal{D}\right)$, based on the selected variables are calculated. Suppose the *same* set of variables is used to calculate the estimated log-odds based on $\mathcal{D}_{(k)}$. Then

$$CSI\,(k) = E\left[l\left(\mathbf{X}^{V(\mathcal{D})}, \mathcal{D}\right) - l\left(\mathbf{X}^{V(\mathcal{D})}, \mathcal{D}_{(k)}\right)\right]^2 \tag{4}$$

is a *conditional* measure of selection influence for case $k$. We refer to (4) as a conditional measure since it is conditional on the set of selected variables, $V(\mathcal{D})$, being kept fixed. The unconditional analogue of (4) is obtained if the selection step is *repeated* after omission of the $k$-th data case. This gives

$$USI\,(k) = E\left[l\left(\mathbf{X}^{V(\mathcal{D})}, \mathcal{D}\right) - l\left(\mathbf{X}^{V\left(\mathcal{D}_{(k)}\right)}, \mathcal{D}_{(k)}\right)\right]^2. \tag{5}$$

As in the non-selection case, the measures (4) and (5) can be estimated parametrically and non-parametrically. In both cases the parametric and non-parametric estimates are nearly identical, and we denote the non-parametric versions by $csi(k)$ and $usi\,(k)$ respectively.
Consider a new entity $(\mathbf{X}, Y)$, and let $Z_i = 1$ if $Y = i$, and 0 otherwise. Write $t_i\,(\mathbf{X}, \mathcal{D}) = t_i\,(\mathbf{X})\,,\ i = 1, 2$. Let

$$R\,(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{2}\left[t_i\,(\mathbf{x}_j) - z_{ji}\right]^2, \tag{6}$$

where $z_{ji} = 1$ iff $y_j = i$, $j = 1, ..., n$; $i = 1, 2$. Then $R\,(\mathbf{t})$ is a measure of the error with which the group membership values of the training data cases are estimated by the posterior probability estimates. Based on this interpretation of $R\,(\mathbf{t})$, we propose the following new measure of selection influence. Let

$$R^{V(\mathcal{D})}(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{2}\left[t_i\left(\mathbf{x}_j^{V(\mathcal{D})}\right) - z_{ji}\right]^2 \tag{7}$$

be the measure when the full data set is used for variable selection, and let

$$R^{V\left(\mathcal{D}_{(k)}\right)}(\mathbf{t}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{2}\left[t_i\left(\mathbf{x}_j^{V\left(\mathcal{D}_{(k)}\right)}\right) - z_{ji}\right]^2 \tag{8}$$

be the measure when case $k$ is omitted from the training data. We propose

$$Q_k(\mathbf{t}) = \left\{R^{V\left(\mathcal{D}_{(k)}\right)}(\mathbf{t}) - R^{V(\mathcal{D})}(\mathbf{t})\right\}/R^{V(\mathcal{D})}(\mathbf{t}) \tag{9}$$

as a measure of the influence of case $k$ in the selection process. Note that $Q_k(\mathbf{t})$ quantifies the relative change in the accuracy of the post-selection posterior probability estimates if case $k$ is omitted from the training data. Therefore a negative value of $Q_k(\mathbf{t})$ would suggest that $\mathbf{t} = (t_1, t_2)$ based

on $V\left(\mathcal{D}_{(k)}\right)$ is a more accurate posterior probability estimator than $\mathbf{t}$ based on $V(\mathcal{D})$. One could also expect this to translate into more accurate prediction/ classification of future cases if the classification rule is based on $\mathbf{X}^{V\left(\mathcal{D}_{(k)}\right)}$ instead of $\mathbf{X}^{V(\mathcal{D})}$ . Considerations such as these suggest that we can modify, and hopefully improve, a given selection technique as follows. Suppose $Q_{(1)}(\mathbf{t}) \leq Q_{(2)}(\mathbf{t}) \leq ... \leq Q_{(n)}(\mathbf{t})$, and that $Q_{(1)}(\mathbf{t}) < 0$. Then we omit the case corresponding to $Q_{(1)}(\mathbf{t})$ and select the variables for further use from the reduced data set. A claim that such a modification improves a given selection technique should, however, be substantiated by means of simulation studies. This falls outside the scope of the current paper.
Why would one prefer $Q_k(\mathbf{t})$ to $usi\,(k)$ as a measure of selection influence? The reason for this is to be found in the fact that the sign of $Q_k(\mathbf{t})$ gives an indication of whether omitting case $k$ leads to an improvement or a deterioration in the fit of the selected model. Omitting cases for which $Q_k(\mathbf{t}) < 0$ leads to an improvement in fit, while $Q_k(\mathbf{t}) > 0$ signifies a deterioration in fit if case $k$ is omitted. A similar interpretation is impossible for $usi\,(k)$, since a large value of this measure merely implies that case $k$ is influential in the selection process, without providing an indication of whether it improves or worsens the fit of the selected model.

## 3 Data analysis

Consider the flea beetle data as discussed in Flury (1997, p.306). It consists of 39 observations on 4 variables pertaining to morphological properties of two species of flea beetles. The training data set, containing 19 observations on individuals of specie type 1 and 20 individuals of specie type 2, was analysed as follows. Firstly, formula (8) of Fung (1995) was used to calculate a non-parametric estimate of $\mathcal{F}(k)$. Table 1 shows some of the values.

| | | Selection | | | | |
|---|---|---|---|---|---|---|
| | No select | Full data set | Omitting case $k$ | | | |
| $k$ | $\mathcal{F}(k)$ | $csi\,(k)$ | $usi\,(k)$ | $Q_k\,(t)$ | Sel var nr | $e_k$ |
| 3 | .0696 | .0887 | .0887 | .1332 | 1,2,4 | .0291 |
| 9 | .5674 | .6973 | .6973 | -.1216 | 1,2,4 | .0270 |
| 16 | .0382 | .0287 | 1.1977 | -.0623 | 1,2,3,4 | .0316 |
| 18 | .0889 | .0191 | 1.5540 | .0283 | 1,2,3,4 | .0340 |
| 19 | .3571 | .4659 | .4659 | -.3124 | 1,2,4 | .0194 |
| 21 | .7464 | .5111 | 2.4411 | -.2216 | 1,2,3,4 | .0242 |
| 27 | 2.6897 | .4756 | 6.1553 | -.7867 | 1,2,3,4 | .0105 |
| 32 | .0483 | .0244 | 1.3816 | .0698 | 1,2,3,4 | .0348 |
| 36 | .8253 | .9916 | .9916 | -.6165 | 1,2,4 | .0151 |
| 38 | .2846 | .0979 | 2.0182 | .0044 | 1,2,3,4 | .0302 |

**TABLE 1: INFLUENCE MEASURES**

According to these values the five most influential points, in order of importance, are numbers 27, 36, 21, 9 and 19. Next, an all possible subsets approach was applied to the full data set. This resulted in variables 1, 2 and 4 being selected. Based on these selected variables, non-parametric estimates, $csi(k)$, $k = 1, ..., 39$, of (4), were calculated (see Table 1). The five most influential points are now 36, 9, 21, 27 and 19. It should be noted that the set of selected variables is kept fixed during calculation of $csi(k)$. Columns 4 and 5 of Table 1 provide values of $usi\,(k)$ and $Q_k(\mathbf{t})$ respectively. To calculate these values, variable selection is repeated for each reduced data set, and the selected variables are displayed in column 6 of Table 1. The five most influential points according to $usi\,(k)$ are 27, 21, 38, 18 and 32. Since variable selection is repeated for the reduced data set when $usi\,(k)$ is calculated, we feel that it is a more appropriate measure of influence in a selection context than either $\mathcal{F}\,(k)$ or $usi\,(k)$. The values of $usi\,(k)$ in column 4 of Table 1 do not, however, give an indication of whether omitting a given point improves the fit of the selected model. Such an indication is provided by the values of $Q_k(\mathbf{t})$ in column 5: a negative value implies an improved fit. Therefore, according to $Q_k(\mathbf{t})$, the five most influential data points are 27, 36, 19, 21 and 3, and omitting any one of the first four of these points leads to an improved fit. It should be noted that for $k = 16, 19$ and 3 the set of selected variables remains unchanged from the set selected using the full data set. A large negative (positive) value of $Q_k(\mathbf{t})$ can therefore be a reflection of a changed set of variables being selected, or merely of an improvement (deterioration) in fit, with the selected variables remaining unchanged.

What should one recommend to a practitioner analysing this data set? Based on the values in Table 1, we would recommend careful inspection of data points 27 and 36. In order to get an indication of whether omitting case 27 before analysis improves the classification performance of the resulting discriminant rule, the following steps were repeated 500 000 times. Firstly, consider the full data set. Randomly select 14 of the 19 cases in group 1 and 15 of the 20 cases in group 2 to form a training data set of 29 cases and a test data set of 10 cases. Apply variable selection to the training data and form a discriminant rule based on the selected variables. Use this rule to classify the cases in the test data set and calculate the misclassification rate. Secondly, omit case 27 from the data set and repeat the above process, now selecting only 14 of the remaining 19 cases in group 2 for the training data set. In both cases, we obtain measures of the classification performance by averaging the 500 000 respective misclassification rates. For the full data set, this average is 0.0295 and for the data set without case 27, it is 0.0105. Denote this error rate estimate by $e_{27}$. It seems that a practitioner would indeed be well advised to omit case 27. A critical reader may well ask what the value of the above error rate estimator would be if any other single case (instead of case 27) is omitted. To answer this, the error rate estimation procedure was repeated, each time omitting one

of the data cases. Column 7 of Table 1 shows a selection of the results. From these numbers, together with the estimates for cases not shown in Table 1, we see that omitting case 27 does indeed result in the lowest $e_k$. Case 27 is followed by cases 36 and 19, which are also ranked second and third respectively according to $Q_k(\mathbf{t})$. The strong correspondence between $Q_k(\mathbf{t})$ and $e_k$ is reflected in a correlation coefficient of 0.91 between the two sets of values.

## 4 Conclusions

Given the importance of variable selection in data analysis, identification of selection influential points is desirable. One approach is to extend existing influence measures, i.e. measures proposed in a non-selection context, for application to cases where variable selection is done. In §2 we showed how measures proposed by Fung (1995) can be extended in this way. We emphasised that a so-called unconditional approach, which explicitly takes a potential selection effect into account, is preferable to a so-called conditional approach, which is equivalent to using a model that was selected data dependently as if it were specified beforehand. Although these measures are useful, they do not indicate whether a point which has been identified as selection influential should be omitted from the data prior to analysis. Such an indication is provided by the new measure also proposed in §2. The analysis of an example in §3 shows that omitting a point with a large negative value of this measure is worthwhile in the sense of improving the classification performance of the resulting discriminant rule. Of course, the analysis of examples cannot prove that such a result is generally true, but we deem the evidence provided by the example (and others not reported on here) to be strongly supportive of the claim.

### References

Flury, B. (1997). *A First course in Multivariate Statistics*, Springer-Verlag New York.

Fung, W.K. (1995). Diagnostics in linear discriminant analysis. *Journal of the American Statistical Association*, **90**, 952-956.

Léger, C. and Altman, N. (1993). Assessing influence in variable selection problems. *Journal of the American Statistical Association,* ***88****, 547-556.*

Le Roux, N.J., Steel, S.J., and Louw, N. (1997). Variable selection and error rate estimation in discriminant analysis. *Journal of Statistical Computation and Simulation*, **59,** 195-219.

McLachlan, G.J. (1992). *Discriminant analysis and statistical pattern recognition,* Wiley New York.

# Using the bootstrap for bias estimation in population dynamics models

Verena Trenkel[1] and Dominique Pelletier[1]

[1] Institut Francais de Recherche pour l'Exploitation de la Mer, laboratoire MAERHA, Rue de l'Ile d'Yeu, BP 21105, 44311 Nantes cedex 3, France

**Abstract:** In the context of fisheries science the question has been raised whether point estimates needed to be bias corrected. It has been found that bias estimates obtained by bootstrapping can be rather large and it is not always clear how reliable this estimated bias is. The performance of the bootstrap for estimating bias in point estimates for model parameters from complex population dynamics models is investigated. In particular the robustness of the model based bootstrap to violations of model assumptions is tested. A nested bootstrap is employed for investigating the stability of bias estimates. An investigation primarily based on simulated datawill be presented.

**Keywords:** population dynamics models; bias; bootstrap.

## 1 Introduction

It has been recognised for some years that ignoring the uncertainity in the data used in fisheries stock assessment could lead to inappropriate or dangerous management decisions (Hilborn and Walters, 1992). Stochastic fluctuations of the environment and model uncertainties further increase the uncertainty fisheries scientists and managers have to deal with.
Most models used in fisheries for fish stock assessments in a wider sense are non-linear and fairly complex. In parallel to developments in statistical inference two schools of thoughts have developped about how observation errors should be incorporated into model-based inference and simulations of fisheries management scenarios. The Bayesian approach specifies conditional likelihood functions for the data given the parameters and makes use of additional information to formulate prior distributions for model parameters. In the case of the frequentist approach, the bootstrap method has become popular for estimating the variance and confidence intervals of parameter estimates (e.g. Pelletier and Gros, 1991).The question has been raised whether point estimates needed to be bias corrected and similarly for confidence intervals. It has been found for fisheries models that bias estimates obtained by bootstrapping can be rather large. It is not always clear how reliable these estimates are and what the causes are for the bias.

## 2 Bias estimation of point estimates

An estimator $g(\cdot)$ of a quantity of interest $\theta$ is called unbiased if $E[g(\cdot)] = \theta$, which means it is expected to give the true value. For some estimators this is true for any sample size $n$, for others $n$ needs to become very large ($n \to \infty$) for the assumption to hold. For small samples, the difference between the true value and the estimate is caused by sampling variability. An estimator is biased if $E[g(\cdot)] \neq \theta$ and not even collecting more information will lead to the true parameter estimate. In both cases of biased or unbiased estimators, the difference between the point estimate and the mean of the bootstrap estimates $\bar{\hat{\theta}}^b$ can be used for correcting the point estimate. The bias-corrected point estimate $\bar{\theta}$ is given by Efron and Tibshirani (1993) as

$$\bar{\theta} = \hat{\theta} - \hat{bias} = 2\hat{\theta} - \bar{\hat{\theta}}^b \tag{1}$$

As far as we know, the statistical properties of the complicated models usually employed for estimation in fisheries have not been studied. However, two inter-linked sources of bias have raised concern in the fisheries literature. The problems are known as error-in-variables and time series bias (Hilborn and Walters, 1992). In many fisheries problems, explanatory variables might be just as uncertain as the dependent variables. In addition to these sources of bias, violations of model error distribution assumptions can lead to biased point estimates.

Independent of the causes of bias in point estimates, two types of bootstrap methods are relevant for bias estimation : the nonparametric sample bootstrap and the model-based bootstrap (Efron and Tibsharani, 1993). Note that the two methods do not estimate the same quantity. Whereas the sample bootstrap creates the empirical distribution of the observations by resampling the data, in the case of the model-based bootstrap the distribution function of the observations is defined by the model and its moments are estimated from the data.

Given that the error distribution of the model is incorrect for whatever reason, the bias estimates obtained with the model based bootstrap will themselves be biased. In contrast, the nonparametric sample bootstrap should lead to consistent estimates of the estimator's bias. However, as calculations are carried out by Monte Carlo simulation, bias estimates can themselves be biased. A double bootstrap technique (nested bootstrap) can be employed for studying the bias in bias and variance estimates obtained from bootstrapping (Davison and Hinkley, 1997). Various other bootstrap techniques can be used for studying the properties of a given estimator.

## 3 Example : Fisheries stock assessment

The deterministic age structured population dynamics model employed in many fisheries stock assessement situations for a given species is

$$N_{t+1,a+1} = N_{t,a} e^{-(M_a + F_{t,a})} \qquad a = 1, \ldots A_{max}, t = 1, \ldots T \tag{2}$$

where $N_{t,a}$ is the number of individuals of age $a$ in year $t$ that belong to the same stock (i.e. are part of the same distinct population). The natural mortality rate at age, $M_a$, corresponds to mortality in the absence of fishing. Mortality induced by fishing is noted as $F_{t,a}$ and the resulting catch $C_{t,a}$ is calculated as

$$C_{t,a} = \frac{N_{t,a} F_{t,a} (1 - e^{-(M_a + F_{t,a})})}{M_a + F_{t,a}} \quad a = 1, \ldots A_{max}, t = 1, \ldots T \tag{3}$$

Given $M_a$, $F_{T,a}$ and $F_{t,A_{max}}$ as well as $C_{t,a}$, for $a = 1, \ldots A_{max}$ and $t = 1, \ldots T$, population numbers at age $N_{t,a}$ and all other fishing mortality rates can be calculated stepwise by solving equation (3) numerically and then inserting the result into equation (2).
External estimates of $M_a$ are required. Abundance indices $I_{t,a}$ are available for maximum likelihood estimation of fishing mortality rates $F_{T,a}$ and $F_{t,A_{max}}$. For normally distributed indices we write

$$I_{t,a} = q_{t,a} N_{t,a} + \varepsilon_{t,a} \quad a = 1, \ldots A_{max}, t = 1, \ldots T \tag{4}$$

where $\varepsilon_{t,a} \sim N(0, {\sigma_a}^2)$ and $q_{t,a}$ are known catchability coefficients. Similar formulas can be obtained for other distributions; the lognormal distribution is rather popular. Note that the observation errors for a given age class are assumed to come from the same distribution although other error models are possible. The error variance $\sigma_a^2$ is unknown and needs to be estimated. An iterative numerical procedure (simplex algorithm) gives maximum likelihood estimates for the unknown fishing mortality rates $F_{T,a}$ and $F_{t,A_{max}}$, $a = 1, \ldots A_{max}, t = 1, \ldots T$, as well as $\sigma_a$. This fish stock assessment approach is commonly called ADAPT (Gavaris, 1988).
A study on simulated data was carried out to investigate the application of the bootstrap for estimating the bias of parameter estimates in the above model. The population and catch data were simulated as for data set 4 in the study conducted by the NRC (1998) for investigating the performance of different stock assessment methods. Abundance indices from survey data were simulated by assuming a stratified sampling protocol and then raising these estimates to obtain the final abundance indices by age class and year. This procedure leads to normally distributed abundance indices.
For the nonparametric sample boostrap the survey strata were resampled with replacement. For both bootstrap methods different observation error distribution models were assumed for parameter estimation : normal (true) and lognormal errors. For the model based bootstrap the choice of standardisation of residuals was also investigated. The stability of all bias estimates was studied using a nested bootstrap.

## References

Davison, A.C. and Hinkley,D.V. (1997). *Bootstrap methods and their application.* Cambridge University Press.

Efron,B. and Tibshirani,R.J. (1993). *An introduction to the bootstrap.* Chapman and Hall.

Gavaris, S. (1988). *An adaptive framework for estimation of population size.* Can. Atl. Fish. Sci. Advisory Comm. (CAFSAC) Research document 88/29.

Hilborn,R. and Walters,C.J. (1992). *Quantitative fisheries stock assessment: Choice, dynamics and uncertainty.* London: Chapman and Hall.

National Research Council (1998). *Improving fish stock assessments*, Washington : National Academy Press.

Pelletier,D. and Gros,P. (1991). Assessing the impact of sampling error on mode-based management advice: comparaison of equilibrium yield per recruit variance estimators. *Canadian Journal of Fisheries and Aquatic Sciences*, **48**, 2129-2139.

# Wavelet Analysis of Seasonal Long Memory

Brandon Whitcher[1]

[1] EURANDOM, P.O. Box 513, 5600 MB Eindhoven, The Netherlands; *whitcher@eurandom.tue.nl*

**Abstract:** The analysis of time series with slowly decaying autocovariances, usually called long-memory processes, has been extensively studied over the past decade. Time series with slowly decaying periodic autocovariances have caught only limited attention recently. We investigate the ability of wavelet transforms, specifically the discrete wavelet packet transform, to analyze and adequately estimate parameters of interest in the case of seasonal long memory. We apply our methodology to atmospheric $CO_2$ measurements collected at the Mauna Loa observatory.

**Keywords:** Atmospheric $CO_2$; Discrete Wavelet Packet Transform; Maximum Likelihood.

## 1 Introduction: Seasonal Long Memory

A popular model for long-memory processes is given by the fractional difference process $\{X_t\}$ such that

$$(1-B)^d X_t = \epsilon_t, \tag{1}$$

where $-1/2 < d < 1/2$, $B$ is the backward shift operator and $\{\epsilon_t\}$ is Gaussian white noise with variance $\sigma_\epsilon^2$. This process is stationary and invertible (Hosking 1981). The spectral density function (SDF) of $\{X_t\}$ has a singularity at zero and is approximately linear on the log scale. The long-memory aspect is seen through the fact that the autocovariance sequence (ACVS) decays at a hyperbolic rate.
A simple generalization of this model was given, in passing, by Hosking (1981) via

$$\left(1 - 2\phi B + B^2\right)^\delta Y_t = \epsilon_t,$$

where $\phi = \cos(2\pi f_G)$ and $\{\epsilon_t\}$ is defined as before. The SDF of $\{Y_t\}$ is given by

$$S_Y(f) = \sigma_\epsilon^2 \{4[\cos(2\pi f) - \phi]^2\}^{-\delta}, \quad \text{for} \quad -\tfrac{1}{2} < f < \tfrac{1}{2}, \tag{2}$$

and hence, exhibits a singularity at $|f_G| < 1/2$ and exhibits an oscillating ACVS that slowly decays. Gray et al. (1989) showed that $\{Y_t\}$ is stationary and invertible for $|\phi| = 1$ and $-1/4 < \delta < 1/4$ or $|\phi| < 1$ and

$-1/2 < \delta < 1/2$. We call $\{Y_t\}$ a seasonally persistent process (SPP) which was originally coined by Anděl (1986). Clearly, the definition of an SPP also includes a fractional difference process. When $\phi = 1$ we have that $\{Y_t\}$ is a fractional difference process given by (1) with parameter $d = 2\delta$. A recent article by Arteche and Robinson (1999) contains relevant background information and an overview of semiparametric estimation procedures concerning seasonal long memory processes. Whitcher (2000) described how wavelet transforms may be used to simulate SPPs.

## 2 Discrete Wavelet Packet Transform

The orthonormal discrete wavelet transform (DWT) is known to approximately decorrelate long-memory processes. It does this through band-pass filtering the process in such a way that the spectrum in each pass band is approximately constant. In order to exploit the approximate decorrelation property for SPPs we need to generalize the partitioning scheme of the DWT. This is easily done by performing the discrete wavelet packet transform (DWPT) on the process; see, e.g., Percival and Walden (2000, Ch. 6). Instead of one particular filtering sequence, the DWPT executes all possible filtering combinations to obtain a wavelet packet tree, denoted by $\mathcal{T} = \{(j,n) : j = 0, \ldots, J; n = 0, \ldots, 2^j - 1\}$. An orthonormal basis $\mathcal{B} \subset \mathcal{T}$ is obtained when a collection of DWPT coefficients is chosen with disjoint ideal band-pass frequencies that cover $[0, 1/2]$.

Let $h_0, \ldots, h_{L-1}$ be the unit scale wavelet (high-pass) filter coefficients from a compactly supported wavelet family of even length $L$; e.g., Daubechies or minimum-bandwidth discrete time wavelets. The scaling (low-pass) coefficients may be computed via the quadrature mirror relationship $g_l = (-1)^{l+1} h_{L-l-1}$, for $l = 0, \ldots, L-1$. Now define

$$u_{n,l} \equiv \begin{cases} g_l, & \text{if } n \bmod 4 = 0 \text{ or } 3; \\ h_l, & \text{if } n \bmod 4 = 1 \text{ or } 2, \end{cases}$$

to be the appropriate filter at a given node of the wavelet packet tree.

Let $\mathbf{Y}$ be a length $N = 2^J$ vector of observations and $\mathbf{W}_{j,n}$, $(j,n) \in \mathcal{T}$, denote the vector of wavelet coefficients associated with the frequency interval $\lambda_{j,n} \equiv \left(\frac{n}{2^{j+1}}, \frac{n+1}{2^{j+1}}\right]$. Let $W_{j,n,t}$ denote the $t$th element of the length $N_j \equiv N/2^j$ vector of wavelet coefficients $\mathbf{W}_{j,n}$. Given the DWPT coefficients $\{W_{j-1,\lfloor \frac{n}{2} \rfloor,t}\}$ of length $N_{j-1}$ we produce $\{W_{j,n,t}\}$ via

$$W_{j,n,t} \equiv \sum_{l=0}^{L-1} u_{n,l} W_{j-1,\lfloor \frac{n}{2} \rfloor, 2t+1-l \bmod N_{j-1}}, \quad t = 0, 1, \ldots, N_j - 1.$$

Since the wavelet coefficients from the previous level $j-1$ are being utilized here, constructing the current series of wavelet coefficients only requires a convolution with $L$ nonzero terms. The vectors $\mathbf{W}_{j,n}$ are concatenated, in

increasing order by $\lambda_{j,n}$, to construct the DWPT vector $\mathbf{W}_{\mathcal{B}}$ corresponding to the orthonormal basis $\mathcal{B}$. To start the recursion set $\mathbf{W}_{0,0} = \mathbf{Y}$. The DWPT is most efficiently computed using a pyramid algorithm that has $O(N \log N)$ operations.

## 3 Approximate Maximum Likelihood Estimation

Let $\mathbf{Y}$ be a realization of a zero mean stationary SPP with unknown parameters $\delta$, $f_G$ and $\sigma_\epsilon^2 > 0$. The likelihood function for $\mathbf{Y}$, under the assumption of multivariate Gaussianity, is well known. To avoid the difficulties in computing the exact MLEs of the parameters of interest, we use the approximate decorrelation of the DWPT as applied to SPPs; i.e., $\Sigma_{\mathbf{Y}} \approx \widehat{\Sigma}_{\mathbf{Y}} \equiv \mathcal{W}_{\mathcal{B}}^T \mathcal{V}_N \mathcal{W}_{\mathcal{B}}$, where $\mathcal{W}_{\mathcal{B}}$ is an $N \times N$ orthonormal matrix defining the DWPT through the basis $\mathcal{B}$ and $\mathcal{V}_N$ is a diagonal matrix containing the coefficients $\nu_{j,n}^2 \equiv \sigma_\epsilon^2 \bar{\nu}_{j,n}^2$, where

$$\bar{\nu}_{j,n}^2 \equiv 2^{j+1} \int_{\frac{n}{2^{j+1}}}^{\frac{n+1}{2^{j+1}}} \frac{1}{\{4[\cos(2\pi f) - \cos(2\pi f_G)]^2\}^\delta} \, df$$

for all $(j,n) \in \mathcal{B}$ such that $\sum_{(j,n)\in\mathcal{B}} N_{j,n} = N$ (here $N_{j,n} = N_j$ for all $n = 0, \ldots, 2^j - 1$). The approximate log-likelihood function is given by

$$\begin{aligned} \widehat{\mathcal{L}}(\delta, f_G, \sigma_\epsilon^2 \,|\, \mathbf{Y}) &= \log\left(|\widehat{\Sigma}_{\mathbf{Y}}|\right) + \mathbf{Y}^T \widehat{\Sigma}_{\mathbf{Y}}^{-1} \mathbf{Y} \\ &= N \log(\sigma_\epsilon^2) + \sum_{(j,n)\in\mathcal{B}} \left[ N_{j,n} \log(\bar{\nu}_{j,n}^2) + \frac{\mathbf{W}_{j,n}^T \mathbf{W}_{j,n}}{\sigma_\epsilon^2 \bar{\nu}_{j,n}^2} \right] \end{aligned} \quad (3)$$

The MLE $\hat{\sigma}_\epsilon^2$ is easily obtained by differentiating (3), setting it to zero and solving for $\sigma_\epsilon^2$ which yields

$$\hat{\sigma}_\epsilon^2 \equiv \frac{1}{N} \sum_{(j,n)\in\mathcal{B}} \frac{\mathbf{W}_{j,n}^T \mathbf{W}_{j,n}}{\bar{\nu}_{j,n}^2}.$$

Replacing $\sigma_\epsilon^2$ with its MLE, we reduce the complexity of (3) to obtain the concentrated log-likelihood

$$\widehat{\mathcal{L}}(\delta, f_G \,|\, \mathbf{Y}) = N \log(\hat{\sigma}_\epsilon^2) + \sum_{(j,n)\in\mathcal{B}} N_{j,n} \log(\bar{\nu}_{j,n}^2). \quad (4)$$

The expression in (4) is now a function of only two parameters $\delta$ and $f_G$, through the numeric integration in $\bar{\nu}_{j,n}^2$, whose space of possible solutions lives on $(-1/2, 1/2) \times (0, 1/2)$. Minimization of (4), over the space of possible solutions, yields the MLEs $\hat{\delta}$ and $\hat{f}_G$.

## 4 Application to Atmospheric $CO_2$ Data

In this section we use the DWPT-based method to model 416 monthly atmospheric $CO_2$ measurements collected at the Mauna Loa Observatory, Hawaii. The entire record spans 1958-1998 and is available via the internet at `http://cdiac.esd.ornl.gov/ftp/ndp001/maunaloa.co2`, but missing observations reduce the continuous record to span June, 1964 to December, 1998. Woodward et al. (1998) analyzed the second difference of a similar time series using a two-factor Gegenbauer ARMA model; that is an SPP with two asymptotes at $f_1 \approx 0.083$ and $f_2 \approx 0.17$. The differencing operation was performed to remove the obvious trend, assumed to be quadratic, in the measurements. We prefer to remove the trend by filtering out the low frequency content of the signal through a wavelet multiresolution analysis (MRA); see Figure 1. The top six series are the wavelet details $\widetilde{\mathcal{D}}_1, \ldots, \widetilde{\mathcal{D}}_6$ with mean zero (plotted on the same vertical scale) and the bottom series is the wavelet smooth $\widetilde{\mathcal{S}}_6$ associated with frequencies $f \in [0, 1/32)$. This MRA is based on the frequency partitioning of the DWT and corresponds to an orthonormal basis consisting of $\{(j,1) : j = 1, \ldots, J\} \cup \{(J,0)\}$. Notice how the trend in the data is succinctly captured in the wavelet smooth. Subtracting the wavelet smooth from the original series produces a new de-trended series. When compared to the second difference series, band-pass filtering through MRA appears to preserve the seasonality in the data without inducing any unwanted effects.

The sample size for the monthly atmospheric $CO_2$ measurements is not dyadic, therefore we padded the time series to 512 observations and performed the DWPT. All wavelet coefficients using any of the 96 extra observations were removed from future calculations.

Basis selection begins by applying the Portmanteau test of white noise to the wavelet coefficient vectors $\mathbf{W}_{j,n}$ for $(j,n) \in \mathcal{T}$. Thus, a value $\mathcal{Q}_{j,n}$ is associated with each node in the wavelet packet tree. We want to include all vectors of wavelet coefficients such that $\mathcal{Q}_{j,n}$ does not exceed a predetermined quantile of the $\chi^2$ distribution. This produces an overcomplete basis in general, and must be orthogonalized by removing the children nodes if the parent is also included in the basis (a bottom-up procedure).

Once an appropriate orthonormal basis $\mathcal{B}$ has been selected, the wavelet variances $\bar{\nu}^2_{j,n}$ may be computed via numeric integration and optimization of the concentrated likelihood proceeds. Allowing for two asymptotes in the SDF of our process, the wavelet-based seasonal long memory model provides the following MLEs: $\hat{\delta}_1 = 0.49$, $\hat{f}_1 \approx 0.079$, $\hat{\delta}_2 = 0.41$, $\hat{f}_2 \approx 0.160$ and suggests the model

$$(1 - 1.758B + B^2)^{0.49}(1 - 1.068B + B^2)^{0.41}(Y_t - \widetilde{\mathcal{S}}_{6,t}) = \epsilon_t.$$

The first $(\delta, f_G)$-pair corresponds to the strong annual component in the data. This is apparent in Figure 1 where the majority of energy is con-

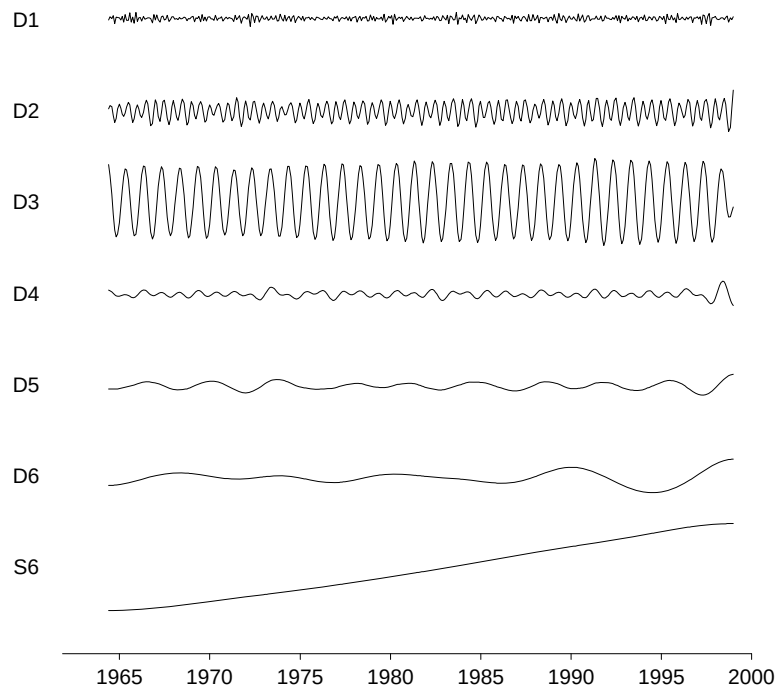



FIGURE 1. Multiresolution analysis of monthly $CO_2$ measurements.

tained in the third wavelet detail, corresponding to the frequency interval $\lambda_{3,1} = (1/16, 1/8]$. The second $(\delta, f_G)$-pair is associated with the first harmonic of the annual frequency and contributes less, as indicated by its smaller fractional difference parameter. These estimates agree with the ones obtained in Woodward et al. (1998).

## 5 Conclusions

As shown here, we believe that wavelet methodology provides a useful tool for all aspects of time series analysis – from exploratory data analysis to computationally intensive procedures such as maximum likelihood estimation. Not only do wavelet transforms simplify computations, they also facilitate a more straightforward interpretation of complex features by breaking the process down on a scale by scale basis. Further enhancements to wavelet-based estimation procedures for seasonal long memory models and

residual analysis, such as allowing short-memory autoregressive or moving average terms, are required in order to round out this methodology.

## References

Anděl, J. (1986). Long memory time series models. *Kybernetika*, **22**, 105–123.

Arteche, J., and P. M. Robinson (1999). Seasonal and cyclical long memory. In S. Ghosh (Ed.), *Asymptotics, Nonparametrics, and Time Series*, pp. 115–148. New York: Marcel Dekker.

Gray, H. L., N.-F. Zhang, and W. A. Woodward (1989). On generalized fractional processes. *Journal of Time Series Analysis*, **10**, 233–257.

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika*, **68**, 165–176.

Percival, D. B. and A. T. Walden (2000). *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.

Whitcher, B. (2000). Simulating Gaussian stationary processes with unbounded spectra. *Journal of Computational and Graphical Statistics*, to appear.

Woodward, W. A., Q. C. Cheng, and H. L. Gray (1998). A $k$-factor GARMA long-memory model. *Journal of Time Series Analysis*, **19**, 485–504.

# Generalized Mixture Autoregressive Model

Chun Shan Wong[1], Wai Keung Li[2]

[1] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: cswonga@hkusua.hku.hk
[2] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: hrntlwk@hku.hk

**Abstract:** We generalize the mixture autoregressive (MAR) model to the generalized mixture autoregressive with exogenous variables (GMARX) model for the modelling of nonlinear time series. The models consist of a mixture of two Gaussian transfer function type models with the mixing proportions changing over time. The advantages of the GMARX model over other nonlinear time series models include a fuller range of shape changing predictive distributions, the ability to handle cycles and conditional heteroscedasticity in the time series, and a better point prediction. The estimation is easily done via a simple EM algorithm and the model selection problem is addressed. The GMARX model is illustrated with the riverflow data of River Jökulsá Eystri of Iceland.

**Keywords:** EM algorithm; Forecasting; Mixture model; Model selection.

## 1 Introduction

In the past two decades, many nonlinear time series models have been proposed in the literature. See Tong (1990) for a comprehensive review. These models usually specify a nonlinear conditional mean and/or variance function. Despite the success of these models reported by many researchers, there is a major limitation of these models in application to some real life time series. This limitation comes from the fact that the conditional distributions of the time series are usually assumed to be Gaussian. Under the normality assumption, the marginal and/or conditional distributions of the time series are unimodal and symmetric. However, in real life many time series display features which seem to violate the normality assumption.
Several attempts have been made to incorporate the shape changing feature of conditional distributions into a nonlinear time series model. Martin (1992) introduced the multipredictor autoregressive time series models which tried to identify the shape of conditional distributions explicity. However, these models are quite difficult to estimate. Le *et al.* (1996) introduced the Gaussian mixture transition distribution models which allow a bit more flexibility in the shape of conditional distributions but these models cannot handle cyclical time series well because of the restrictions inherent in the models (Wong and Li, 2000).

Recently, Wong and Li (2000) generalize the idea of mixture distributions to the nonlinear time series context. The proposed mixture autoregressive (MAR) model is actually a mixture of $K$ Gaussian autoregressive models. As a mixture model, the conditional distributions of the time series given the past history are changing over time. These distributions can be multimodal. Despite the success in applications to some real data, several extensions of the MAR model are possible.

We generalize the MAR models to the generalized mixture autoregressive with exogenous variables (GMARX) models. The GMARX models extend the MAR models in two ways. First, they incorporate the information given by the exogenous variables. Second, unlike the MAR models, the mixing proportions can be changing over time depending on the past history of the variable of interest as well as the exogenous variables. While these new GMARX models retain the abilities in describing the shape of conditional distributions and capturing heteroscedasticity (Engle, 1982), the quality of the point forecasts generated is also much better than those generated from the MAR models. The selection of the GMARX model can be aided by a minimum Bayesian information criterion procedure and a hypothesis testing procedure. The GMARX model is illustrated with the riverflow data of River Jökulsá Eystri of Iceland.

## 2 The GMARX model

Suppose we are interested in the time series $\{y_t\}$ and we have also observed $l$ time series of exogenous variables, $\{x_{i,t}, i = 1, \ldots, l\}$. The GMARX model under consideration is defined by

$$\begin{aligned} &F\left(y_t|\mathcal{F}_{t-1}, \Omega_t\right) = \textstyle\sum_{k=1}^{2} \alpha_{k,t}\Phi(e_{k,t}/\sigma_k), \\ &e_{k,t} = y_t - \mu_{k,t} = y_t - \phi_{k0} - \textstyle\sum_{i=1}^{p_k} \phi_{ki} y_{t-i} - \sum_{i=1}^{l} \sum_{j=0}^{q_{ki}} \delta_{kij} x_{i,t-j}, \\ &\log(\alpha_{1,t}/\alpha_{2,t}) = \beta_0 + \beta_1 v_{1t} + \ldots + \beta_m v_{mt}. \end{aligned} \tag{1}$$

Here $F(y_t|\mathcal{F}_{t-1}, \Omega_t)$ is the conditional cumulative distribution function of $Y_t$ given the information in the sets $\mathcal{F}_{t-1}$ and $\Omega_t$, evaluated at $y_t$; $\mathcal{F}_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$; $\Omega_t = \{x_{1,t}, x_{1,t-1}, \ldots; \ldots; x_{l,t}, x_{l,t-1}, \ldots\}$; $\Phi(\cdot)$ is the (conditional) cumulative distribution function of the standard Gaussian distribution; $\alpha_{k,t}$ is the mixing proportion for the $k$th component, $\alpha_{1,t} + \alpha_{2,t} = 1$; $v_{it} \in \mathcal{F}_{t-1} \cup \Omega_t$ for $i = 1, \ldots, m$. We call the last equation in model (1) the logistic equation.

The GMARX model is actually a mixture of two Gaussian transfer function type models. The shape of the conditional distributions of $Y_t$ changes over time as the conditional means of the two components are dependent on past values of $x_{i,t}$ and $y_t$ as well as the mixing proportion, $\alpha_{k,t}$. The conditional distributions can be either unimodal or bimodal. The conditional expectation of $Y_t$ is given by $E(Y_t|\mathcal{F}_{t-1}, \Omega_t) = \alpha_{1,t}\mu_{1,t} + \alpha_{2,t}\mu_{2,t}$. This conditional

expectation may not be the best predictor of the future values. For example, if $\alpha_{1,t} = \alpha_{2,t} = 0.5$ and the difference between the means of the two components are large, $\mu_{1,t} \gg \mu_{2,t}$ say, this conditional expectation should not be a good forecast of $Y_t$ as it is located between the two modes of the conditional distribution and the probability mass around this conditional expectation would be quite small. Note that the merit of the mixture type time series model lies in its ability to describe the conditional distribution of the time series.

The GMARX model is capable of capturing changing conditional variances. The conditional variance of $Y_t$, is given by

$$\mathrm{var}(Y_t|\mathcal{F}_{t-1}, \Omega_t) = \sum_{k=1}^{2} \alpha_{k,t}\sigma_k^2 + \sum_{k=1}^{2} \alpha_{k,t}\mu_{k,t}^2 - \left(\sum_{k=1}^{2} \alpha_{k,t}\mu_{k,t}\right)^2.$$

This conditional variance of $Y_t$ is changing over time and will depend on the mixing proportions at time $t$ as well as the conditional means of the two components.

For the purpose of forecasting, suppose we are interested in evaluating the function $g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r})$ for $r \geq 1$. If $g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}) = E(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r})$, we are interested in the $r$-step point forecast. If $g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}) = F(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r})$, we are interested in the $r$-step ahead predictive distribution. Note that we assume here that the information of the exogenous variables is known up to time $t + r$. Granger and Teräsvirta (1993) give an account for $r$-step forecasts based on general nonlinear models. We discuss three approaches here, namely the naive, the exact and the Monte Carlo approaches.

In the naive approach, we have

$$g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}) = g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}, \hat{y}_{t+1}, \ldots, \hat{y}_{t+r-1})$$

where $\hat{y}_{t+1}, \ldots, \hat{y}_{t+r-1}$ are the naive forecasts of $y_{t+1}, \ldots, y_{t+r-1}$ given the information sets $\mathcal{F}_t$ and $\Omega_{t+r}$. In the exact approach, we have

$$g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}) = \int\cdots\int g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}, y_{t+1}, \ldots, y_{t+r-1}) \times f(y_{t+1}, \ldots, y_{t+r-1}|\mathcal{F}_t, \Omega_{t+r}) dy_{t+1} \ldots dy_{t+r-1}$$

where $f(y_{t+1}, \ldots, y_{t+r-1}|\mathcal{F}_t, \Omega_{t+r})$ is the joint density function of $y_{t+1}, \ldots, y_{t+r-1}$ given the information sets $\mathcal{F}_t$ and $\Omega_{t+r}$. In the Monte Carlo approach, we have

$$g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}) = \frac{1}{N}\sum_{j=1}^{N} g(Y_{t+r}|\mathcal{F}_t, \Omega_{t+r}, y_{t+1}^j, \ldots, y_{t+r-1}^j)$$

where $\{y_{t+1}^j, \ldots, y_{t+r-1}^j\}$ are sampled from $f(y_{t+1}, \ldots, y_{t+r-1}|\mathcal{F}_t, \Omega_{t+r})$. The naive approach is convenient but the Monte Carlo approach is preferable when a higher degree of accuracy is required.

## 3 Estimation and Model Selection

For the estimation of the parameters of a GMARX model, the EM algorithm (Dempster *et al.*, 1977) is employed. The standard errors of the parameter estimates can be computed by the Missing Information Principle (Louis, 1982). Extensive simulation studies indicate that the EM estimation method has small bias and reasonable standard errors, and the theoretical standard errors computed using Louis' method match the empirical standard errors well.

Model selection of the GMARX model can be done via a minimum Bayesian information criterion (BIC) procedure. Hypothesis testing concerning parameters in the logistic equation of the GMARX model can be done with the usual likelihood ratio test procedure. Simulation studies reveal that the minimum BIC procedure and the likelihood ratio test have good properties.

## 4 Example: River Jökulsá Eystri data

We fit the GMARX model to the riverflow data of River Jökulsá Eystri of Iceland. The data consist of the daily riverflow ($y_t$) in $\mathrm{m}^3\mathrm{s}^{-1}$, the daily precipitation ($x_{1,t}$) in mm, and the mean daily temperature ($x_{2,t}$) in °C at the meteorological station at Hveravellir from 1 January, 1972 to 31 December, 1974. There are 1,096 observations. An important hydrological feature of this river is that there is a glacier on the drainage area. Consequently, temperature has certain influence on the riverflow besides melting the snow. See Tong (1990) for a description on the data.

The orders $p_1, p_2, q_{11}, q_{12}, q_{21}, q_{22}$ and $m$ as well as the variables $v_{it}$ of the GMARX model were chosen by the minimum BIC procedure. The fitted GMARX model is

$$F(y_t|\mathcal{F}_{t-1}, \Omega_t) = \Phi(e_{1,t}/.7356) + \Phi(e_{2,t}/8.3592),$$

$$\begin{aligned} e_{1,t} = {} & y_t - 3.6866 - .9324y_{t-1} + .0705y_{t-2} - .0382x_{1,t} - .0663x_{2,t} \\ & +.0396x_{2,t-1}, \end{aligned}$$

$$\begin{aligned} e_{2,t} = {} & y_t - 3.7124 - 1.1984y_{t-1} + .4938y_{t-2} - .3317y_{t-3} + .1418y_{t-4} \\ & -.3894x_{1,t} + .2780x_{1,t-1} - .6642x_{2,t} - 1.1506x_{2,t-1} + .6707x_{2,t-2} \\ & +.6580x_{2,t-3}, \end{aligned}$$

$$\log(\alpha_{1,t}/\alpha_{2,t}) = 8.1542 - .3527y_{t-1} + .1466y_{t-3} - .1154x_{1,t} - .1987x_{2,t}.$$

We compare the GMARX model with the open-loop threshold autoregressive (TARSO) model (Tong, 1990), the nonlinear additive autoregressive with exogenous variables (NAARX) model (Chen and Tsay, 1993), the nested threshold autoregressive (NeTAR) model (Astatkie *et al.*, 1997) and the linear transfer function (TF) model. We compare these models based on the out-sample forecast performance and the ability in describing predictive distributions. We divided the data into two parts: the initialization

TABLE 1. Mean squared prediction errors of forecasts for the River Jökulsá Eystri data.

| Lead time | GMARX | TARSO | NAARX | NeTAR | TF |
|---|---|---|---|---|---|
| 1 | 59.90 | 62.38 | 61.64 | 48.79 | 67.68 |
| 2 | 141.77 | 148.60 | 140.82 | 115.72 | 166.36 |
| 3 | 179.74 | 191.93 | 176.90 | 164.58 | 217.65 |
| 4 | 208.30 | 225.86 | 199.83 | 208.50 | 257.12 |
| 5 | 226.47 | 250.33 | 215.28 | 242.24 | 277.87 |
| 6 | 248.78 | 271.91 | 230.84 | 273.75 | 290.04 |

TABLE 2. Empirical coverages of the $(1-\alpha)100\%$ one-step ahead prediction intervals for the River Jökulsá Eystri data.

| | $(1-\alpha)100\%$ | | | | | |
|---|---|---|---|---|---|---|
| Model | 95 | 90 | 80 | 70 | 60 | 50 |
| GMARX | 91.78 | 89.32 | 84.10 | 77.53 | 67.95 | 58.08 |
| TARSO | 92.88 | 90.41 | 86.03 | 78.36 | 70.96 | 63.56 |
| NAARX | 90.68 | 89.32 | 85.21 | 79.73 | 75.89 | 70.96 |
| NeTAR | 91.78 | 89.86 | 87.95 | 83.56 | 77.53 | 69.04 |
| TF | 91.51 | 90.68 | 86.58 | 81.37 | 78.36 | 72.88 |

(years 1972 to 1973) and the testing (year 1974) parts. We first fitted these models to the initialization part. We then generated one-step to six-step ahead forecasts and one-step and two-step prediction intervals, based on these fitted models, for the data in the testing part.

Table 1 show the mean squared prediction error of the one-step to six-step ahead forecasts for the riverflow generated by each model. From the table, the NeTAR model performs best when the lead time is small while the NAARX model performs best when the lead time is large. For small lead times, the performance of the GMARX model is comparable with that of the NAARX model. For larger lead times, the performance of the GMARX model is only slightly inferior to that of the NAARX model.

The empirical coverages of the one-step and two-step ahead prediction intervals for the riverflow generated by each model are shown in Tables 2 and 3. From the tables, the empirical coverages of the GMARX-based and the TARSO-based prediction intervals are closer to the nominal coverages than the prediction intervals generated by other models.

## References

Astatkie, T., Watts, D.G., and Watt, W. E. (1997). Nested threshold autoregressive (NeTAR) models. *International Journal of Forecasting*, **13**, 105-116.

TABLE 3. Empirical coverages of the $(1-\alpha)100\%$ two-step ahead prediction intervals for the River Jökulsá Eystri data.

| | $(1-\alpha)100\%$ | | | | | |
|---|---|---|---|---|---|---|
| Model | 95 | 90 | 80 | 70 | 60 | 50 |
| GMARX | 93.13 | 90.38 | 84.62 | 76.65 | 70.05 | 60.99 |
| TARSO | 90.93 | 89.01 | 83.79 | 73.90 | 66.48 | 56.04 |
| NAARX | 89.84 | 88.17 | 82.14 | 76.37 | 72.53 | 69.51 |
| NeTAR | 92.31 | 89.56 | 84.62 | 79.95 | 72.80 | 65.93 |
| TF | 91.21 | 88.46 | 84.89 | 79.95 | 75.00 | 66.48 |

Chen, R., and Tsay, R.S. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association*, **88**, 955-967.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1-38.

Engle, R.F. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987-1007.

Granger, C.W.J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. New York: Oxford University Press.

Le, N.D., Martin, R.D., and Raftery, A.E. (1996). Modeling flat stretches, bursts, and outliers in time series using mixture transition distribution models. *Journal of the American Statistical Association*, **91**, 1504-1514.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **44**, 226-233.

Martin, V.L. (1992). Threshold time series models as multimodal distribution jump processes. *Journal of Time Series Analysis*, **13**, 79-94.

Tong, H. (1990). *Non-Linear Time Series*. New York: Oxford University Press.

Wong, C.S., and Li, W.K. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society, Series B*, **62**, 95-115.

# Non- and Semiparametric Identification of Seasonal Nonlinear Autoregression Models

Lijian Yang[1] and Rolf Tschernig[2]

[1] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824,U. S. A.
[2] Institut für Statistik und Ökonometrie, Humboldt Universität zu Berlin, Spandauer Str.1, D-10178 Berlin, Germany

**Abstract:** Estimation and lag selection methods are proposed for non- and semiparametric seasonal nonlinear autoregression models, using either local constant or local linear estimation. For the semiparametric models, after preliminary estimation of the seasonal parameters, the nonparametric function estimation and lag selection are the same as for standard models. The semiparametric methods are applied to UK public investment data and indicate strong nonlinear dynamics.

**Keywords:** Final prediction error; Lag selection; Periodic nonlinear autoregression; Seasonal dummy nonlinear autoregression; Seasonal shift nonlinear autoregression.

## 1 Nonparametric Method

In nonlinear time series analysis, nonparametric estimators provide great flexibility since no parametric function class must be chosen a priori. This flexibility has not been available for seasonal time series because the necessary stationarity condition is violated when seasonality is present. A popular approach to remove seasonal nonstationarity is to use seasonally adjusted data. This, however, is not justified for nonlinear modelling for at least four reasons. First, the effect of such seasonal filters on data exhibiting nonlinearities is unclear as virtually all seasonal adjustment procedures have been designed for linear processes. Second, Ghysels, Granger and Siklos (1996) showed that some of these procedures such as X-11 involve nonlinear transformations which may change the properties of the original data. Third, data adjusted with most model-based seasonal adjustment procedures and procedures with model-based interpretation are noninvertible. Such procedures include those used by official agencies. See e.g. Maravall (1995) for details. Thus, an additional approximation error is introduced if finite order AR models, either parametric or nonparametric, are used. Finally, using seasonally adjusted data is misleading if an orthogonal decomposition of the original data into a trend, a seasonal and irregular component does not exist. See the examples in Franses (1996, Chapter 6),

who promoted periodic linear autoregressive models with autoregression parameters that vary with the seasons. Therefore, standard nonparametric models are not appropriate for seasonally adjusted data.
Yang and Tschernig (1998) considered a realization $\{Y_t\}_{t=0}^n$ of sample size $n+1$ of a process which has a stationary distribution for each of the $S$ seasons. We write index $t$ as $t = s + S\tau$ where $s = 0, 1, \ldots, S-1$ denotes the season and $\tau = 0, 1, \ldots$ represents a new time index.
The most general seasonal process which we consider is the seasonal nonlinear autoregressive (SNAR) model given by

$$Y_{s+\tau S} = f_s(X_{s+\tau S}) + \sigma_s^{1/2}(X_{s+\tau S})\xi_{s+\tau S} \tag{1}$$

where $X_t = (Y_{t-i_1}, Y_{t-i_2}, ..., Y_{t-i_m})^T$ is the vector of all the correct lagged values, $i_1 < \cdots < i_m$, $\xi_t$'s are i.i.d. with $E(\xi_t) = 0$, $E(\xi_t^2) = 1$, $t = s + \tau S = i_m, i_m+1, ....$ In contrast to the standard nonlinear autoregression model the regression functions $\{f_s\}_{s=0}^{S-1}$ here are allowed to vary with the $S$ seasons. This is a nonlinear generalization of the periodic AR (PAR) model

$$Y_{s+\tau S} = b_s + \sum_{i=1}^{p} \alpha_{is} Y_{s+\tau S-i} + \epsilon_{s+\tau S}.$$

For this reason, one can also view the SNAR model as a periodic nonlinear autoregression. Yang and Tschernig (1998) provided local constant and local linear estimators $\left\{\widehat{f}_s\right\}_{s=0}^{S-1}$ for functions $\{f_s\}_{s=0}^{S-1}$ in (1), with a data-driven bandwidth selected as in Yang and Tschernig (1999). They then defined the Final Prediction Error (FPE) as the following functional

$$FPE\left(\{\widehat{f}_s\}_{s=0}^{S-1}\right) = \frac{1}{S}\sum_{s=0}^{S-1} E\left[\left\{\breve{Y}_{s+\tau S} - \widehat{f}_s(\breve{X}_{s+\tau S})\right\}^2 w(\breve{X}_{s+\tau S,M})\right]$$

where $w$ denotes a weight function and $\{\breve{Y}_t\}$ is another series with exactly the same distribution as $\{Y_t\}$ but independent of $\{Y_t\}$. This was an extension of the final prediction error lag selection criteria of Tschernig and Yang (2000) to SNAR model. It was proposed that lags $\widehat{i}_1, \ldots, \widehat{i}_{\widehat{m}}$ be selected to minimize the FPE, and it was shown that this would lead to consistent selection of the correct set of lags $i_1 < \cdots < i_m$ as $n \to \infty$.

## 2 Semiparametric Methods

Since estimation and lag selection of model (1) are conducted within each season, the effective sample size is $n/S$, which may be too small for some macroeconomic applications. One may, however, restrict the seasonal variation of the functions between the $s$-th and the 0-th season to a constant

shift $b_s$: $f_s(\cdot) = f(\cdot) + b_s$, $s = 0, 1, 2, \ldots, S-1$. By definition $b_0 = 0$. The resulting process

$$Y_{s+\tau S} = f(X_{s+\tau S}) + b_s + \sigma_s^{1/2}(X_{s+\tau S})\xi_{s+\tau S} \tag{2}$$

is a seasonal dummy nonlinear autoregressive (SDNAR) model since it generalizes the seasonal dummy linear autoregressive (SDAR) model

$$Y_{s+\tau S} = b_s + \sum_{i=1}^{p} \alpha_i Y_{s+\tau S-i} + \epsilon_{s+\tau S}. \tag{3}$$

A set of estimators $\left\{\widehat{b}_s\right\}_{s=1}^{S-1}$ was proposed in Yang and Tschernig (1998) for the dummy parameters $\{b_s\}_{s=1}^{S-1}$ by averaging the differences $\widehat{f}_s(x) - \widehat{f}_0(x)$ where the functions $\left\{\widehat{f}_s\right\}_{s=0}^{S-1}$ are obtained for the general model (1). After removing $\left\{\widehat{b}_s\right\}_{s=1}^{S-1}$ from $Y_{s+\tau S}$'s for all $s$, the effective sample size for SDNAR model (2) is $n$ for function estimation.
Another way of restricting the seasonal nonlinear autoregression model (1) is to assume that the seasonal process is additively separable into a seasonal mean shift $\delta_s$, $s = 0, 1 \ldots, S-1$, and a nonseasonal nonlinear autoregression $\{U_t\}$, i.e. $Y_{s+\tau S} = \delta_s + U_{s+\tau S}$. One may call

$$\begin{aligned} Y_{s+\tau S} - \delta_s &= f(Y_{s+\tau S-i_1} - \delta_{\{s-i_1\}}, \ldots, Y_{s+\tau S-i_m} - \delta_{\{s-i_m\}}) \\ &\quad + \sigma^{1/2}(Y_{s+\tau S-i_1} - \delta_{\{s-i_1\}}, \ldots, Y_{s+\tau S-i_m} - \delta_{\{s-i_m\}})\xi_{s+\tau S} \end{aligned} \tag{4}$$

a seasonal shift nonlinear autoregressive (SHNAR) model, where for any integer $a$, we define $\{a\}$ as the unique integer between 0 and $S-1$ so that $a - \{a\}$ is a multiple of $S$. For identifiability, one assumes that $\delta_0 = 0$. This seasonal shift model is another way of generalizing the SDAR model (3) where the constants $\delta_1, ..., \delta_{S-1}$ of the linear model are obtained up to an additive constant via the system of linear equations $b_s = \delta_s - \sum_{i=1}^{p} \alpha_i \delta_{\{s-i\}}, s = 0, 1, ..., S-1$. Again, Yang and Tschernig (1998) proposed a method to estimate all parameters $\delta_1, ..., \delta_{S-1}$ and then remove them from the process in order to achieve an effective sample size of $n$. An alternative to the shift model (4) is analysed by Burman and Shumway (1998) who allow the seasonal shifts to be multiplied by a nonlinear function of time, however, at the cost of assuming the nonseasonal component to be linear.

## 3 An Example and Discussion

The three proposed models cover various kinds of deterministic seasonality. Nonstationarity due to stochastic seasonality has to be removed prior to

TABLE 1. Semiparametric lag selection of fourth differences of UK public investment

| Model | Criterion | Selected lags |
|---|---|---|
| dummy | FPE | 1,6,5 |
| shift | FPE | 1,6,5 |
| linear | FPE | 1,2,4,6 |
| linear | AIC | 1,2,4,6,8 |
| linear | SC | 1,2,4 |

Note: The maximal lag considered is 8 and all possible lag combinations are considered. The AIC stands for Akaike Information Criterion, SC the Schwarz Criterion.

the non- or semiparametric modelling (just like trends). In order to avoid overdifferencing and thus a noninvertible series one may use the HEGY-test (Hylleberg et al., 1990). Simulation study of Yang and Tschernig (1998) showed that the proposed seasonal lag selection methods work even in small samples.

We apply the semiparametric methods to the quarterly UK public investment in 1985 prices from 1962:1 to 1988:4 taken from Osborn (1990). A detailed analysis using linear periodic models and seasonal unit root testing by Franses (1996) can be found at *http://www.cs.few.eur.nl/few/people/franses/research/book1.htm.* The analysis is done with log data.

Franses (1996, p. 66-72) used the HEGY procedure and extensions and found for the UK data unit roots at all seasonal frequencies and the zero frequency. We therefore investigate the series after applying the filter $(1 - B^4)$, and then divide the series by its standard deviation.

The selection results for the fourth differences of the UK public investment data are presented in Table 1. After having applied the fourth order differencing filter, there is no relevant seasonality left in the data since removing seasonal shifts from the series does not change its variance. One may therefore expect the SDNAR and SHNAR model to perform similarly. Indeed, the chosen lag vector 1,6,5 is the same for both models. In contrast, all linear criteria contain the different vector 1,2,4.

Residual diagnostics showed no sign of misspecification. The surfaces of the estimated regression function of the SDNAR model are shown in Figure 1 where the value of lag 5 is fixed at $-0.10$, 0, 0.10. All surfaces look quite smooth. The deviation from linear hyperplanes is not quite pronounced but still evident. To reject linearity formally, nonparametric testing procedure is needed which is not yet available.

## References

Burman, P. and Shumway, R.H. (1998). Semiparametric modeling of seasonal time series, Journal of Time Series Analysis 19, 127-145.

Franses, H.F. (1996). Periodicity and Stochastic Trends in Economic Time Series, Oxford University Press, Oxford.

Ghysels, E., Granger, C.W. and Siklos, P. (1996). Is seasonal adjustment a linear or nonlinear data filtering process?, Journal of Business and Economics Statistics 14, 374-386.

Hylleberg, S., Engle, R.F., Granger, C.W.J., Yoo, B.S. (1990). Seasonal integration and cointegration, Journal of Econometrics 44, 215-38.

Maravall, A. (1995). Unobserved components in economic time series, Handbook of Applied Econometrics.

Tschernig, R. and Yang, L. (2000). Nonparametric lag selection for time series, Journal of Time Series Analysis, in press.

Yang, L. and Tschernig, R. (1998). Non- and semiparametric identification of seasonal nonlinear autoregression models, Discussion Paper 114, SFB 373, Humboldt Universität zu Berlin.

Yang, L. and Tschernig, R. (1999). Multivariate bandwidth selection for local linear regression, Journal of the Royal Statistical Society, Series B, 61, 793-815.

**Acknowledgements:** This research was supported in part by the Deutsche Forschungsgemeinschaft via Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" at Humboldt Universität zu Berlin. Yang's research has also been partially supported by NSF grant DMS 9971186. The authors thank Richard Blundell and Herman van Dijk for suggesting to investigate the seasonal shift model and Jörg Breitung and Helmut Lütkepohl for valuable comments on a later draft.

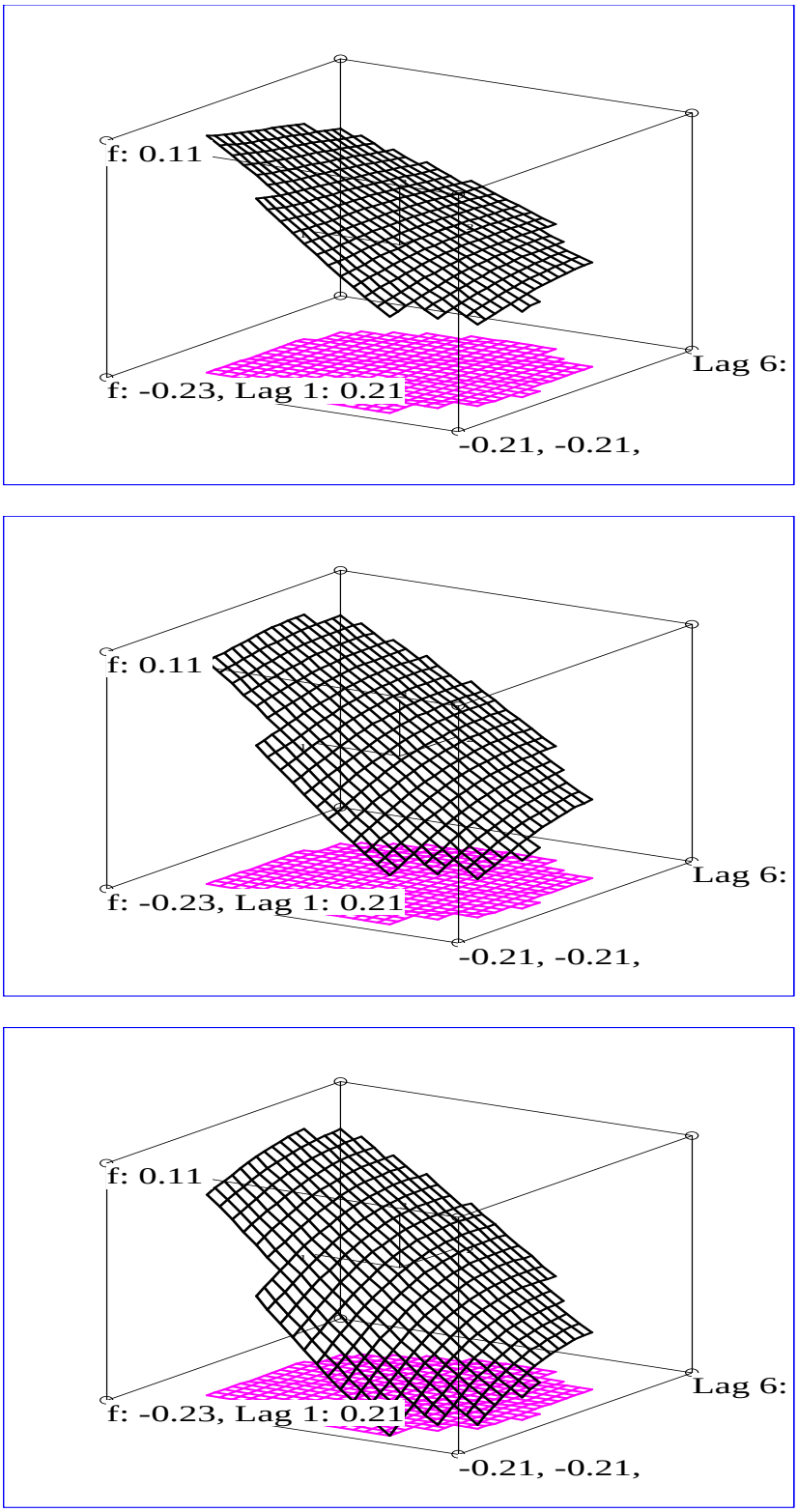



FIGURE 1. UK public investment, dummy model function of $Y_{t-1}$ and $Y_{t-6}$ with $Y_{t-5} = -0.1, 0, 0.1$

# Estimation of latent Gaussian ARMA models for categorical behaviour data

David Allcroft[1] and Chris Glasbey[2]

[1] Department of Mathematics and Statistics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland; d.allcroft@bioss.sari.ac.uk
[2] Biomathematics and Statistics Scotland, The King's Buildings, Edinburgh, EH9 3JZ, Scotland

**Abstract:** We consider the fitting of latent Gaussian models to categorical time series of cow feeding data. We derive a spectral quasi-likelihood for the data, and compare it with least squares fits to autocorrelations and MCMC estimators of the parameters in thresholded ARMA processes. We show that the spectral method is more efficient than least squares and far faster than MCMC.

**Keywords:** ARMA process; Cow feeding data; MCMC; Spectral likelihood; Tetrachoric correlation.

## 1 Introduction

The analysis of categorical behaviour data (Haccou and Meelis, 1994) poses many challenges for statisticians. In particular, we consider binary feeding data, collected for 30 days in May 1995 as part of a longer experiment at Langhill Dairy Cattle Research Centre, Edinburgh (Tolkamp et al., 1998). Thirty-six cows had continuous access to food in electronic feeders, and time spent at feeders was recorded automatically. Figure 1 shows three days of feeder-visit data for one cow, together with a derived variable, feeding data, obtained by suppressing short intervals away from feeders (Tolkamp et al., 1998). We prefer to model this variable as it is less susceptible than the feeder-visit data to herd and dominance effects. The data are recorded in continuous time, which we have discretised at one minute intervals.

We postulate a latent, Gaussian-distributed, physiological variable, with feeding occurring when this variable exceeds a threshold. Latent variables may be stochastically linked with categorical data, such as a logistic response (Keenan, 1982), or deterministically linked (Cox and Snell, 1992), as in our case. For categorical data, latent variables offer a more flexible approach than the use of either stochastic compartment models or hidden Markov models (MacDonald and Zucchini, 1997), because they simplify the inclusion of diurnal cycles, covariates and multivariate dependencies between animals.

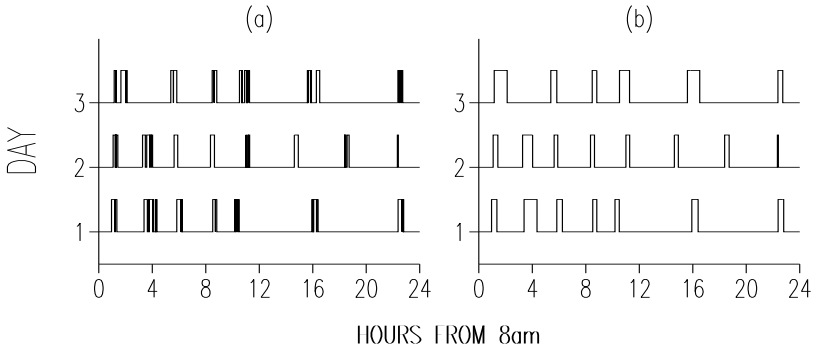



FIGURE 1. Three days of data for one cow: (a) feeder-visit data, (b) feeding data.

In Section 2, we relate the autocorrelations of the observed and latent processes, taking diurnal pattern into account. Then, in Section 3 we derive a spectral quasi-likelihood and estimate parameters in latent ARMA models using both it and a least squares fit to autocorrelations. In Section 4, we compare the efficiencies of these estimators. Finally, in Section 5 we discuss the results.

## 2 Autocorrelations

From data such as Figure 1, there is evidence for a diurnal feeding effect. We estimate the probability of feeding at a particular time of day by averaging observations at nearby times for all days. We used cross-validation to select an optimal window width of approximately one hour, by omitting each day's data in turn and then predicting it.

We estimate autocorrelations in the latent Gaussian process $\{y_t\}$, denoted $\hat{\rho}_l^{(G)}$ at lag $l$, from the observed binary process $\{x_t\}$, by numerically maximising quasi-likelihoods of the form

$$\sum_t \log \left[ \int_{L_t} \int_{L_{t+l}} f(y_t, y_{t+l}; \rho_l^{(G)})\ dy_t\ dy_{t+l} \right].$$

Here, the integration interval, $L_t$, is $(-\infty, T_t)$ if $x_t = 0$ and $(T_t, \infty)$ if $x_t = 1$, threshold $T_t$ is chosen so that the probability of not feeding, $\Phi_{T_t}$, matches the diurnal trend, where $\Phi_T$ is the standard normal integral from $-\infty$ to $T$, and $f$ denotes the bivariate Gaussian density with zero mean and unit variance.

In the absence of trend, $T$ is a constant and $\hat{\rho}_l^{(G)}$ simplifies to a tetrachoric correlation, estimable from the functional relationship

$$\hat{\rho}^{(B)} = (\Phi_{T,T}(\hat{\rho}^{(G)}) - \Phi_T^2)/(\Phi_T - \Phi_T^2), \tag{1}$$

where $\hat{\rho}^{(B)}$ denotes a sample autocorrelation of the binary process and $\Phi_{T,T}(\hat{\rho}^{(G)})$ is the bivariate normal integral from $-\infty$ to $T$ with correlation $\hat{\rho}^{(G)}$. Expression (1) can also be used to compute expected autocorrelations of the binary process, $\rho^{(B)}$, from those of the Gaussian process.

## 3 Model estimation

Inspection of $\hat{\rho}^{(G)}$ for the cows indicates that the simplest form of ARMA process that could provide an adequate model is an ARMA(2,1) model, with the additional benefit that this process has a continuous time analogue.
ARMA parameters can be estimated in an ad-hoc way via least squares, i.e. minimise $\sum_{l=1}^{n'/2}(\hat{\rho}_l - \rho_l)^2$, using either binary or Gaussian autocorrelations and some choice of $n'$. However, sample autocorrelation coefficients are highly correlated, so this is not necessarily an efficient estimation procedure. An alternative is to transform to independent statistics, for which the natural choice is by the Fourier transform. Whittle (1953) derived the spectral approximation for the log-likelihood, $\mathcal{L}$, of an $m$-dimensional stationary multivariate Gaussian process of length $n$, which has the form of a set of independent complex Wishart distributions (Brillinger, 1975):

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log|S_k| - \frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \mathrm{trace}[{S_k}^{-1}\hat{S}_k]. \tag{2}$$

Here $S_k$ and $\hat{S}_k$ are, respectively, the $m \times m$ complex matrices of cross-spectral and cross-periodogram coefficients at frequency $2\pi k/n$, so in our case,

$$S_k = \sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \rho_l^{(G)} e^{-2\pi i k l/n} \qquad \text{for } k = -\frac{n}{2}, \ldots, \frac{n}{2} - 1. \tag{3}$$

For a rigorous proof, see Coursol and Dacunha-Castelle (1983). Note that, in our application the Gaussian process is latent, so (2) can only be considered as a quasi-likelihood, and we also consider the same functional expression but with $\rho^{(G)}$ replaced by $\rho^{(B)}$.
We have extended the proof, to show that, for short-memory processes such as ARMA models, $\mathcal{L}$ can be approximated by a 'restricted' likelihood, $\mathcal{L}'$:

$$\mathcal{L}' = -\frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S'_k| - \frac{n}{2n'}\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \mathrm{trace}[{S'_k}^{-1}\,\hat{S'}_k],$$

where $n' \ll n$, with considerable computational saving. Here $S'_k$ and $\hat{S'}_k$ are obtained as the discrete Fourier transforms of cross-correlations up to lag $n'/2$ only, by replacing $n$ by $n'$ in (3).

The proof relies on re-expressing $\mathcal{L}$ as

$$\mathcal{L} = -\frac{1}{2}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log|S_k| - \frac{1}{2}\sum_{l=-\frac{n}{2}}^{\frac{n}{2}-1} \text{trace}[\alpha_l\ \hat{\rho}_l^{(G)}],$$

and similarly $\mathcal{L}'$, where $\alpha_l$ is the inverse autocorrelation coefficient at lag $l$, defined as

$$\alpha_l = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} {S_k}^{-1} e^{-2\pi i k l/n} \qquad \text{for } l = -\frac{n}{2}, \ldots, \frac{n}{2} - 1. \tag{4}$$

Here we use a multivariate generalisation of the univariate case considered by Cleveland (1972) and Chatfield (1979). We define $\alpha_l'$ similarly, but replacing $n$ by $n'$ and $S$ by $S'$ in (4).
The proof requires that $\rho_l^{(G)}$ and $\alpha_l$ are negligible for $|l| > n'/2$, and that $S_k$ is a continuous function of $k$. These conditions hold for ARMA processes, typically for small values of $n'$, because autocorrelations and inverse autocorrelations decay exponentially (Chatfield, 1979; Box and Jenkins, 1976). The proof then follows by showing that $\alpha_l' \approx \frac{n'}{n}\alpha_l$, $S_k' \approx S_{\frac{n}{n'}k}$ and

$$\sum_{k=-\frac{n'}{2}}^{\frac{n'}{2}-1} \log|S_k'| \approx \frac{n'}{n}\sum_{k=-\frac{n}{2}}^{\frac{n}{2}-1} \log|S_k|.$$

Figure 2 shows $\rho^{(G)}$ and $\hat{\rho}^{(G)}$ for one cow with ARMA parameter values of $\hat{\phi} = (1.9716, -0.9728), \hat{\theta} = -0.9927$, obtained by numerically maximising the Gaussian spectral likelihood. Similar values were obtained by least squares estimators for a range of values of $n'$ and with fits directly to the binary process.

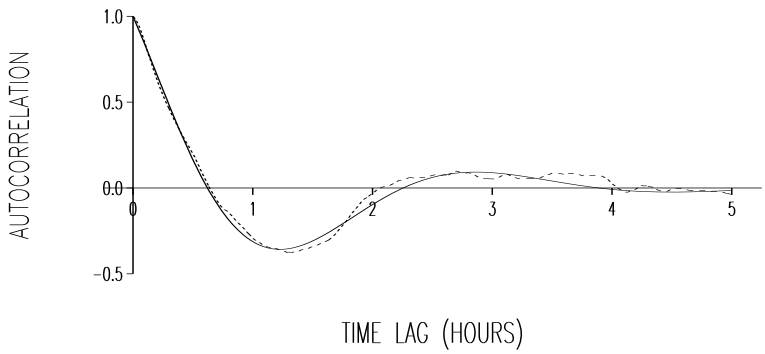



FIGURE 2. Gaussian autocorrelations for one cow: (—) $\rho^{(G)}$, (- - -) $\hat{\rho}^{(G)}$.

## 4 Efficiencies of estimators

We compare efficiencies of least squares and spectral estimators of ARMA parameters by simulation, for a range of models, of values of $T$, $n$ and $n'$ and use of either binary or Gaussian correlations. MCMC, combining Gibbs sampling with a Metropolis-Hastings algorithm, was also used for smaller values of $n$, to obtain a fully efficient estimator by maximum likelihood. Note that, in general, the MCMC method uses too much CPU time to be a practical alternative. A representative set of results is given in Table 1. For the AR(1) processes, all the methods are seen to be nearly as efficient as MCMC and so the spectral approach offers no clear benefit. For the other processes the spectral method is generally seen to be more efficient than the least squares, in some cases the improvement being quite substantial. The optimal value of $n'$ for least squares is quite crucial, typically 2 or 4, whereas for the the spectral method, the exact choice of $n'$ is not important, it simply has to be sufficiently large, typically greater than about 10.

| Model / Parameter values | $n$ | $T$ | LS (B) | LS (G) | Spectral (B) | Spectral (G) | MCMC |
|---|---|---|---|---|---|---|---|
| AR(1) | 100 | 0 | 145 | 146 | **142** | 145 | 140 |
| $\phi = 0.6$ | | 1 | 183 | 190 | **181** | 191 | 172 |
| | 1000 | 0 | **36** | 37 | **36** | 38 | 34 |
| | | 1 | **43** | 48 | **43** | 50 | 39 |
| MA(1) | 100 | 0 | 291 | 291 | 256 | **247** | 143 |
| $\theta = -0.6$ | | 1 | 327 | 339 | **321** | 333 | 202 |
| | 1000 | 0 | 148 | 148 | **95** | 104 | * |
| | | 1 | 135 | 155 | **110** | 155 | * |
| ARMA(1,1) | 100 | 0 | 487 | 489 | 451 | **311** | 275 |
| $(\phi, \theta) = (0.6, -0.3)$ | | 1 | 621 | 612 | 586 | **431** | 300 |
| | 1000 | 0 | 107 | 107 | 104 | **82** | * |
| | | 1 | 243 | 212 | 234 | **122** | * |
| CPU time (seconds) | | | 0.9 | 2.3 | 3.7 | 7.8 | 50000 |

TABLE 1: 1000 $\times$ RMSE of parameter estimates, averaged over 100 simulations, at optimal value of $n'$. The smallest value in each row, excluding MCMC, is highlighted. (MCMC was too slow to apply for MA(1) and ARMA(1,1) processes when $n = 1000$.) CPU times are for a SunUltra2 to process 100 series of length 100 for each estimation method.

## 5 Discussion

Latent Gaussian processes are flexible models for categorical behaviour data. In particular, we have seen that a latent ARMA(2,1) process shows

promise as a model for cow feeding data. We have explored alternative estimators, both analytically and by simulation, and found that the use of a spectral quasi-likelihood is both computationally quick and more efficient than least squares alternatives.
Further work will involve a more comprehensive simulation study for ARMA(2,1) processes. Also, we will explore ways of modelling a group of cows simultaneously in a multivariate framework, and the relationship between the latent Gaussian model and stochastic compartment and hidden Markov models.

**Acknowledgements:** We thank Ilias Kyriazakis, Bert Tolkamp and Langhill Dairy Cattle Research Centre, Edinburgh, for the data. We also acknowledge Colin Aitken and Elizabeth Austin for their input into this work and the Scottish Executive Rural Affairs Department for financial support.

## References

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis Forecasting and Control.* Holden-Day, San Francisco.

Brillinger, D. R. (1975). *Time Series: Data Analysis and Theory.* Holt, Rinehart and Winston, New York.

Chatfield, C. (1979). Inverse autocorrelations. *Journal of the Royal Statistical Society, Series A* **142**, 363-377.

Cleveland, W. S. (1972). The inverse autocorrelations of a time series and their applications. *Technometrics* **14**, 277-293.

Coursol, J. and Dacunha-Castelle, D. (1983). Remarks on the approximation of the likelihood function of a stationary Gaussian process. *Theory of Probability and its Applications* **27**, 162-167.

Cox, D.R. and Snell, E.J. (1992). *Analysis of Binary Data.* Second Edition. Chapman & Hall, London.

Haccou, P. and Meelis, E. (1994). *Statistical Analysis of Behavioural Data: An Approach Based on Time-structured Models.* Oxford University Press, Oxford.

Keenan, D.M. (1982). A time series analysis of binary data. *Journal of the American Statistical Association* **77**, 816-821.

MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman & Hall, London.

Tolkamp, B.J., Allcroft, D.J., Austin, E.J., Nielsen, B.L. and Kyriazakis, I. (1998). Satiety splits feeding behaviour into bouts. *Journal of Theoretical Biology* **194**, 235-250.

Whittle, P. (1953). The analysis of multiple stationary time series. *Journal of the Royal Statistical Society, Series B* **15**, 125-139.

# Tree-based Algorithms for Multiple Imputation of Missing Data

M.J. Bárcena[1] and F. Tusell[1]

[1] Facultad de CC.EE. y Empresariales. Avenida Lehendakari Aguirre, 83. 48015 BILBAO. e-mail: mb@alcib.bs.ehu.es

**Abstract:** We address the problem of completing a data matrix in which some observations are missing. Two algorithms, which use regression and/or classification trees, are proposed to impute the missing data. The algorithms have some desirable properties like making few assumptions and being well suited to multiple imputation.

**Keywords:** Missing data; Multiple imputation; Binary trees; File matching.

## 1 Introduction

Let $X$ be a $N \times (p+q)$ data matrix, with entries partly missing. This includes: i) Scattered missing entries which were never recorded, or were subsequently lost; ii) A full block missing, say the $N_2 \times q$ block consisting of the last $N_2$ rows and $q$ columns.
A problem of practical relevance is that of drawing inferences from such an incomplete data set, and a considerable body of literature exist on this issue. A landmark is the monograph Rubin (1987), setting up a methodology and advocating the use of multiple imputation. A recent monograph is Schafer (1997) which develops algorithms for imputation based on the EM algorithm and data augmentation.
The use of Schafer's algorithms implies the specification of a parametric model and a (possibly non-informative) prior on the parameters. Our intent has been to produce a good all-purpose nonparametric method, capable of coping with situations where little is known about the underlying data generation mechanism. We have explored methods which use binary trees (see Bárcena and Tusell (1999)) to perform the imputation, and developed two algorithms suitable for different situations.

## 2 Tree-based Algorithms

Consider the situation labelled ii) in the Introduction. Call $\mathbf{X_{obs}}$ the vector of fully observed variables for all $N_1 + N_2$ cases (the first $p$ columns) and $\mathbf{X_{mis}}$ the vector of the last $q$ variables incompletely observed. We would like to impute $\mathbf{X_{mis}}$ for the last $N_2$ subjects with a method:

1. Making as little assumptions as feasible on the joint distribution of the $(p+q)$ variables;

2. Allowing for multiple imputations, and

3. Taking into account the structure of the $q$ variables that are imputed. So the $q$ variables will be imputed jointly.

Binary trees are a flexible tool to capture the relationship between a response and a set of predictors. However, there is no widely available methodology to construct a tree with multivariate response. Ciampi(1991) requires the especification of a likelihood. Recently, Siciliano and Mola(2000) propose a splitting criteria to grow decision trees with multivariate response. The next two algorithms, built around univariate response binary trees, are designed to meet the three goals above.

### 2.1 The forest climbing algorithm

Let $\mathcal{Y}_{x|\mathbf{z}}$ be a binary tree "regressing" $x$ on the variables in $\mathbf{z}$. The *forest climbing* algorithm proceeds as follows:

1. Build trees $\mathcal{Y}_{X_{p+1}|\mathbf{X_{obs}}}, \ldots, \mathcal{Y}_{X_{p+q}|\mathbf{X_{obs}}}$ using the CART methodology (see Breiman et al. (1984)) and the $N_1$ complete observations.

2. Drop each of the $N_2$ incomplete cases down the $q$ trees constructed. Let case $i$ fall in the terminal nodes labelled $(\ell_{i,1}, \ldots, \ell_{i,q})$ of trees (respectively) $\mathcal{Y}_{X_{p+1}|\mathbf{X_{obs}}}, \ldots, \mathcal{Y}_{X_{p+q}|\mathbf{X_{obs}}}$. Call $(\ell_{i,1} \cap \ldots \cap \ell_{i,q})$ the subset of the $N_1$ complete cases which also end in said leaves. If $(\ell_{i,1} \cap \ldots \cap \ell_{i,q}) \neq \emptyset$, impute the missing values of case $i$ by those of one in $(\ell_{i,1} \cap \ldots \cap \ell_{i,q})$. If multiple imputation is desired, sample $k$ cases out of that intersection.

3. If $(\ell_{i,1} \cap \ldots \cap \ell_{i,q}) = \emptyset$, iteratively replace leaves by their ancestors ("climb the trees") until a non empty intersection is found from which one or more complete cases can be drawn. We climb first the tree where the replacement of a node by its ancestor leads to the least possible increase in deviance.

The idea is disarmingly simple. Take any tree $\mathcal{Y}_{X_k|\mathbf{X_{obs}}}$, $p < k \leq p+q$. The leaves of that tree are classes of a partition of the predictor space such that, within each class, knowledge of $\mathbf{X_{obs}}$ cannot help us in further refining our prediction of $X_k$ (otherwise, the leave would have been splited). It then makes sense that if subject $i$ with unknown $X_k$ ends in leave $\ell_{i,k}$ when dropped down the tree $\mathcal{Y}_{X_k|\mathbf{X_{obs}}}$, its $X_k$ value be predicted by a function of the $X_k$ values of subjects in the training sample which ended in the same leave.

Since we want to jointly impute all values in $\mathbf{X_{mis}}$ for the subject at hand, we would like to use complete cases in $(\ell_{i,1} \cap \ldots \cap \ell_{i,q})$, and this is exactly what the algorithm above does.
This algorithm can be considered as a nearest neighbour method in which "nearness" is defined as "falling in the same leaves than". Similar ideas exist in the literature, under the name of predictive mean matching.

### 2.2 The cascade algorithm

Jointly imputing $\mathbf{X_{mis}}$ given the values in $\mathbf{X_{obs}}$ is easy as soon as we have the conditional distribution $f(\mathbf{X_{obs}}|\mathbf{X_{mis}})$: we only have to draw from that distribution to perform multiple imputation. To generate observations with approximate distribution $f(\mathbf{X_{obs}}|\mathbf{X_{mis}})$, the *tree cascade* algorithm follows the steps:

1. Using the $N_1$ complete observations, construct a sequence of $q$ trees:

$$\{\mathcal{Y}_{X_{p+1}|\mathbf{X_{obs}}}, \mathcal{Y}_{X_{p+2}|\mathbf{X_{obs}},X_{p+1}}, \ldots, \mathcal{Y}_{X_{p+q}|\mathbf{X_{obs}},X_{p+1},\ldots,X_{p+q-1}}\}.$$

2. For each incomplete observation with observed $\mathbf{X_{obs}}$,
   - Drop $\mathbf{X_{obs}}$ down the first tree. Sample the leave where it ends to obtain a value $X_{p+1}$.
     A tree can be regarded as a mechanism generating observations with a given conditional distribution. So in this step we generate an approximate random drawing from $f(X_{p+1}|\mathbf{X_{obs}})$.
   - Since $f(X_{p+1}, X_{p+2}|\mathbf{X_{obs}}) = f(X_{p+1}|\mathbf{X_{obs}})f(X_{p+2}|\mathbf{X_{obs}}, X_{p+1})$, to generate an approximate random drawing from $f(X_{p+1}, X_{p+2}|\mathbf{X_{obs}})$:
     Drop $(\mathbf{X_{obs}}, X_{p+1})$ down the second tree in the sequence. Sample the leave where it ends to obtain a vector of imputed values $(X_{p+1}, X_{p+2})$.
   - Do likewise for the rest of the trees in the sequence. The last tree produces an approximate random drawing from $f(\mathbf{X_{mis}}|\mathbf{X_{obs}})$.

There are several ways to determinate the order in which each missing variable is used as a response variable in each tree of the sequence. We have investigated two different alternatives: *best first* and *best last*. In the first case, the trees are used in order of decreasing goodness of fit; the rationale being that, since each imputed variable can be input in subsequent trees, we want the values imputed earlier to be of the best possible quality.
On the other hand, in order to ensure consistency of the imputed variables, the whole vector $X_{p+1}, \ldots, X_{p+q}$ is imputed at the last step, which makes desirable a high quality tree at the end of the cascade.

## 3 Implementation and simulated results

We have written functions to implement our methods in the statistical language R (see Venables at al.(1997) for a description). We have used the functions in the `rpart` package (see Therneau and Atkinson(1997)) as building blocks. For the purpose of comparison, we used the routines in the package `norm`, a port to R (by Alvaro Novo, and available at CRAN, `http://cran.ar.r-project.org`) of the programs of the same name described in Schafer(1997).

We have generated data from a multivariate normal distribution $N_{15}(\mathbf{0}, \Sigma)$ with $\Sigma$ exhibiting moderate correlation among variables. The variables were standardized to have variance equal to one. Each of the two hundred replications generated contains $N = 500$ observations. The last $N_2 = 50$ observations of the last $q = 5$ variables were deleted and then their values imputed using the remaining $N_1 = 450$ complete observations as the training sample.

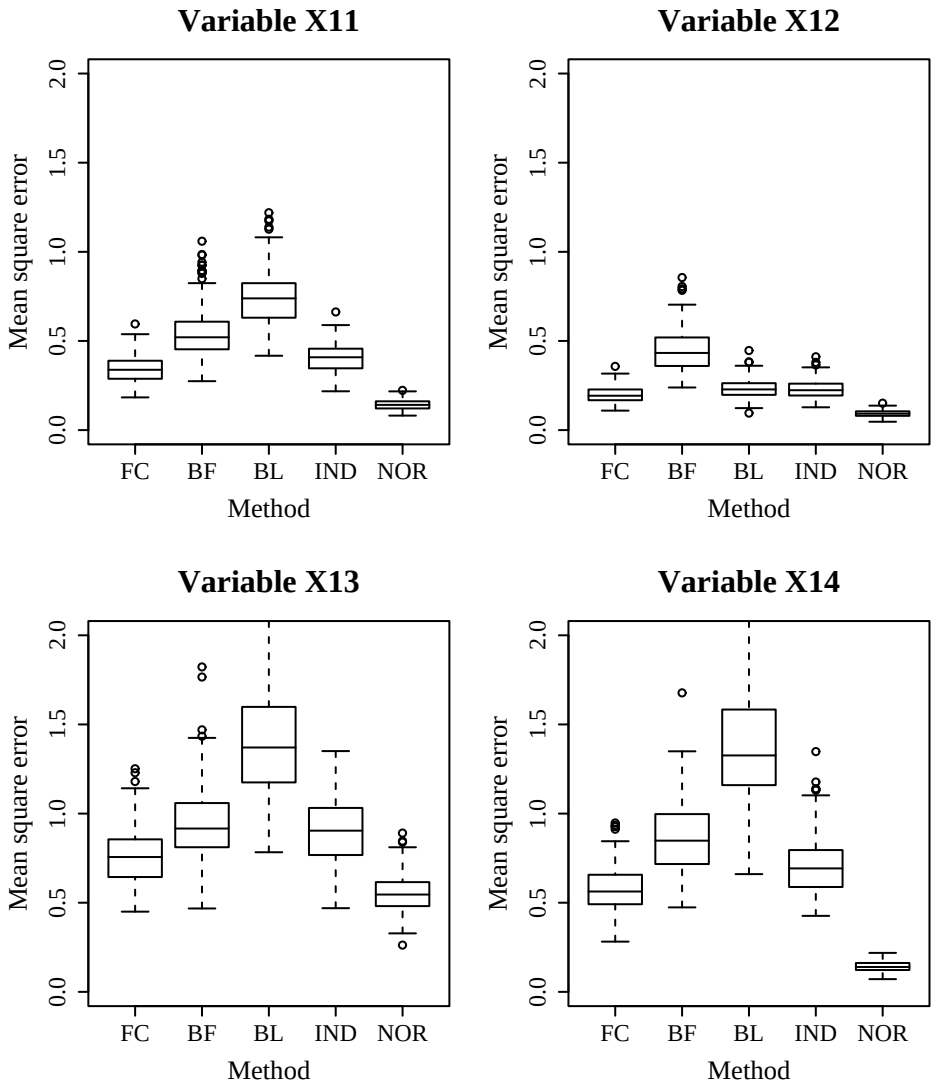
We have simulated the behaviour of the forest climbing algorithm (FC) and the cascade algorithm, both with best first (BF) and best last (BL) orderings and joint imputation (that is, all of the missing values are imputed at once, thus ensuring compatibility of the imputed values). We have also simulated the behaviour of the cascade algorithm with BF order and individual imputation of each variable (IND). Finally, we have simulated Schafer's method (NOR), using the EM algorithm to find the maximum likelihood estimates of the parameters conditional on $\mathbf{X_{obs}}$ and subsequently drawing random observations from that conditional distribution $f(\mathbf{X_{mis}}|\mathbf{X_{obs}})$.

Figure 1 shows the mean square error (MSE) of imputation for four of the variables, averaged over the 200 replications of the experiment (the fifth variable, not shown for lack of space, behaves similarly). Notice that a naïve strategy of imputing with a random complete subject from the sample *(cold deck)* would give a MSE of 2. Naïve imputation using the mean would give a MSE of 1. Since data is generated following a multivariate normal model, we can expect the parametric method (NOR) to perform best, and this is indeed the case. What is interesting is that the forest climbing algorithm is always a second best. When imputing using the cascade algorithm, the minimum MSE is of course obtained imputing each variable separately. Of the remaining two cascade algorithms, neither order BF or BL seems uniformly better (see for example the results for variable X11 and X12 in Figure 1). Additional more extensive results are available from the authors.

## 4 Some remarks

Both the forest climbing and cascade algorithms meet the goals enumerated in Section 2: they are all-around methods making almost no assumptions, take into account the structure of the variables to impute and provide for easy multiple imputation.

FIGURE 1. Imputation results for four variables and two hundred replications with $N_1 = 450$, $N_2 = 50$, $p = 10$, $q = 5$ and multivariate normally distributed data. See text for description of the methods.




Both methods scale well, and can be used with fairly large samples. The largest portion of time is devoted to constructing the trees. Subsequent imputation is very fast. Typically, only a fraction of cases require climbing in the forest climbing algorithm.

We remark in closing that generalizations are possible to the case of irregularly missing observations, situation labelled i) in the Introduction. The generalized algorithms use the completely observed cases to construct a tree for each incomplete variable as a function of all variables but itself. The set of trees obtained is used as indicated for the situation ii).
When dropping a partially observed case down one of the trees, we might eventually need the value of a variable which is not available for that case. Full use is made then of surrogate splitting. This achieves the same flexibility that the sweep algorithm provides in parametric methods, where regression on a different subset of regressors is required for each missingness pattern. Here, a single tree is constructed for each variable, with missing "regressors" taken up by surrogates if need be.

**Acknowledgements:** We thank for support the Spanish MEC (grants PB95-0346 and PB98-0149).

## References

Bárcena, M.J. and Tusell, F. (1999). Enlace de encuestas: una propuesta metodológica y aplicación a la Encuesta de Presupuestos de Tiempo. *Questiió*, **23**, 297-320.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, California.

Ciampi, A. (1991). Generalized regression trees.. *Computational Statistics and Data Analysis*, **12(1)**, 57-78.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* John Wiley and Sons, New York.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, London.

Siciliano, R. and Mola, F. (2000). Multivariate data analysis and modeling through classification and regression trees. *Computational Statistics and Data Analysis*, **32**, 285-301.

Therneau, T.M. and Atkinson, E.J. (1997). *An introduction to recursive partitioning using the RPART routines.*, Technical Report, Mayo Foundation.

Venables, B., Smith, D., Gentleman, R. and Ihaka, R. (1997). *Notes on R: A Programming Environment for Data Analysis and Graphics.*, Auckland: Dept. of Statistics, University of Adelaide and University of Auckland. Available at `http://cran.at.r-project.org`.

# Goodness-of-fit Tests for Bivariate Ordinal Regression Models

Rami Bustami[1], Emmanuel Lesaffre[1], Geert Molenberghs[2]

[1] Biostatistical Centre, Catholic University of Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium
[2] Biostatistics, Limburg University Centre, University Campus-Building D, B-3590 Diepenbeek, Belgium

**Abstract:** Two different methods are proposed to test the goodness of-fit of the BDM (Bivariate Dale Model, Dale, 1986) which is an example of a bivariate generalized linear model. The first is based on the technique of le Cessie and van Houwelingen (1991). This test is based on a weighted sum of kernel smooth estimates for both marginal and association standardized residuals. The asymptotic properties of the test are also given. The second method is an extension of Tsiatis' (1980) goodness-of-fit test for logistic regression models. The test is based on partitioning the space of covariates into distinct regions and calculating a test statistic which is a quadratic form of the observed cumulative counts minus the expected cumulative counts. The usefulness of both methods is illustrated on data from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR).

**Keywords:** Bivariate Dale Model; Global Cross-ratio; Kernel Smoothing; Tsiatis' Goodness-of-Fit Test; Additive Bivariate Dale Model

## 1 Introduction

After fitting a model, there is a need to assess the adequacy of the model fit. For logistic regression models, a variety of goodness-of-fit tests have been proposed. Some tests are based on partitioning the covariates into groups or regions (e.g. Tsiatis, 1980), other tests are based on smoothing methods. Le Cessie and van Houwelingen (1991) proposed a goodness-of-fit test based on an unbiased kernel estimate of the standardized residuals of the model.

The BDM is an extension of the logistic regression model to two dimensions and is useful for the analysis of bivariate categorical data. Bivariate categorical data arise in longitudinal studies with ordinal outcomes. However, in this paper we will consider studies where the bivariate ordinal response is obtained in a non-longitudinal way. Examples with bivariate ordinal responses are: the severity of a disease recorded on one or both eyes, ears, kidneys, etc. and recorded as: none, mild, moderate, or severe. In such studies, there is an interest in measuring the dependence on covariates for each outcome separately and for the association between outcomes.

In this paper we extend both le Cessie and van Houwelingen's (1991) and Tsiatis' (1980) goodness-of-fit tests to the BDM. The extension of le Cessie and van Houwelingen's goodness-of-fit test is computed by summing kernel smooth estimates of the standardized residuals which are obtained from the cumulative counts and probabilities for each part of the BDM. The extension of Tsiatis' goodness-of-fit test is based on a score test and is implemented by partitioning the space of covariates into distinct regions.

## 2 The Bivariate Dale Model

Let $\boldsymbol{Y} = (Y_1, Y_2)^T$ be a random vector that takes on values $(k, l)$, $1 \leq k \leq c_1$, $1 \leq l \leq c_2$. The data can be arranged as $c_1 \times c_2$ contingency tables, representing pairs of ordered categorical variables with $c_1$ and $c_2$ levels in the presence of a covariate vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$. Assume that there are $N$ different values of the covariate vector, i.e. $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$ each comprising $n_i$ individuals. Hence, we assume $N$ contingency tables pertaining to the unique values of the covariate vector $\boldsymbol{x}_i$. Let $Z_i^*(d_1, d_2)$ be the number of individuals whose observed response vector is $(d_1, d_2)$. When modeling ordered categorical data, it is more convenient to work with cumulative counts which are defined for each contingency table $i$ as $Z_i(k, l) = \sum_{d_1 \leq k, d_2 \leq l} Z_i^*(d_1, d_2)$. Thus, $Z_i(k, l)$ is the number of individuals wh! ose observed response vector is less than or equal to $(k, l)$. The corresponding cumulative bivariate probabilities are (omitting the dependence of the different terms on the subscript $i$ for ease of notation) $\mu(k, l \mid \boldsymbol{x}) = P(Y_1 \leq k, Y_2 \leq l \mid \boldsymbol{x})$. The cumulative marginal probabilities are obtained by summing over subscripts: $\mu_1(k \mid \boldsymbol{x}) = P(Y_1 \leq k \mid \boldsymbol{x})$ and $\mu_2(l \mid \boldsymbol{x}) = P(Y_2 \leq l \mid \boldsymbol{x})$. The BDM consists of two marginal parts and an association part. A popular choice for the link functions is the following:

$$\begin{aligned} \operatorname{logit}(\mu_1(k \mid \boldsymbol{x})) &= \alpha_{1k} + \boldsymbol{\beta}_1^T \boldsymbol{x} \\ \operatorname{logit}(\mu_2(l \mid \boldsymbol{x})) &= \alpha_{2l} + \boldsymbol{\beta}_2^T \boldsymbol{x} \\ \log(\psi(k, l \mid \boldsymbol{x})) &= \Delta + \alpha_{ka} + \beta_{la} + \delta_{kl} + \boldsymbol{\beta}_3^T \boldsymbol{x}, \end{aligned} \tag{1}$$

$(k = 1, \ldots, c_1 - 1; l = 1, \ldots, c_2 - 1)$, $\alpha_{ka}$, $\beta_{la}$ and $\delta_{kl}$ are association row and column effects and cell-specific effects, respectively, with certain uniqueness constraints (see Dale, 1986). The global cross-ratios in (1) are defined in terms of the cumulative marginal and bivariate probabilities by (suppressing the dependence on $\boldsymbol{x}$ for simplicity) $\psi_{kl} = \{\mu_{kl}(1 - \mu_{1k} - \mu_{2l} + \mu_{kl})\}/\{(\mu_{1k} - \mu_{kl})(\mu_{2l} - \mu_{kl})\}$, with

$$\mu_{kl} = \begin{cases} (1 + [\mu_{1k} + \mu_{2l}](\psi_{kl} - 1) - \omega_{kl})/2(\psi_{kl} - 1) & \text{if } \psi_{kl} \neq 1, \\ \mu_{1k}\mu_{2l} & \text{otherwise,} \end{cases}$$

and $\omega_{kl} = \{[1 + (\psi_{kl} - 1)(\mu_{1k} + \mu_{2l})]^2 + 4\psi_{kl}(1 - \psi_{kl})\mu_{1k}\mu_{2l}\}^{1/2}$. Model fitting proceeds via Newton-Raphson or Fisher scoring techniques (see Dale, 1986).

## 3 Le Cessie and van Houwelingen's GOF Test for the Bivariate Dale Model

An extension of le Cessie and van Houwelingen's goodness-of-fit test to a BDM with estimated regression parameters can be defined as: $\hat{T} = N^{-1}\sum_{i=1}^{N}[\boldsymbol{W}_i^T\hat{\boldsymbol{V}}^{-1/2}(\boldsymbol{Z}-\boldsymbol{n}\hat{\boldsymbol{\mu}})]^2$, where $\boldsymbol{Z} = (\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_N)^T$, $\hat{\boldsymbol{V}} = \mathrm{cov}(\boldsymbol{Z})$, $\boldsymbol{n} = \boldsymbol{1}n_i$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_N)^T$ and $\boldsymbol{1}$ is a column vector of ones. $\boldsymbol{W}_i$ is the column vector with $j$th element the coefficient $w_{ij}/(\sum_k w_{ik}^2)^{1/2}$ such that $w_{ij} = \prod_{u=1}^{p} K[(x_{iu}-x_{ju})/h_u]$. The function $K$ is a one-dimensional non-negative symmetric bounded kernel function on $[-a,a]$. It is normalized according to $\int K(q)dq = 1$ and $\int (K(q))^2 dq = 1$. The bandwidth $h_u$ satisfies $h_u = s_u h$, where $s_u$ is the standard deviation of the $u$th covariate, $u = 1,\ldots,p$. Under the correct BDM, the expression for the mean of $\hat{T}$ for each part of the model is similar to that obtained by le Cessie and van Houwelingen. That is

$$E(\hat{T}) = N^{-1}\sum_{i=1}^{N}\boldsymbol{W}_i^T\boldsymbol{V}_0^{-1/2}(\boldsymbol{I}-\boldsymbol{R})\boldsymbol{V}_0(\boldsymbol{I}-\boldsymbol{R})^T\boldsymbol{V}_0^{-1/2}\boldsymbol{W}_i. \tag{2}$$

Under certain regularity conditions on the design points, and using a uniform kernel function defined as: $K(q) = 1$ if $|q| \leq 1/2$ and 0, otherwise, the asymptotic variance of $\hat{T}$ is obtained for each part of the BDM by

$$\mathrm{var}(\hat{T}) = \frac{2}{N}\left(\frac{2}{3}\right)^d \sum_i \hat{f}(x_i)\prod_u h_u, \tag{3}$$

with $d$ the number of smoothed covariates and $\hat{f}(x_i) = (N\prod h_u)^{-1}\sum_j w_{ij}$ (le Cessie and van Houwelingen, 1991). The matrix $\boldsymbol{R}$ in expression (2) should be replaced by the corresponding matrix for the BDM which is defined by $\boldsymbol{R} = \boldsymbol{G}_0^{-1}X\left[X^T\boldsymbol{n}(\boldsymbol{G}_0^T)^{-1}\boldsymbol{V}_0^{-1}\boldsymbol{G}_0^{-1}X\right]^{-1}X^T(\boldsymbol{G}_0^T)^{-1}\boldsymbol{V}_0^{-1}$, where $\boldsymbol{G}_0 = \partial\boldsymbol{\eta}_0/\partial\boldsymbol{\mu}_0$ and $X$ is a block-diagonal design matrix accounting for each part of the BDM.

## 4 Tsiatis' GOF Test for the Bivariate Dale Model

Consider the BDM (1). The space of the covariates $X_1,\ldots,X_p$ is partitioned into $m = 2^p$ distinct regions in the $p$-dimensional space denoted by $R_1,\ldots,R_m$. The indicator functions $I^{(g_u)}$ are defined by $I^{(g_u)} = 1$, if $X_j \in R_{g_u}$ and 0, otherwise, $(j = 1,\ldots,p)$, with $g_u \in \{0,1\}$ indicating whether the value of the covariate $u$ is less or greater than the average of that covariate, $u = 1,\ldots,p$. An extension of Tsiatis' goodness-of-fit test to

the BDM follows. Define the following BDM

$$\begin{aligned} \text{logit}(\mu_1(k \mid \boldsymbol{x})) &= \alpha_{1k} + \boldsymbol{\beta}_1^T \boldsymbol{x} + \boldsymbol{\gamma}_1^T \boldsymbol{I} \\ \text{logit}(\mu_2(l \mid \boldsymbol{x})) &= \alpha_{2l} + \boldsymbol{\beta}_2^T \boldsymbol{x} + \boldsymbol{\gamma}_2^T \boldsymbol{I} \\ \log(\psi(k,l \mid \boldsymbol{x})) &= \Delta + \alpha_{ka} + \beta_{la} + \delta_{kl} + \boldsymbol{\beta}_3^T \boldsymbol{x} + \boldsymbol{\gamma}_3^T \boldsymbol{I}, \end{aligned} \tag{4}$$

where $\boldsymbol{I}^T = (I^{(1)}, \ldots, I^{(m)})$ and the vectors of partitioning parameters $\boldsymbol{\gamma}_1^T = (\gamma_{11}, \ldots, \gamma_{1m})$, $\boldsymbol{\gamma}_2^T = (\gamma_{21}, \ldots, \gamma_{2m})$, and $\boldsymbol{\gamma}_3^T = (\gamma_{31}, \ldots, \gamma_{3m})$. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\gamma}_3)^T$, Tsiatis' goodness-of-fit test for the BDM consists of testing the hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{0}$. Denoting the vector of regression parameters by $\boldsymbol{\theta}$, where $\boldsymbol{\theta}^T = (\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2, \Delta, \boldsymbol{\alpha}_a, \boldsymbol{\beta}_a, \boldsymbol{\delta}, \boldsymbol{\beta}_3)$. The log-likelihood of the BDM (4) can be defined as: $\ell(\boldsymbol{\gamma}, \boldsymbol{\theta}; \boldsymbol{Z}) = \sum_{i=1}^N (\boldsymbol{C}_i \boldsymbol{Z}_i)^T \log(\boldsymbol{C}_i \boldsymbol{\mu}_i(\boldsymbol{\gamma}, \boldsymbol{\theta}))$, with $\boldsymbol{Z}_i$, $\boldsymbol{\mu}_i$ the vectors of non-redundant cumulative cell counts and probabilities for contingency table $i$, respectively, $i = 1, \ldots, N$. The matrix $\boldsymbol{C}_i$ is a matrix with elements either 0, 1 or $-1$. The extension of Tsiatis' goodness-of-fit test is based on the score test for the BDM defined by: $T = \boldsymbol{U}^T \boldsymbol{\zeta}^{-1} \boldsymbol{U}$, with $\boldsymbol{U} = \partial \ell / \partial \boldsymbol{\gamma}$, the score vector defined by

$$\boldsymbol{U} = \sum_{i=1}^N (\partial \boldsymbol{\eta}_i / \partial \boldsymbol{\gamma})^T \left[ (\partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i)^T \right]^{-1} \boldsymbol{V}_i^{-1} (\boldsymbol{Z}_i - n_i \boldsymbol{\mu}_i).$$

Define the $m \times m$ matrix $\boldsymbol{\zeta}$ by $\boldsymbol{\zeta} = \boldsymbol{A} - \boldsymbol{B}\boldsymbol{W}^{-1}\boldsymbol{B}^T$, where the matrix $\boldsymbol{B} = E(-\partial^2 \ell / \partial \boldsymbol{\gamma} \partial \boldsymbol{\theta}^T)$, defined by

$$\boldsymbol{B} = \sum_{i=1}^N n_i \boldsymbol{Q}_{1i}^T \left[ (\partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i)^T \right]^{-1} \boldsymbol{V}_i^{-1} \left[ (\partial \boldsymbol{\eta}_i / \partial \boldsymbol{\mu}_i) \right]^{-1} \boldsymbol{Q}_{2i}, \tag{5}$$

where $\boldsymbol{Q}_{1i} = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\gamma}$ and $\boldsymbol{Q}_{2i} = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\theta}$. The matrix $\boldsymbol{A} = E(-\partial^2 \ell / \partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T)$ and is obtained by replacing $\boldsymbol{Q}_{2i}$ in (5) by $\partial \boldsymbol{\eta}_i / \partial \boldsymbol{\gamma}$. If $\boldsymbol{Q}_{1i} = \boldsymbol{Q}_{2i} = \partial \boldsymbol{\eta}_i / \partial \boldsymbol{\theta}$ then we obtain the matrix $\boldsymbol{W} = E(-\partial^2 \ell / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T)$, minus the expected information matrix of the BDM. All the above expressions are evaluated at $\boldsymbol{\gamma} = \boldsymbol{0}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, wher! e $hat\boldsymbol{\theta}$ is the MLE of $\boldsymbol{\theta}$. Under $H_0$ the statistic $T$ is asymptotically distributed as $\chi^2_{(\varphi)}$, with $\varphi = \text{rank}(\boldsymbol{\zeta})$.

Results from a limited simulation study (not shown) showed that the extension of le Cessie and van Houwelingen's goodness-of-fit test has more power to detect misspecification of link functions than the extension of Tsiatis' goodness-of-fit test which showed a reasonably high power to detect deviations in a BDM where a quadratic term of a covariate is forgotten.

## 5 Example: The WESDR

The aim of the WESDR was to identify risk factors of diabetic retinopathy among insulin-taking younger-onset diabetics. Originally, severity of diabetic retinopathy of 996 younger-onset diabetics was measured for the

right and left eye and graded on a 10 point ordinal scale. The 10 classes were then combined to four: none, mild, moderate and proliferative. After excluding patients with missing values, data from 720 diabetics were used. Several eye-specific and person-specific covariates were measured.

TABLE 1. The WESDR

| Regression parameters | MLE | SE | $P$-value |
|---|---|---|---|
| *For margins* | | | |
| Cutpoint 1 ($\alpha_1$) | -0.732 | 0.083 | |
| Cutpoint 2 ($\alpha_2$) | 1.518 | 0.093 | |
| Cutpoint 3 ($\alpha_3$) | 3.311 | 0.150 | |
| Age at diagnosis (years) | 0.002 | 0.009 | 0.83 |
| Duration of diabetes (years) | 0.124 | 0.009 | <0.001 |
| Glycosylated haemoglobin (%) | 0.093 | 0.027 | <0.001 |
| Diastolic blood pressure (mm Hg) | 0.041 | 0.007 | <0.001 |
| Gender (male=0, female=1) | -0.317 | 0.137 | 0.02 |
| Proteinuria (absent=0, present=1) | 0.876 | 0.206 | <0.001 |
| Macular oedema (absent=0, present=1) | 1.289 | 0.232 | <0.001 |
| *For association* | | | |
| Intercept ($\Delta$) | 3.638 | 0.180 | |
| Age at diagnosis (years) | -0.038 | 0.023 | 0.10 |
| Doses of insulin per day (1,2,3) | -0.889 | 0.350 | 0.01 |
| Gender (male=0, female=1) | -0.943 | 0.357 | 0.01 |

*Log-likelihood*= $-1155.22$

We used a common marginal model for the right and left eye retinopathy levels. The likelihood-ratio test indicated no evidence against this simplification ($P = 0.918$). Observe that the covariates were centered around their averages. The results of fitting the BDM with link functions as in (1) are shown in Table 1. The common cutpoints are denoted by $\alpha_t$, $t = 1, 2, 3$.
To check the BDM fit of Table 1, the extension of le Cessie and van Houwelingen's goodness-of-fit test $\hat{T}$ is computed and its mean and asymptotic variance are calculated from expressions (2–3), using a uniform kernel function and a bandwidth $h$ of 0.4 for all parts of the BDM. The results showed a lack of model fit in the marginal part ($P$-values are: 0.033 and 0.252 for the marginal and the association part, respectively). The extension of Tsiatis' goodness-of-fit test yielded a $P$-value of 0.025. The results suggest that there is substantial evidence that the BDM of Table 1 does not provide an adequate fit to the WESDR data.
Bustami *et al.* (1999) used these data to illustrate the use of the Additive Bivariate Dale Model (ABDM) which is a nonparametric version of the BDM based on a natural extension of the generalized additive model

(GAM) of Hastie and Tibshirani (1990). By fitting an ABDM to the data they showed a significant departure of linearity in the marginal part of the model for the covariates *duration of diabetes* and *age at diagnosis of diabetes*. They also showed that the estimated marginal smooth functions for those covariates can be approximated by quadratic functions.

TABLE 2. The WESDR

| Regression parameters | MLE | SE | $P$-value |
|---|---|---|---|
| Cutpoint 1 ($\alpha_1$) | -1.854 | 0.137 | |
| Cutpoint 2 ($\alpha_2$) | 0.940 | 0.122 | |
| Cutpoint 3 ($\alpha_3$) | 3.089 | 0.178 | |
| Age at diagnosis (years) | 0.023 | 0.010 | 0.02 |
| Age$^2$ | -0.005 | 0.001 | <0.001 |
| Duration of diabetes (years) | 0.241 | 0.014 | <0.001 |
| Duration$^2$ | -0.007 | 0.001 | <0.001 |
| Glycosylated haemoglobin (%) | 0.124 | 0.028 | <0.001 |
| Diastolic blood pressure (mm Hg) | 0.023 | 0.007 | 0.001 |
| Gender (male=0, female=1) | -0.396 | 0.142 | 0.01 |
| Proteinuria (absent=0, present=1) | 0.850 | 0.208 | <0.001 |
| Macular oedema (absent=0, present=1) | 1.355 | 0.244 | <0.001 |

*Log-likelihood*= -1072.35

Table 2 shows the marginal part of the BDM fit after adding age$^2$ and duration$^2$. Both age$^2$ and duration$^2$ are highly significant. For the BDM of Table 2, both goodness-of-fit tests provided no evidence for a lack of fit. The extension of le Cessie and van Houwelingen's goodness-of-fit test yielded $P$-values of 0.330 and 0.123 for the marginal and the association part, respectively. The extension of Tsiatis' goodness-of-fit test yielded a $P$-value of 0.472. The likelihood-ratio test comparing the BDM fits of Tables 1 and 2 showed a highly significant improvement in the model fit ($P < 0.001$).

## References

Bustami, R., Lesaffre, E., Molenberghs, G., and Loos, R., Danckaerts, M., Vlietinck, R. (1999). Modeling bivariate responses smoothly: examples from ophthalmology and genetics. *Revised.*

Dale, J.R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Hastie, T.J., and Tibshirani, R.J. (1990). *Generalized Additive Models.* London: Chapman & Hall.

Le Cessie, S. and van Houwelingen, J.C. (1991). A goodness of fit test for binary regression models, based on smoothing methods. *Biometrics*, **47**, 1267–1282.

Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, **67**, 250–251.

# Specification Testing of Univariate Continuous-Time Interest Rate Models

Renato G. Flôres Jr.[1] and Cristian Huse[2]

[1] EPGE/FGV, Praia de Botafogo, 190, Rio de Janeiro, Brazil
[2] EPGE/FGV, Praia de Botafogo, 190, Rio de Janeiro, Brazil and IBMEC, Av Rio Branco, 108, Rio de Janeiro, Brazil

**Abstract:** We propose a general framework for specification testing of univariate stationary continuous-time interest rate models. Based on the Pearson system of distributions, we define a class of stationary densities that encompasses many of those in the most used models of the finance literature, such as the Vasicek (1977) and the Cox-Ingersoll-Ross (1985) ones. By rejecting a general class given by the corresponding differential equation, one can strongly reject the models which are nested within this particular class. This avoids ad hoc choices of interest rate models and the mispricing of interest rate derivative securities.

**Keywords:** Specification Testing; Diffusions; Interest Rate Models; Pearson System; Derivative Securities.

## 1 Introduction

One undesirable feature concerning the estimation of continuous-time interest rate models is the *ad hoc* choice of the process to be estimated, a fact that may lead to the mispricing of derivative securities.
The usual way one estimates continuous-time interest rate models is, first, to choose the specification to be estimated, according to what one believes is the shape of the volatility function, as most of the interest rate models assume a linear mean-revearting drift. After estimating the model chosen, one may assess its adequacy using the Conditional Moments Test (Newey (1985)). The drawback of such a procedure is the *ad hoc* choice of the moments being considered - usually one considers only moments up to order $p$, where $p$ is usually small. The question to be addressed is, assuming there is misspecification of the model chosen, what if the moment to be causing this misspecification is of order $p+q$, while the moments being considered are up to order $p$. If one considers the density of the process, instead of its first $p$ moments, there is an informational gain, as all moments of the process are being considered.
More specifically, consider a continuous time univariate model given by

$$dx_t = A(x_t)dt + \sqrt{B(x_t)}dW_t, \tag{1}$$

where $A(.)$ and $B(.)$ are respectively the drift and the volatility of the process. In the first section of the paper, we present a general class of processes, which we call Pearson class, which encompasses many of the continuous time interest rate models - the case when the drift A(.) is affine and the volatility function B(.) is quadratic:

$$A(x) = \alpha_0 + \alpha_1 x, \tag{2}$$

$$B(x) = \beta_0 + \beta_1 x + \beta_2 x^2, \tag{3}$$

It is easily seen that some of the most used interest rate models are nested in this more general class:

TABLE 1 - Models nested in the Pearson class

| Model | Restrictions on the Drift | Restrictions on the Volatility |
|---|---|---|
| Vasicek | - | $\beta_1 = \beta_2 = 0$ |
| CIR | - | $\beta_0 = \beta_2 = 0$ |
| Courtadon | - | $\beta_0 = \beta_1 = 0$ |
| Duffie-Kan | - | $\beta_2 = 0$ |
| Merton | $\alpha_1 = 0$ | $\beta_1 = \beta_2 = 0$ |
| Dothan | $\alpha_0 = \alpha_1 = 0$ | $\beta_0 = \beta_1 = 0$ |
| GBM | $\alpha_0 = 0$ | $\beta_0 = \beta_1 = 0$ |

By providing a class os densities encompassing the models from Table 1, we provide an answer to a common question addressed by researchers and practitioners in the field: "Which model to consider ?". Thus, given that the density of the process at stake - according to the test - is a member of the Pearson class, one may compare alternative continuous-time models.
This paper is divided as follows. In section 2, we describe de Pearson class of distributions. In the third section, we present the drawbacks of Aït-Sahalia's (1996a) nonparametric specification test, so as to justify our approach, which we present in section 4. After considering both size and power of the test in section 5, and implementing the methodology for a interest rate time series in section 6, we conclude.

## 2 The Pearson Equation

In this section we propose a class that encompasses some of the most used continuous time interest rate models. Consider the Kolmogorov forward, or Fokker-Planck, equation, which describes the transition densities of continuous-time Markov processes without jump:

$$\frac{\partial}{\partial t} p(x,t,y,t') = -\frac{\partial}{\partial x}(A(x).p(x,t,y,t')) + \frac{\partial^2}{\partial x^2}(B(x).p(x,t,y,t')), \tag{4}$$

where: $p(x,t,y,t')$:= transition density from point (y,t') to (x,t); $A(x)$:= drift of the process; $B(x)$:= volatility of the process. Assume that the boundary conditions are those corresponding to reflecting barriers:

$$-A(x).p(x,t,y,t') + \frac{\partial^2}{\partial x^2}(B(x).p(x,t,y,t')) = 0, \tag{5}$$

on an interval I with endpoints $x_1$ and $x_2$ .
An alternative representation of the diffusion described by (4) could be given by the Itô representation

$$dx_t = A(x_t)dt + \sqrt{B(x_t)}dW_t, \tag{6}$$

where $\{W_t, t \geqslant 0\}$ is a standard Brownian motion. This representation is the most used in applications to finance, although it leads to some unusual facts in the bivariate case (see Flôres and Huse (2000)).
Wong (1964) considers the first-order probability density functions $\pi(x)$ satisfying the Pearson equation

$$\frac{d\pi(x)}{dx} = \frac{a_0 + a_1 x}{b_0 + b_1 x + b_2 x^2}\pi(x), \tag{7}$$

hereafter, (7) will be called the Pearson factor. It is shown that by suitably identifying the functions $A(x)$ and $B(x)$ in (4) with the polynomials $a_0 + a_1 x$ and $b_0 + b_1 x + b_2 x^2$ in (7), a class of stationary Markov processes is constructed for which

$$\lim_{t\to\infty} p(x|x_0,t) = \int_{x_1}^{x_2} \pi(x_0).p(x|x_0,t)dx_0 = \pi(x), \tag{8}$$

where $\pi(x)$ satisfies (7). It is shown (Wong (1964)) that the identification scheme is such that the Pearson equation (7) uniquely specifies the Fokker-Planck equation (4). Furthermore, it is shown, after separation of variables, that

$$B(x) = K(b_0 + b_1 x + b_2 x^2), \tag{9}$$

and

$$A(x) = K(a_0 + a_1 x) + \frac{dB(x)}{dx} = K(a_1 + 2b_2)x + K(a_0 + b_1), \tag{10}$$

where $K$ is a positive (time) scaling factor - Wong (1964) sets $K$=1 without loss of generality. It is straightforward to see that this class of processes nests all the process with affine drift and quadratic volatility. After estimating the diffusion, it is possible to obtain the Pearson equation (7) - the density $\pi(.)$ - from the volatility and the drift of the diffusion (9)-(10). Thus our procedure consists on estimating $A(.)$ and $B(.)$, so as to recover the Pearson equation associated to the process at stake. It is important to stress that, by rejecting that the process satisfies (7), one will be rejecting the hypotheses that the process at stake has affine drift and quadratic volatility.

## 3 Nonparametric Testing

An elegant way to construct specification tests using a nonparametric framework has been proposed by Aït-Sahalia (1996a). Following his idea, one defines a testing strategy so as to check specification by matching density functionals, i.e. testing the null hypothesis

$$H_0 : \exists \theta_0 \in \Theta s.t. \pi(., \theta_0) = \pi(.), \tag{11}$$

against the alternative that there is no such $\theta_0 \in \Theta$. It is worth to notice that if one estimates the density function using nonparametric kernel methods, one gets consistent estimates under both the null and the alternative hypotheses; the parametric density estimator, however, is consistent only under the null. A natural way to construct a test is by comparing the distance between both parametric and nonparametric estimates under a suitable metric. For that purpose, consider the distance

$$\Psi = \int (\pi(x, \theta) - \pi(x))^2 \pi(x) dx, \tag{12}$$

The asymptotic distribution of the test statistic is derived using Aït-Sahalia' s (1994) functional delta method The null is rejected whenever the test statistic is "large". The distance (11) weighs the difference between the densisites according to their relevance, putting more weight on values of the process more likely to appear. Conversely, discrepancies between the densities for values not so likely to appear are less penalized.
One appealing feature of this approach is that, opposed to the framework of a Conditional Moment Test, where the moments to be considered are chosen in a rather informal way, the test statistic here proposed considers the entire density of the process under study, which means more information, as all moments are being considered simultaneously. Thus if a specification has to be rejected due to its k-th moment, the nonparametric test will be able to reject it, while the CM test will do so if, and only if, this particular moment is included in the test by the researcher.
Though extremely elegant, Aït-Sahalia's test has a remarkable drawback, as pointed out in Pritsker (1998). The time series of US interest rates is dependent and highly persistent, but the asymptotic distribution of the test statistics treats observations as if they were iid, understating its variance and possibly causing the test true interest rate models to often i.e. according to Pritsker, there is a high probability of Type I errors associated to the test when dealing with persistent series. In fact, using simulated series, Pritsker shows that there is evidence of size distortion for the test, as well as modest power.

## 4 Back to the Parametric World

Given the drawbacks of Aït-Sahalia's approach, we propose in this section an alternative testing procedure for diffusions.

### 4.1 Diffusion Estimation

Most of the papers concerning the estimation of continuous-time models discretize the model before estimating it (see Chan et al (1992)) and use either the Generalized Method of Moments or Maximum Likelihood to obtain the parameters. First of all, one should recall that the Pearson equation has 5 parameters, so that the usual GMM using only the first two moments (four moment conditions) is not identified. Moreover, fitting some of the members of the Pearson class can get so difficult that it is almost never attempted (Johnson and Kotz (1994), chapter 12). An alternative way of estimating the process at stake is to use the Generalized Least Squares method (see Carrol and Ruppert (1987) and references therein). One should recall that GLS and ML are asymptotically equivalent estimators - the GLS estimator being inefficient only if the distributional assumption of the ML estimator is correct. Thus, we will use use the parameters estimated by GLS to simulate processes driven by a member of the Pearson family of distributions.

### 4.2 Simulation of the Diffusion Sample Paths

The adequacy of the model is tested via simulated-based inference methods, i.e. simulating sample paths based on the parameters previously estimated.

## 5 Conclusion

Besides testing the adequacy of the most widely used interest rate models, our approach leads to and interesting and promising research topic: models with nonlinear drifts, such as in Constantinides (1992). Therefore, one needs generalizations of the Pearson equation, such as the Multimodal Distributions by Cobb et al (1983).

## References

Aït-Sahalia, Y. (1994). The Delta Method for Nonparametric Kernel Functionals. Mimeo, University of Chicago.

Aït-Sahalia, Y. (1996a). Testing Continuous Time Interest Rate Models of the Spot Interest Rate. Review of Financial Studies.

Aït-Sahalia, Y. (1996b). Nonparametric Pricing of Interest Rate Derivative Securities. *Econometrica*, **64**, 527-560.

Carrol, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Monographs on Statistics and Applied Probability, 30. London: Chapman and Hall.

Chan, K. C., Karolyi, G.A., Longstaff, F.A. and Sanders, A.B. (1992). An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *Journal of Finance*, **47**, 1209-1227.

Cobb, L., Koppstein, P. and Chen, N.H. (1983). Estimation and Moment Recursion Relations for Multimodal Distributions of the Exponential Family. *Journal of the American Statistical Association*, **78**, 124-130.

Constantinides, G. M. (1992). A Theory of the Nominal Term Structure of Interest Rates. *Review of Financial Studies*, **5**, 531-552.

Cox, J., Ingersoll, J. and Ross, S. (1985). A Theory of the Term Structure of Interest Rates. *Econometrica*, **53**, 385-407.

Flôres, R. G. and Huse, C. (2000). Exploratory Semiparametric Analysis of Two-Dimensional Diffusions in Finance. Mimeo, EPGE/FGV.

Johnson, N. L. and Kotz, S. (1994). *Continuous Univariate Distributions.* New York: John Wiley & Sons.

Newey, W. (1985). Generalized Method of Moments Specification Testing. *Journal of Econometrics*, **29**, 229-256.

Pritsker, M. (1998). Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models. *Review of Financial Studies*, **11**, 449-487.

Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics*, **5**, 177-188.

Wong, E. (1964). The Construction of a Class of Stationary Markov Processes. In *Stochastic Processes in Mathematical Physics and Engineering*, Proceedings of Symposia in Applied Mathematics, Vol. XVI, ed. By R. Bellman. Providence, RI: American Mathematical Society, 264-276.

# Bayesian P-Splines

Stefan Lang[1] and Andreas Brezger[1]

[1] Institute of Statistics, University of Munich, Ludwigstr. 33, D-81541 Munich

**Abstract:** P-splines are an attractive approach for modelling nonlinear smooth effects of covariates within the generalized additive and varying coefficients model framework. In this paper we propose a Bayesian version for P-splines. Compared to the traditional approach, Bayesian P-splines have several advantages. Among others, these are simultaneous estimation of smooth functions and the smoothing parameters and easy extendability of the models, e.g. to mixed models with random effects for serially or spatially correlated response. In a second step we generalize the P-spline approach for one dimensional curves to two dimensional surface fitting for modelling interactions between metrical covariates. We illustrate our approach by a thorough analysis of a forest health survey.

**Keywords:** generalized additive models, MCMC, P-Splines, surface fitting, varying coefficients

## 1 Indroduction

Consider the *additive model* (AM) with predictor

$$E(y|x) = \mu = \gamma_0 + f_1(x_1) + \cdots + f_p(x_p) = \eta$$

where the mean of a response variable $y$ is assumed to be the sum of smooth functions $f_j$. The metrical response $y$ is assumed to follow a Gaussian distribution. To allow for non-Gaussian response the AM is extended to *generalized additive models* (GAM) by assuming that the distribution of the response belongs to an exponential family and that the mean $\mu$ is related to the predictor by a response function $h(\eta) = \mu$. Several proposals are available for modelling and estimating the smooth functions $f_j$, see e.g. Hastie, Tibshirani, (1990) for an overview. An attractive approach, based on *penalized regression splines* (P-splines), has been presented by Eilers and Marx (1996). Here it is assumed that the effect $f$ of a covariate $x$ can be approximated by a polynomial spline written in terms of a linear combination of B-spline basis functions. To ensure enough flexibility a moderate number of equally spaced knots (20-40) within the domain of $x$ is chosen. Sufficient smoothness of the fitted curve is achieved through a difference penalty on adjacent B-spline coefficients.
This paper presents a Bayesian version of the P-splines approach by Eilers and Marx. This is achieved by replacing difference penalties by their

stochastic analoga, which serve as smoothness priors for the unknown regression coefficients. In a second step we generalize the P-spline approach for one dimensional curves to two dimensional surface fitting by assuming that the unknown surface can be approximatd by the tensor product of one dimensional B-splines. Smoothness is now achieved by smoothness priors common in spatial statistics. The proposed methods are a generalization of an approach by Fahrmeir, Lang (1999) who use the close relationship between dynamic models and nonparametric regression for Bayesian inference in GAM's. Software for fitting the models in this paper is included in the program *BayesX* for Bayesian inference via MCMC. The program is available via internet under `http://www.stat.uni-muenchen.de/~lang/`. Our methods are illustrated by an application to a forest health survey.

## 2 Bayesian GAM's based on P-Splines

Consider regression situations, where observations $(y_i, x_i, w_i)$, $i = 1, \ldots, n$ on a response $y$, a vector of metrical covariates $x = (x_1, \ldots, x_p)$ and a vector of further covariates $w = (w_1, \ldots, w_q)$ are given. Generalized additive models (Hastie, Tisbshirani 1990) assume that, given $x_i$ and $w_i$ the distribution of $y_i$ belongs to an exponential family, with mean $\mu_i = E(y_i|x_i, w_i)$ linked to a semiparametric additive predictor $\eta_i$ by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + w_i'\gamma. \tag{1}$$

Here $h$ is a known response function and $f_1, \ldots, f_p$ are unknown smooth functions of the covariates. The linear combination $w_i'\gamma$ corresponds to the usual parametric part of the predictor. In the P-splines approach by Eilers and Marx (1996) it is assumed that the unknown functions $f_j$ can be approximated by a spline of degree $l$ with equally spaced knots $\zeta_{j0} = x_{j,min} < \zeta_{j1} < \cdots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$ within the domain of $x_j$. It it well known that such a spline can be written in terms of a linear combination of $m = r + l$ B-spline Basis functions $B_{j\rho}$, i.e.

$$f_j(x_j) = \sum_{\rho=1}^{m} \beta_{j\rho} B_{j\rho}(x_j).$$

The Basis functions $B_{j\rho}$ are defined only locally in the sense that they are nonzero only on domain spanned by $2 + l$ knots. It would be beyond the scope of this paper to go into the details of B-splines and their properties, see e.g. Eilers and Marx (1996) for a nice introduction. By defining the $n \times m$ design matrices $X_j$, where the element in row $i$ and column $\rho$ is given by $X_j(i, \rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor (1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + W'\gamma. \tag{2}$$

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jm})$, $j = 1, \ldots, p$ corresponds to the vectors of unknown regression coefficients. The matrix $W$ is the usual design matrix for fixed effects. In a simple regression spline approach the unknown regression coefficients are estimated using standard maximum likelihood algorithms for generalized linear models. However, a crucial point with simple regression splines is the choice of the number of knots. If the number of knots is too small, the resulting spline may be not flexible enough to capture the variability of the data. If, on the contrary, a too large number of knots is chosen estimated curves tend to overfit the data and, as a result, too rough functions are obtained. As a remedy to these problems Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves.
In a Bayesian approach, as considered in this paper, unknown parameters $\beta_j$ and $\gamma$ are considered as random variables and have to be supplemented with appropriate prior distributions. For the fixed effects parameters $\gamma$ we usually assume diffuse priors, i.e. $\gamma_j \propto const$. The stochastic analoga of difference penalties are first and second order random walks

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{3}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or $\beta_{j1}$ and $\beta_{j2} \propto const$, for initial values, respectively. The amount of smoothness is controlled by the additional variance parameters $\tau_j^2$, which are the analoga of smoothing parameters in a frequentist approach. The priors (3) can be equivalently written in form of a global smoothness prior

$$\beta_j \propto exp\left(\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right)$$

with appropriate penalty matrix $K_j$.
For full Bayesian inference the unknown variance parameters $\tau_j^2$ are also considered as random and estimated simultaneously together with the unknown $\beta_j$. Therefore, hyperpriors are assigned to them in a further stage of the hierarchy by highly dispersed inverse gamma priors $p(\tau_j^2) \sim IG(a_j, b_j)$. Our Bayesian model is completed by assuming conditional independence of observations $y_i$ given the parameters and mutual independence of parameter vectors $(\beta_j, \tau_j^2)$, $j = 1, \ldots, p$ and $\gamma$.

## 3 Modelling interactions

The models considered so far are not appropriate for modelling interactions between covariates. A special way to deal with interactions are *varying coefficients models* (VCM). Here nonlinear terms $f_j(x_{ij})$ are generalized to

$f_j(x_{ij})z_{ij}$, where $z_j$ may be a component of $x$ or $w$ or a further covariate. Covariate $x_j$ is called the effect modifier of $z_j$ because the effect of $z_j$ varies smoothly over the range of $x_j$. Estimation of VCM's poses no further difficulties, since only the design matrices $X_j$ in (1) have to be redefined by multiplying each element in row $i$ of $X_j$ with $z_{ij}$. VCM's are particularly useful if the interacting variable $z_j$ is categorical. Consider now situations where both interacting covariates are metrical. In that case the interaction between two covariates $x_j$ and $x_s$ may be modelled by a two dimensional smooth surface $f_{js}(x_j, x_s)$ leading to a predictor of the form

$$\eta_i = \cdots + f_j(x_{ij}) + f_s(x_{is}) + f_{js}(x_{ij}, x_{is}) + \ldots \quad .$$

Here we assume that the unknown surface can be approximated by the tensor product of the two one dimensional B-splines, i.e.

$$f_{js}(x_j, x_s) = \sum_{\rho=1}^{m} \sum_{\nu=1}^{m} \beta_{js\rho\nu} B_{j\rho}(x_j) B_{s\nu}(x_s).$$

Priors for $\beta_{js} = (\beta_{js11}, \ldots, \beta_{jsmm})$ are now based on spatial smoothness priors common in spatial statistics. Since there is no natural ordering of parameters, priors have to be defined by specifying the conditional distributions of $\beta_{js\rho\nu}$ given neighboring parameters. The most simplest prior specification based on the 4 nearest neighbors can be defined by

$$\beta_{js\rho\nu} \sim N\left(\frac{1}{4}(\beta_{js\rho-1,\nu} + \beta_{js\rho+1,\nu} + \beta_{js\rho,\nu-1} + \beta_{js\rho,\nu+1}), \frac{\tau_{js}^2}{4}\right)$$

for $\rho, \nu = 2, \ldots, m-1$ and appropriate changes for corners and edges. This is a direct generalization of a first order random walk in one dimension. Another choice is based on the Kronecker product $K_{js} = K_j \otimes K_s$ of penalty matrices of the main effects. A nice interpretation of such priors in terms of locally polynomial fits is given in Besag, Kooperberg (1995).

## 4 Bayesian inference via MCMC

Bayesian inference is based on the posterior

$$p(\beta, \tau^2, \gamma) \propto L(y, \beta, \gamma) \prod_{j=1}^{p} \left(p(\beta_j|\tau_j^2)p(\tau_j^2)\right) p(\gamma) \tag{4}$$

with $\tau^2 = (\tau_1^2, \ldots, \tau_p^2)$ and $L(y, \beta, \gamma)$ denoting the likelihood. For notational simplicity the dependence from additional parameter vectors $\beta_{js}$ for two dimensional surfaces has been omitted in (4). MCMC simulation is based on drawings from full conditionals of blocks of parameters given the rest and the data. In general, we partition the vectors $\beta_j$, $j = 1, \ldots, p$ into smaller

blocks $\beta_j[u,v]$ and draw random samples from the unnormalized full conditionals $p(\beta_j[u,v]|\cdot)$ by Metropolis-Hastings (MH) steps with *conditional prior proposals*. Details can be found in Fahrmeir, Lang (1999). However, in some important special cases, more sophisticated algorithms are possible. If the response is Gaussian the full conditionals for $\beta_j$ are Gaussian, thus allowing to update $\beta_j$ in one large block by drawing samples directly from the Gaussian full conditional. Numerical efficiency is guaranteed by applying Cholesky decompositions for band matrices, see Fahrmeir, Lang (2000). If the response is binomial and a probit link is chosen, sampling of parameters can be simplified by considering latent Gaussian utilities associated with the two categories of the response. In this case the full conditionals for $\beta_j$ are once again Gaussian and the efficient sampling schemes developed for Gaussian response can be used. The sampling schemes based on latent utilities are particularly attractive for GAM's with *multicategorical response*. Important examples are the independent probit model for nominal response and the cumulative threshold model for ordinal response, see Fahrmeir, Lang (2000) for details.

## 5 Application to forest health data

In this longitudinal study on the state of trees, we analyse the influence of calendar time, age of trees, canopy density and location of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 1997. There are 80 observation points with occurence of beeches spread over the whole area. We use the degree of defoliation as a binary indicator for damage state, with $y_{it} = 1$ for "light or distinct damage" of tree $i$ in year $t$, $y_{it} = 0$ for "no damage". Covariates used here are defined as follows:

- $A$ age of tree at the beginning of the study in 1983, measured in 3 categories $A^{(1)}$ = "below 50 years", $A^{(2)}$ = "between 50 and 120 years", and $A^{(3)}$ = "above 120 years" (reference category);
- $C$ Canopy density at the stand measured in percentages.

Based on previous results, we estimate a logit model with predictor

$$\eta_{it} = f_1(t) + f_2(t)A_i^{(1)} + f_3(t)A_i^{(2)} + f_4(CD_{it}) + b_i,$$

where $t$ is calendar time in years. The nonlinear functions $f_1, \ldots, f_4$ are modelled by P-splines of degree 3 with second order random walk penalties. To account for unobserved spatial heterogeneity our predictor contains an additional tree specific random effect $b_i$. We make the usual assumption that the $b_i$'s are i.i.d. Gaussian, $b_i|\tau_b^2 \sim N(0, \tau_b^2)$ and define a highly dispersed inverse Gamma prior for $\tau_b^2$. Estimation of such a *mixed model* via

MCMC poses no further difficulties, see e.g. Fahrmeir, Lang (1999). Figure 1 shows estimated effects of nonlinear functions. Figure a) corresponds to the time trend of old trees over 120 years. We observe that trees recovered after 1986 until 1992, then the probability for damage increases again. For young trees the time varying effect is declining over the observation period and significantly negative, i.e., in comparison to old trees, young trees have lower probabilities of being damaged and they seem to recover better. The effect for trees of medium age is similar. The monotonic decrease of the effect of canopy densities $\geq 30\%$ gives evidence that beeches get more shelter from bad environmental influences in stands with high canopy density.

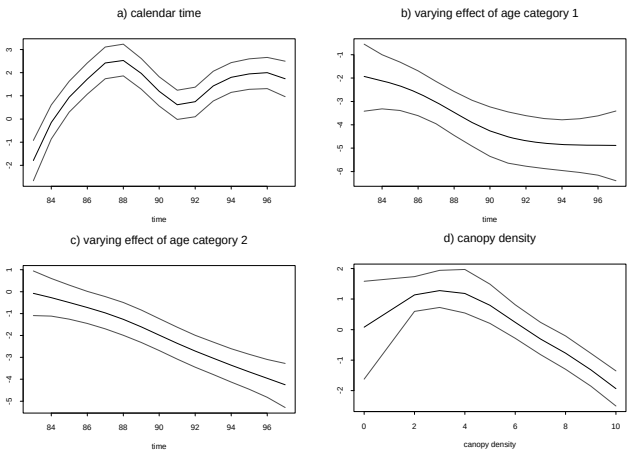



FIGURE 1. Estimated nonparametric functions. Shown is the posterior mean within 80 % credible regions.

## References

Besag, J., Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, **82**, 733-746.

Eilers, P.H.C., Marx, B.D. (1996). Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statist. Sci.*, 11 (2), 89-121.

Fahrmeir, L., Lang, S. (1999). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. Discussion Paper 169, SFB 386, LMU-Munich.

Fahrmeir, L., Lang, S. (2000). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. Discussion Paper 194, SFB 386, LMU-Munich.

Hastie, T., Tibshirani, R. (1990). *Generalized additive models.* Chapman & Hall, London.

# LS estimation in a Semiparametric Additive Regression Model with dependent errors

Germán Aneiros Pérez[1]

[1] Universidad de La Coruña Departamento de Matemáticas Facultad de Informática Campus de Elviña s/n La Coruña SPAIN e-mail: ganeiros@udc.es

**Abstract:** Suppose that $y_i = \zeta_i^T \beta + m(t_i) + \varepsilon_i$, $i = 1, ..., n$, where the $(p \times 1)$-vector $\beta$ and the function $m$ are unknown, and the errors $\varepsilon_i$ pertain to a strong mixing process. The problem of bandwidth selection for a LS estimator of $\beta$, based on residuals from nonparametric fits, is addressed here. We obtain an insightful representation for a bandwidth optimizing terms of lower orden than $n^{-1}$.

**Keywords:** Partial linear models; Bandwidth selection.

## 1 Introduction

We consider that we have an observed data set $\{(y_i, \zeta_i^T, t_i)^T\}_{i=1}^n$generated by the semiparametric model

$$y_i = \zeta_i^T \beta + m(t_i) + \varepsilon_i, i = 1, ..., n, \tag{1}$$

where $(y_i, \zeta_i^T, t_i)^T \in R \times R^p \times [0, 1]$, $\beta$ is a $(p \times 1)-$vector of unknown parameters, $m(\cdot)$ is an unknown real-valued function defined on $[0, 1]$ (without loss of generality) and $\varepsilon_i$ a sample of errors identically distributed. This model, called a "partial linear model" because is a mixture of a linear and a nonparametric part in a regression model with $p + 1$ regressor variables, was introducted by Engle et al. (1986) to study the effect of weather on electricity demand.

We focus on estimation of $\beta$. Based on residuals from nonparametric fits (by kernel estimation) of the observation vector $\mathbf{y}$ and the design matrix $\mathbf{X}$ for the linear parametric part against the covariate vector of the nonparametric component, Speckman (1988) and Gao (1995) study the asymptotic normality and the consistence of the LS estimator of $\beta$. Speckman supposes that the errors $\varepsilon_i$ are i.i.d. while Gao works with a class of linear processes for $\varepsilon_i$.

Semiparametric estimators tipically require the selection of a smoothing parameter $h$ to guarantee a good behaviour of the same. There are not too many works about this crucial problem. Speckman suggested to use generalized cross-validation to choose the bandwidth that is used in the estimation of the regression function $r(\zeta_i^T, t_i) = \zeta_i^T \beta + m(t_i)$ and Aneiros

and Quintela (1999) proved an optimality property of the modified cross-validation criterion when the errors are strong mixing. Of course, these bandwidths no necessarily are optimal for $\beta$'s estimation. For the LS estimator based on the local polynomial regression scheme of order q, Linton (1995), assuming independence in the errors, obtain the expression of an optimal bandwidth (in a mean squared error sense) by means of second order approximations for the MSE. The main point of this paper is an analysis of how we should select the bandwidth for the LS estimator based on kernel estimation when we assume a strong mixing condition for the errors. Our results are qualitatively similar to those developed in Linton (1995).

## 2 The estimator and the results

If we consider the model (1) without the lineal component,

$$y_i = m(t_i) + \varepsilon_i, \tag{2}$$

a kernel type estimation can be written as

$$m_{n,h}(t) = \sum_{i=1}^{n} w_{n,h}(t, t_i) y_i, \tag{3}$$

with $w_{n,h}(\cdot, t_i)$ a weight function (obtained from a kernel function $K(\cdot)$, and indexed by a smoothing parameter $h$) that can take different forms, providing thus different estimators. When the function to be estimated has bounded support, kernel estimators are affected by boundary effects when estimating near the extremes of the observation interval. We can remedy these effects using boundary kernels near the boundary of the observation interval (see Gasser and Müller (1984)).
In this paper, we focus on the Gasser-Müller (1984) estimate, but our results can be extended to use other different types of kernel estimators. Thus, we work with

$$w_{n,h}(t, t_i) = h^{-1} \int_{(i-1)/n}^{i/n} K(\frac{t-u}{h}) du \tag{4}$$

where $h > 0$, $t_i = (i - 0.5)/n$ and $K$ is a function with support $[-1, 1]$. If $t = qh \in [0, h)$ or $t = 1 - qh \in (1 - h, 1]$ (i.e., $t$ is a boundary point), where $h < 1/2$, then we change $K(\cdot)$ by a boundary kernel $K_q(\cdot)$ (see Gasser and Müller (1984)).
If $\beta$ is known, having present (1), (2) and (3), a natural estimator of $m(\cdot)$ is

$$\widehat{m}_{n,h,\beta}(t) = \sum_{i=1}^{n} w_{n,h}(t, t_i)(y_i - \zeta_i^T \beta).$$

Based on the model

$$y_i = \zeta_i^T \beta + \widehat{m}_{n,h,\beta}(t_i) + \varepsilon_i$$

the LS estimator $\widehat{\beta}_h$ of $\beta$ can be defined by

$$\sum_{i=1}^{n} (y_i - \zeta_i^T \widehat{\beta}_h - \widehat{m}_{n,h,\widehat{\beta}_h}(t_i))^2 = \min!.$$

It's easy to obtain that

$$\widehat{\beta}_h = (\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \widetilde{\mathbf{y}} \tag{5}$$

where $\mathbf{y} = (\mathbf{I} - \mathbf{W})\mathbf{y}$, $\mathbf{X} = (\mathbf{I} - \mathbf{W})\mathbf{X}$, $\mathbf{W}$ is a smoothing matrix with elements $\{w_{n,h}(t_i, t_j)\} = \{w_{ij}\}$ (the dependence of W on n and h is suppressed), $\mathbf{y} = (y_1, ..., y_n)^T$ and $\mathbf{X}_{n\times p} = (x_{ij})_{i,j} = (\zeta_1, ..., \zeta_n)^T$.
Between $x_{ij}$ and $t_i$ we will suppose the relation

$$x_{ij} = g_j(t_i) + \eta_{ij},$$

where the $g_j(\cdot)$ $(j = 1, ..., p)$ are unknown real-valued functions defined on $[0, 1]$ and $\eta_{ij}$ random variables with mean zero (Speckman, Linton).
We will consider the standardized quantity $T = \sigma^{-1} n^{1/2}(\widehat{\beta}_h - \beta)$ for any $(p \times 1)-$vector $\mathbf{c} \neq \mathbf{0}$, where $\sigma^2 = \sigma_\varepsilon^2 \mathbf{c}^T \Sigma_\eta^{-1} \mathbf{c}$, $\sigma_\varepsilon^2 = Var[\varepsilon_i]$ and $\Sigma_\eta = E[\eta_i \eta_i^T]$ (we denote $\eta_i^T = (\eta_{i1}, ..., \eta_{ip})$ ). Using Linton's notation, it can be seen that

$$T = T^* + R_T. \tag{6}$$

With a bandwidth $h = h_n \in [an^{-\pi}, bn^{-\pi}]$ (where $0 < a < b < \infty$ are constants and $\pi = 2/(4v+1)$; here we suppose that the regression functions $m(\cdot)$ and $g_j(\cdot)$ have $v$ continuous derivatives), we have that $R_T = o_p(n^{-2\mu})$, where $\mu = (4v - 1)/(2(4v + 1))$. Under suitable assumptions, we obtain:
**Theorem**
*(a) $T$ and $T^*$ have the same distribution to order $n^{-2\mu}$.*
*(b) $sup_n E[T^{*2}] < \infty$.*
*(c) $E[T^*] = n^{1/2} h^{2v} B + o(n^{-\mu})$ and $Var[T^*] = 1 + n^{-1} h^{-1} V + o(n^{-2\mu})$.*
*Here, the bounded quantities $B$ and $V$ are:*

$$B = \sigma^{-1} c^T \Sigma_\eta^{-1} \frac{\alpha_v^2}{(v!)^2} \int_0^1 \mathbf{g}^{(v)}(t) m(t)\, dt \tag{7}$$

*and*

$$V = (1 + 2 \sum_{k=1}^{\infty} \gamma(k)) \int K_*^2(u)\, du \tag{8}$$

*where* $\mathbf{g}^{(v)}(t) = (g_1^{(v)}(t), ..., g_p^{(v)}(t))^T, \alpha_v(K) = \int_{-1}^{1} u^v K(u)du, K_*(u) = K * K(u) - 2K(u)$ *is the "twiced" kernel, in which* $K*K(u) = \int K(u-s)K(s)ds$ *is the convolution of* $K$ *with itself and* $\gamma(k) = \sigma_\varepsilon^{-2} Cov(\varepsilon_i, \varepsilon_{i+k})$.

**Corollary** *The asymptotically optimal* $h_{n0}$ *for* $c^T\widehat{\beta}_h$ *is given by* $h_{n0} = A_0 n^{-\pi}$, *where* $\pi = 2/(4v+1)$ *and* $A_0 = (V/(4vB^2))^{\frac{1}{4v+1}}$. *With this bandwidth, we have that the AMSE of* $c^T\widehat{\beta}_{h_n}$ *is*

$$\sigma^2\{n^{-1} + [(4v)^{-4v/(4v+1)} + (4v)^{1/(4v+1)}]B^{2/(4v+1)}V^{4v/(4v+1)}n^{-1-2\mu}\}.$$

We will report a simulation study for to examine our approximations and a plug-in estimation of $h_{n0}$.

**Acknowledgements:** This work has been partially supported by the Xunta de Galicia (Spain) and the DGES (Spain) under research projects XUGA 10503A98 and PB98-0182-C02-01, respectively.

## References

Aneiros, G. and Quintela, A, (1999). Modified cross-validation in semiparametric regression models with dependent errors, *In revision.*

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Nonparametric estimates of the relation between weather and electricity sales, *Journal of the American Statistical Association,* **81,** 310-320.

Gao, J (1995). Asymptotic theory for partly linear models, *Communications in Statistics-Theory and Methods,* **24,** 1985-2009.

Gasser, T. and Müller, H.G. (1984). Estimating regression functions and their derivatives by the kernel method, *Scand. J. Statist*, **11**, 171-185.

Linton, O. (1995). Second order approximation in the partially linear regression model, *Econometrica,* **63,** 1079-1112.

Speckman, P. (1988). Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society ser. B*, **50,** 413-436.

# Model Choice in the Bayesian Framework

Ursula Berger[1]

[1] Institut für Medizinische Statistik und Epidemiology, Technische Universität München, D- 81675 München, Germany

**Abstract:** In this work we investigate different approaches for Bayesian model selection based on the Bayes factor-theory and the Bayesian deviance. We discuss the different criteria and draw parallels between them. In an application to Breast Cancer data we point out their performance when used to discriminate Bayesian survival models with dynamic predictors.

**Keywords:** Bayes Factors, Posterior of the Likelihood, Bayesian deviance, Bayesian Hierarchical Models, Survival Data

## 1 Introduction

Markov Chain Monte Carlo (MCMC) methods and the increasing computational power have made Bayesian modelling popular in the last decade, since they allow now the calculation of flexible, high dimensional models mirroring complex structures. While these models often result in a better fit, their increased dimensionality makes interpretation harder. In application one is therefore often interested to keep the model as parsimonious as possible and additional complexity must be justified by a sufficient gain in the fit. This demands for a measure which allows for the comparison of competing Bayesian models of different dimensions.

## 2 Methods for Bayesian Model Comparison

Within the classical (frequentist) modelling framework, comparison of a set of $m$ models $M_j$, $j \in 1, \ldots, m$, typically takes place using the likelihood ratio respectively the deviance statistic, which is penalised by some function of the model dimension. In the Bayesian framework, additionally the specification of the prior densities $P_j(\theta_j)$ for the unknown parameter vector $\theta_j$, as well as the prior model probabilities $\pi_j$ have to be taken into account. Moreover, in complex (hierarchical) models the effective dimension is not clearly defined. The first suggestion arising for Bayesian model choice between a model $M_1$ and a model $M_2$ is the consideration of the model posterior ratio

$$\frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})} = \frac{L_1(\mathbf{y}|M_1)}{L_2(\mathbf{y}|M_2)} \frac{\pi_1}{\pi_2}$$

where $\pi_1/\pi_2$ is the prior odds and $B = L_1(.)/L_2(.)$ is known as the *Bayes factor*, describing the weight of evidence provided by the data for model $M_1$ over model $M_2$. The component $L_j(.)$ is thereby the marginal probability of the data $\mathbf{y}$, with

$$L_j(\mathbf{y}|M_j) = \int L_j(\mathbf{y}|\theta_j, M_j)P_j(\theta_j)d\theta_j \,,$$

which is also thought of as the "prior mean" of the likelihood $L_j(\mathbf{y}|\theta_j, M_j)$. Often, model diagnostics is restricted to the Bayes factor only, when in lack of information it is assumed, a priori, that all models are equally likely ($\pi_1 = \ldots = \pi_m = 1/m$). However, the analytical calculation of $L_j(.)$, i.e. the integration over the parameter priors $P_j(\theta_j)$, is quite difficult, if not impossible, for complex models. Even the different MCMC methods, which have been proposed to simulate or integrate $L_j(.)$, are not straightforward and typically demand considerable effort in implementation and computing time (Han and Carlin, 2000). In addition, the Bayes factor has frequently been criticized due to its distinct sensitivity on the prior specifications of any of the parameters, which leads to the "Lindley Paradox" when non-informative parameter-priors are used (Lindley, 1957).
Aitkin (1991) proposes a *posterior Bayes factor* $B = L_1^A(.)/L_2^A(.)$ based on the ratio of the "posterior mean" of the likelihood

$$L_j^A(\mathbf{y}|M_j) = \int L_j(\mathbf{y}|\theta_j, M_j)P_j(\theta_j|\mathbf{y})d\theta_j \,,$$

in order to overcome these problems. This idea goes along with Dempster's suggestion to plot representations of the posterior distribution of the log-likelihood under each competing model (Dempster 1997), respectively the posteriori density of the deviance $D(\theta) = -2\ln L(\mathbf{y}|\theta, M) + \text{const}$. As a one-point measure for model rating, the posterior mean $\overline{D(\theta)}$ can again be considered. Since the likelihood and the deviance are simple functions of $\theta$, their posterior means have the same inferential status as the posterior mean of $\theta$. In particular one finds both as a by-product of the MCMC-simulation. Recently Spiegelhalter et al. (1998) contributed a further approach by picking up Dempster's suggestion and taking a Bayesian look at the classical information criteria. They introduced the Bayesian deviance information criterion (DIC):

$$DIC(\theta) = \overline{D(\theta)} + df_D \,,$$

where the measure of the fit is again the posterior mean of the deviance $\overline{D(\theta)}$, but now penalized by the effective model complexity $df_D$. They give an asymptotic justification for deriving the effective number of parameters by $df_D = \overline{D(\theta)} - D(\bar{\theta}) = \mathrm{E}_{\theta|y}[D(\theta)] - D(\mathrm{E}_{\theta|y}[\theta])$. This allows to transform the DIC into Akaike criterion-structure by $DIC(\theta) = D(\bar{\theta}) + 2df_D$, where $D(\bar{\theta})$ can be seen as the deviance of the "final model" being penalized by 2-times the effective number of parameters.

## 3 Application to Breast Cancer data

TABLE 1. Model rating for the breast cancer study

| Model | | | | | $D(\bar{\theta})$ | DIC | $\overline{D(\theta)}$ | $\frac{L_j^A}{L_{M*}^A}$ |
|---|---|---|---|---|---|---|---|---|
| | tumor | hormo | T1 | KI-67 | | | | |
| **Static Solutions** | | | | | | | | |
| | - | - | - | - | 255.49 | 260.47 | 257.98 | 2e-7 |
| | $\beta$ | $\beta$ | $\beta$ | $\beta$ | 236.98 | 250.00 | 243.49 | 6e-4 |
| | $\beta(t)$ | $\beta(t)$ | $\beta(t)$ | $\beta(t)$ | 219.28 | 244.82 | 232.05 | 1.01 |
| **Selected Model M*** | | | | | | | | |
| | $\beta$ | $\beta(t)$ | $\beta(t)$ | $\beta$ | 220.35 | 240.83 | 230.59 | 1.00 |
| **M* with Grading** | | | | | | | | |
| $\beta$ | $\beta$ | $\beta(t)$ | $\beta(t)$ | $\beta$ | 219.17 | 241.26 | 230.21 | 1.43 |
| $\beta(t)$ | $\beta$ | $\beta(t)$ | $\beta(t)$ | $\beta$ | 219.09 | 243.59 | 231.34 | 1.11 |

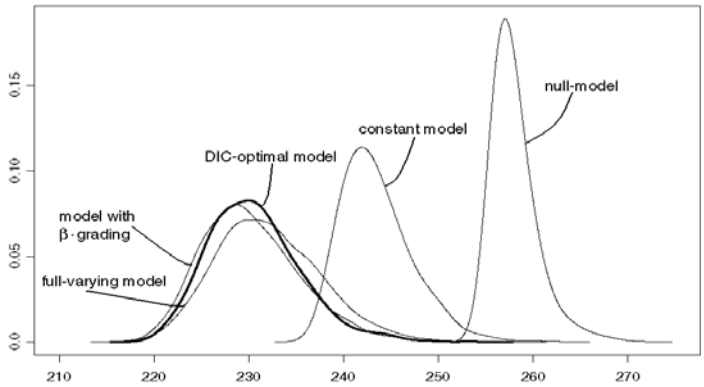



FIGURE 1. The posterior density of the deviance for different models

We use the different model selection criteria in a study of the survival of node-negative breast cancer patients, where it is of great interest to identify high-risk patients as early as possible for further systematic adjuvant therapy. For 97 postoperative patients, tumor size, hormone receptor state, the proliferation marker KI-67 and a newly detected protein T1 are measured

in the tumor tissue. We estimate survival models with time-constant and time-varying effects using a MCMC-algorithm for hierarchical models with state space structure. The final aim is, to decide upon an optimal predictor, where only those covariates are modeled in a complex fashion, which show clear evidence for this.

Tabel 1 shows the results for six selected models, where the first three models represent static solutions, either containing no covariate or assuming time-constant resp. time-varying effects for all covariates. Model M* was selected in a stepwise algorithm optimizing the DIC. It also shows the smallest posterior mean $\overline{D(\theta)}$ within the models considered in this step.

In a next step, tumor grading was additionally included in the model. Since it is a staging of the tumor with respect to its proliferation characteristics, we expect it only to increase the model complexity while contributing no extra information in addition to KI-67. The DIC supported this assumption, while the posterior mean $\overline{D(\theta)}$ doesn't sufficiently adjust for complexity and hence doesn't increase. The model deviance $D(\bar{\theta})$, as expected, generally declines with increasing dimension. Also the posterior Bayes factor, which is calculated over model M*, seems to give too much weight to more complex models.

Figure 1 shows the posterior density of the deviance $D(\theta)$ for the three static solutions, model M* and the model including tumor grading with a time-constant effect. It shows that all models considering time-variation are clearly superior to the time-constant model resp. the null-model. However, it also points out the difficulties of distinction of different models with close fit.

## References

Aitkin, M. (1991). Posterior Bayes Factor (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 111-142.

Dempster, A.P. (1997). The direct use of likelihood for significance testing (with discussion). *Statistics and Computing*, **7**, 247-252. (Reprint of 1973).

Han, C., Carlin, B. (2000). MCMC Methods for Computing Bayes Factors: A Competitive Review. *Research Report*, Division of Biostatistics, Uni. of Minnesota.

Lindley, D.V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

Spiegelhalter, D., Best, N., Carlin, B. (1998). Bayesian deviance, the effictive number of parameters, and the comparison of arbitrarily complex models. *Research Report 98-0009*, Division of Biostatistics, Uni. of Minnesota.

# Incorporating model selection uncertainty into statistical inference: a simple example

Cecilia Candolo[1], Anthony Davison[1] and Clarice Demétrio[2]

[1] Dep. of Mathematics, Swiss Federal Institute of Technology, 1015, Lausanne, Switzerland
[2] Dep. of Exact Sciencies, São Paulo University, 13418-900 Piracicaba, Brazil

**Abstract:** Buckland *et al.* (1997) discuss how model selection uncertainty may be incorporated into statistical inference. We apply their approach in a simple example of linear regression. Through a simulation study we analyse and compare the behaviour of the variance estimation and the bias. We conclude that the bootstrap gives a better variance estimate.

**Keywords:** Bootstrap; Linear Regression; Model Selection.

## 1 Introduction

Buckland *et al.* (1997) discuss how model selection uncertainty may be incorporated into statistical inference using weighting methods, where the weights are obtained from information criteria or by using the bootstrap. It is assumed that each fitted model provides an estimate of the parameter of interest $\theta$, common to all models. If each model $M_k$, $k = 1, \ldots, K$, provides a weight $w_k$, scaled so that $\sum w_k = 1$, and estimates $\hat{\theta}_k$ for $\theta$, the overall estimate of $\theta$ is taken to be

$$\hat{\theta} = \sum_k w_k \hat{\theta}_k. \tag{1}$$

The weights $w_k$ can be obtained using Akaike's Information Criterion, AIC, the Bayes Information Criterion, BIC, by the bootstrap, in which case $w_k$ is estimated by the proportion of resamples in which $M_k$ is identified as best . We use AIC, $I = -2 \log L + 2p$ ,where $L$ is the likelihood function evaluated by replacing parameters by their maximum likelihood estimates and $p$ the number of parameters, and this gives

$$W_k = \frac{\exp(-I_k/2)}{\sum_{i=1}^{K} \exp(-I_i/2)}, \qquad k = 1, \ldots, K.$$

To estimate the variance of (1) in the non-bootstrap case, Buckland *et al.* (1997) suppose that the weights $w_k$ are known, that the bias in estimation

of $\theta$ is zero and that the estimators $\hat{\theta}_k$ and $\hat{\theta}_l$, $k \neq l$, are perfectly correlated. They obtained

$$Var(\hat{\theta}) = \Big\{ \sum_k w_k \sqrt{Var(\hat{\theta}_k \mid \beta_k) + \beta_k^2} \Big\}^2, \tag{2}$$

which can be estimated by replacing the bias $\beta_k = E(\hat{\theta}_k) - \theta$ by $\hat{\beta}_k = \hat{\theta}_k - \hat{\theta}$, and using $\widehat{Var}(\hat{\theta}_k \mid \beta_k)$, found by usual inference methods. Our aim is to consider (1) in simple linear regression, to calculate its variance without making the above-mentioned assumptions, and to find its bias.

## 2 Simple linear regression

Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$, $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The aim is to predict $y$ for a value of $x$, $x_+$. We consider two estimators, $\hat{\theta}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_+$ and $\hat{\theta}_2 = \bar{y}$, corresponding to models $M_1$ and $M_2$. From (1) we get $\hat{\theta} = W_1\hat{\theta}_1 + W_2\hat{\theta}_2 = \bar{y} + W_1\hat{\beta}_1(x_+ - \bar{x})$, and the expectation, variance and bias of $\hat{\theta}$ are

$$E(\hat{\theta}) = E(\bar{y}) + (x_+ - \bar{x})E(W_1\hat{\beta}_1) = \beta_0 + \beta_1\bar{x} + (x_+ - \bar{x})E(W_1\hat{\beta}_1), \tag{3}$$

$$Var(\hat{\theta}) = Var(\bar{y}) + (x_+ - \bar{x})^2 Var(W_1\hat{\beta}_1) + 2(x_+ - \bar{x})Cov(\bar{y}, W_1\hat{\beta}_1), \tag{4}$$

$$E(\hat{\theta}) - \theta = \beta_1(\bar{x} - x_+) + E(W_1\hat{\beta}_1)(x_+ - \bar{x}). \tag{5}$$

We are concerned with $W_1\hat{\beta}_1$, and to derive its expectation and variance we write it as

$$W_1\hat{\beta}_1 \overset{\text{D}}{=} \left[ 1 + e^1 \left\{ 1 + \left\{ \frac{Z}{\sqrt{C/(n-2)}} + \delta \right\}^2 / (n-2) \right\}^{-\frac{n}{2}} \right]^{-1} [V^{1/2}(Z + \delta)],$$

where $Z \sim N(0,1)$, $C \sim \chi^2_{n-2}$, $\delta = \beta_1/V^{1/2}$ is a non-centrality parameter and $V$ is the variance of $\hat{\beta}_1$ . So, $W_1\hat{\beta}_1$ is a function of $Z$, $C$ and $\delta$, and its expectation and variance can be obtained by simulation of $Z$ and $C$. When $\delta$ increases, $W_1 \to 1$, and we can expect that $E(W_1\hat{\beta}_1) \to 1$, given that $E(\hat{\beta}_1) = \beta_1$ under the true model. And so we can expect that $E(\hat{\theta}) - \theta \to 0$.

### 2.1 Simulation and results

We conducted a series of Monte Carlo experiments based on data sets of size $n$ = 11, 21, 41, 101 and 201. The values of $x$ were equally spaced in the interval $[-1, 1]$, $\beta_0 = 1$ and $\beta_1 = 2$, the responses were generated as $\varepsilon \sim N(0,1)$ with $x+ = 1.1$. $St.Err.(\hat{\theta})$ and $E(\hat{\theta})$ were first calculated

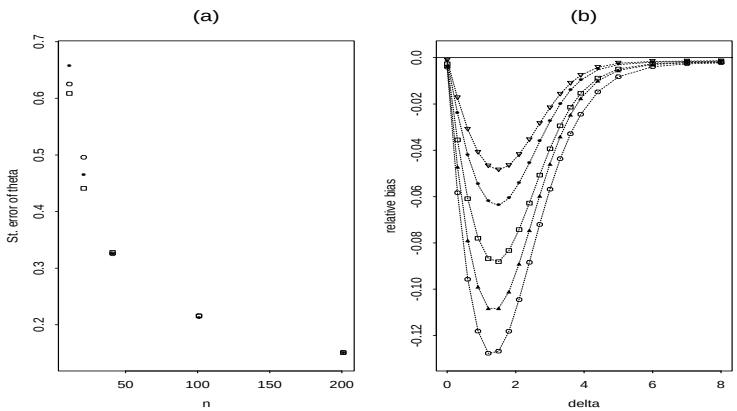



FIGURE 1. *(a) Standard errors of $\hat{\theta}$ calculated from (4) (•), average values from (2) (□) and average bootstrap values, resampling pairs, (○). (b) Relative bias for different sample sizes as a function of $\delta$, for $n = 11$ (○), $n = 21$ (▲), $n = 41$ (□), $n = 101$ (•) and $n = 201$ (▽).*

from (3) and (4), generating 10,000 values of $Z$ and $C$ for the computation of $E(W_1\hat{\beta}_1)$ and $St.Err.(W_1\hat{\beta}_1)$. These values agreed well with the values from 10,000 simulations of the previous model, for all sample sizes. Secondly the same quantities were calculated by bootstrapping, getting $\hat{\theta}$ in each resample, resampling pairs and resampling residuals from $M_1$, the true model, for $R = 999$ resamples. The results from resampling pairs and resampling residuals were similar. $St.Err.(\hat{\theta})$ were also calculated from (2), and for better comparison, we get the average values of $St.Err.(\hat{\theta})$ in 20 differents samples, for (2) and for bootstrap resampling pairs. Finally, we also performed a simulation study to verify the behaviour of $W_1$, $W_1\hat{\beta}_1$ and the bias when $\delta$ increases.

The results are sumarized in Figure 1. Plot (a) represent the values of $St.Err.(\hat{\theta})$ calculated from (4) and the means for the bootstrap (resampling pairs) and (2), as functions of $n$. Despite some small sample fluctuation, these values agree. We also plotted the log of these average variances against log n, and this confirmed our conclusion. In (b) the relative bias for the different sample sizes are plotted as a function of $\delta$. The relative bias decreases as $n$ increases and tends to zero as $\delta$ gets bigger.

We performed all theses calculations including a third model in $\hat{\theta}$, $\hat{\theta}_3 = \hat{\beta}'_0 + \hat{\beta}'_1 x_{1+} + \hat{\beta}_2 x_{2+}$, and the results were the same.

## 3 Application

We take data on the relationship of immunoglobulinG (IgG) with age, in children from 6 months to 6 years. The aim was to establish reference

centiles for the serum concentration of certain immunoglobulins in children. This study was discussed by Royston and Altman (1994), who applied regression using fractional polynomials. They take the square root of IgG as the response $Y$, which appears to eliminate the skewness of the data, and compared fractional polynomials with polynomial regression. We take 3 fractional polynomials and 2 polynomials to estimate $\theta$, the prediction of $Y$ at 6.2 years. The fractional polynomials were $M_1$: $y = \beta_0 + \beta_1 \log x$, $M_2$: $y = \beta_0 + \beta_1 x^{-2} + \beta_2 x^2$ and $M_3$: $y = \beta_0 + \beta_1 x^{1/2} + \beta_2 x$. $M_4$ and $M_5$ correspond to cubic and a quartic polynomials. The more parsimonious model was $M_2$. We calculate $\hat{\theta}$ in two situations: for the whole data set of 298 observations and for a subsample of 53 observations. The data, fitted models and $\hat{\theta}$ for these two situations are in Figure 2. For the first we get $St.Err.(\hat{\theta})$ 0.3053 using (2) and 0.2449 by resampling pairs. For the second, we get 0.4534 and 0.6118. The values for the whole sample are close, but using bootstrap we made no suppositions and in addition we can get confidence intervals and other properties for all quantities involved in $\hat{\theta}$.

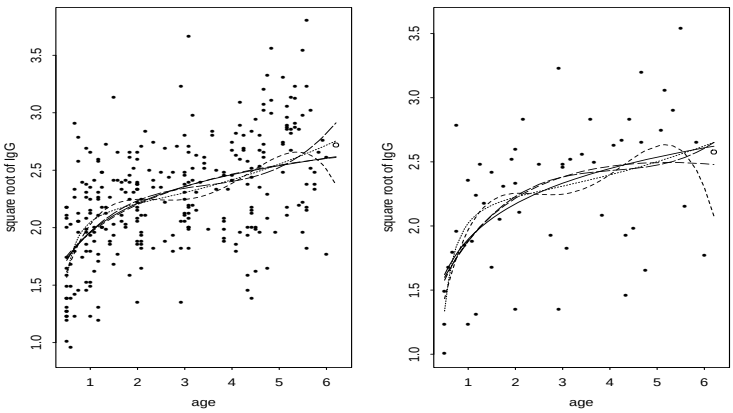



FIGURE 2. *Fits for IgG data (left whole data set, right subsample): $M_1$ (solid); $M_2$ (dotted); $M_3$ (small dashes); $M_4$ (medium dashes); $M_5$ (large dashes); $\hat{\theta}$ ($\circ$).*

## References

Buckland, S. T., Burnham, K. P. and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, **53**, 603-618.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Appl. Statist.* **43**, 429-467.

# Residential Site Choice by Ethnicity: Demand Side Estimation

Mercedes G. Escribano[1]

[1] Ph.D. student at University of Chicago e-mail: mgarciae@midway.uchicago.edu

**Abstract:** Efforts to estimate the demand side for residential site have been numerous, especially since Rosen's analysis of implicit markets in 1974, but non satisfactory due to the difficulty any of these approaches has to face when trying to deal with heterogeneous consumers. This paper proposes an alternative approach that consists of applying a discrete choice model to the demand for residential site choice. In particular, I estimate the rental housing demand of each ethnic group -Blacks, Whites, Asians and Hispanics- for a residential site that is characterized by a bundle of housing dwelling and site traits. The estimation coefficients show that each ethnic group differs in the valuation of housing traits such as the racial composition of the neighborhood.

## 1 Introduction

Efforts to estimate the demand side for residential site have been numerous, especially since Rosen's analysis of implicit markets in 1974. According to Rosen's analysis of implicit markets, the observed price of a housing unit depends on the implicit prices of its residential -site and dwelling-attributes since market housing prices are the result of matching consumer's willingness to pay for traits and producers' minimum acceptance price for those traits. Then, the hedonic price function is the joint envelope of the family of value functions and the family of offer functions.

The hedonic price function would directly reveal the demand side for characteristics if consumers were homogeneus. As Rosen states, only if consumers were identical, the family of value functions would collapse to a unique function that coincides with the hedonic function. If this were the case, the observed price differences would be exactly equalizing across consumers and the hedonic price function would identify the structure of the demand.However, households are heterogeneous; different age, income, race and family size make tastes to differ across individuals. Therefore, the estimation of the demand for residential characteristics using the hedonic approach becomes difficult if not impossible. Examples of these efforts are Linneman (1982) and Bloomquist and Berger (1988).

Because of the difficulties to estimate the demand for residential characteristics by regressing the hedonic price function on these traits, I propose

an alternative approach that consists of applying, to the housing market, a discrete choice model of product differentiation (Berry, Levinshon, Pakes (1995); Nevo (1997); and Berry (1994)). In particular, using the discrete choice technique I estimate for the renter occupied housing market, the demand of each ethnic group for a residential site as a function of its dwelling and site traits. Therefore, this approach directly reveals the preferences of different types of households for certain residential characteristics.
To deal with this objective, this paper is organized as follows. After the presentation of the discrete choice approach for modeling demand of each ethnicity for residential site choice, section III describes the data set used in the estimation and the main results.

## 2 Demand side for residential choice

In this section, I follow Berry, Levinshon and Pakes 1995; Nevo 1997.
Individual household behavior is modeled with a discrete choice model in which the products -housing units– are differentiated geographically and physically. That is, housing units are characterized by site traits like ethnic composition of the neighborhood, and dwelling traits like number of rooms or age of buildings.
Let $h = 1, ....., H$ be housing units in the area A. Then, at time t, the indirect utility that household i would receive if he decided to rent $h$ is:

$$U_{it}^h = \phi_i(X_t^h) - \alpha_i R_{it}^h + \xi_t^h + \epsilon_{it}^h$$

where $X_t^h$ is $D \times 1$ vector of observable (by the econometrician) dwelling and site characteristics of $h$ at time $t$; $R_{it}^h$ is the rental payment that household $i$ faces for $h$ at time $t$; $\xi_t^h$ is the mean valuation of unobservable (by the econometrician) characteristics of $h$ at time $t$; and $\epsilon_{it}^h$ captures household heterogeneity in the valuation of unobservable characteristics of $h$ at time $t$.
In what follows, I omit the time subscript. Further, I assume that each household resides at most in one housing unit; housing and site characteristics are publicly known; $\epsilon_i^h$ iid over (i ,h) with a Type I extreme value distribution; rental payment for $h$ at $t$ is the same for all type of households, i.e., there is no price discrimination; and $\phi_i(X_t^h)$ is linear in the characteristics. Under these assumptions, household's i indirect utility from residing at $h$ takes the form:

$$U_i^h = X^h \beta_i - \alpha_i R^h + \xi^h + \epsilon_i^h$$

Household $i$ decides to reside at housing unit $h$ if and only if the utility that he would obtain at $h$ is at least greater than the utility he would obtain at any other housing unit.

The vector of household's $i$ tastes for housing characteristics, $(\beta_i, \alpha_i)$, can be expressed as

$$(\beta_i, \alpha_i) = (\beta_e, \alpha_e) + \Pi\kappa_i + \Sigma\nu_i$$

where the subscript $e$ indicates the race of household $i$; $(\beta_e, \alpha_e)$, is the vector of mean tastes for housing traits of the households of ethnicity $e$; $\Pi$ is a $(D+1) \times d$ matrix that allows to measure the effect of demographics different from race on the valuation of characteristics; $\kappa_i$ is the $dx1$ vector of household's $i$ demographics other than race; $\Sigma$ allows for heterogeneity in tastes not related with demographics; and $\nu_i$ $N(0, I_{D+1})$.
Then, the indirect utility equation can be written as

$$U_i^h = \delta_e^h + \mu_i^h + \epsilon_i^h$$

where $\delta_e^h = X^h\beta_e - \alpha_e R^h + \xi^h$, that is, it is the mean utility associated with housing unit h for households of race $e$; and $\mu_i^h = [X^h R^h] \times [\Pi\kappa_i + \Sigma\nu_i]$. Further, assuming all relevant consumer heterogeneity for housing decision is reduced to the demographic race, the term $\mu_i^h$ would equal zero.
Under these assumptions, the share of the housing unit $h$ rented by a household of race $e$ in the renter market of housing units, called $S_e^h$, can be expressed as

$$S_e^h = \frac{exp(\delta_e^h)}{1 + \sum_{k=1}^{H}(\delta_e^k)}$$

The market of housing units for rent considered corresponds to an area $B$, with $A \subset B$. Therefore, this expression takes into account that each household $i$ of race $e$ could decide not to rent any of the housing units at $A$ but still rent another housing unit in the market. This is the outside alternative.
At the true values of $\delta_e$ the share predicted by the model for housing unit $h$, $S_e^h$, must equal the observed shares for $h$, $\hat{S}_e^h$. Following Berry (1994), the $\delta_e$ satisfying this requirement is unique and it can be computed from the observed shares as

$$\delta_e^h = ln(\hat{S}_e^h) - ln(\hat{S}_e^0)$$

where $\hat{S}_e^0$ is the share in the market of housing units for rent corresponding to the outside alternative for race $e$.
Using the computed mean utility associated with housing unit $h$ for households of race $e$, $\delta_e^h$, it is possible to estimate the vector of unknown parameters, $(\beta_e, \alpha_e)$, with the equation

$$\delta_e^h = X^h\beta_e - \alpha_e R^h + \xi^h$$

that is linear in housing characteristics.

## 3 Data and results

The data set used to estimate the model presented here has been obtained from the Summary Tape File 1A, Summary Tape File 3A, and Summary Tape File 3C1 of the1990 Census of Population and Housing. The area B considered, that is, the market of housing units for rent, matches Chicago-Gary-Lake County Metropolitan Statistical Area. Area A corresponds to Community Areas 1 to 76 of the city of Chicago. The level of extraction of the files to obtain area A data is Census tracts.

The variables used in the estimation are the number of years of the housing unit; number of bedrooms; proportion of Blacks, Hipanics, Asians and Whites in the neighborhood; a dummy for proximity to the lake; distance to the city center; rent per bedroom; a dummy for North location to the North; and the mean utility households of race e would obtain if rented that housing unit. To control for the endogeneity problem, I instrument the variable rent per bedroom with: a constant, the number of bedrooms of owned occupied housing units in that neighborhood, the median household income in the neighborhood, and the mean number of rooms of owned occupied housing units in the neighborhood.

The most remarkable findings of a 2SLS estimation for the demand for housing of each of the four ethnic groups are the following: (i) Ethnic groups differ in the valuation, in magnitude and sign, of dwelling and site characteristics. (ii) Rent per bedroom is only statistically significantly for Asians and Hispanics. (iii) As the number of bedrooms of a housing unit increases, it is more difficult to rent it. (iv) Each ethnic group prefers renting a housing unit in areas where other households of the same race live. (v) The demand for renting a housing unit by Asians, Hispanics, or White households decreases as the proportion of Blacks in the neighborhood increases. (vi) The proportion of Whites in an area decreases the demand for renting a residential unit by Blacks in that location. (vii) There is a remarkable preference of Hispanics and Asians for the proportion of Hispanics and the proportion of Asians, respectively, living in a neighborhood.

## References

Berry, S. (1994). Estimating discrete-choice models of product differentiation. *RAND Journal of Economics*, **25**, 242–262.

Berry, S., Levinshon, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, **64**, 841–890.

Diamond, D. B. and Tolley, G. S. (1982). *The Economic Role of Urban Amenities*. New York: Academic Press.

Rosen, S. (1974). Hedonic Prices and Implicit Markets. *Journal of Political Economy*, **82**, 34–55.

# A modified cross-validacion bandwidth to estimate the hazard function under dependence

Graciela Estévez[1], Alejandro Quintela[1] and Philippe Vieu[2]

[1] Universidad de A Coruña, A Coruña. Spain. graci@udc.es
[2] Université Paul Sabatier, Toulouse. France. vieu@cict.fr

**Abstract:** In this paper, we estimate the hazard function through a kernel estimator, and we make use a modification of the cross-validation bandwidth studied by Estévez and Quintela (1999), under $\alpha - mixing$ conditions. The modified bandwidth cross-validation is also optimal but it make estimations better. Some simulations and a practical application to real data are also shown.

**Keywords:** Kernel estimation; Hazard function; Bandwidth selection; Modified Cross-validation; Mixing processes.

## 1 Nonparametric estimation of hazard function

Let $X$ be a nonnegative real random variable with absolutely continuous distribution function $F$, and probability density function $f$. There are others ways for to study the distribution of $X$, as the failure rate function $r$. This function, also called "hazard rate function" or "risk function", is defined by

$$r(x) = \lim_{\Delta_x \to 0} \frac{P(x \leq X < x + \Delta_x / X \geq x)}{\Delta_x},$$

that by the definition of conditional probability density it can be written

$$r(x) = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{\bar{F}(x)},$$

considered when $\bar{F}(x) > 0$, that is, $r$ is defined in the set $S = \{x \in R / \bar{F}(x) > 0\}$.
Thus, if $X$ measures a failure time, $r(x)\Delta_x$ can be interpreted as the approximate probability of one object "fails" in the time interval $[x, x + \Delta_x)$, given the object has survived to time $x$, i.e. the instantaneous probability of failure at $x$. In medicine it has been study mainly for censored and/or truncate dates (see, e.g. Lee (1992)). In a seismology context, $r(x)$ might be thought of as the instantaneous risk of the occurrence of an earthquake

in the moment $x$, known the last earthquake has happened in the moment 0 (Rice and Rosenblatt (1976)). In these situations, as well as in others practical situations, a problem of considerable interest is the estimation of $r$ from a random sample $X_1, ..., X_n$ of $F$.

A typical form to estimate this function consists in to specify a parametric form of a known distribution, such as a Weibull or a Gamma, and to estimate the parameters of the supposed distribution (Cox and Oakes (1984)). However, at many times, there not exists sufficient information to precise the distribution of the random variable $X$ of interest. In such situations, it is certainly desirable to have nonparametric tools to estimate such function, without having to restrict us to some particular class of distributions. Several nonparametric methods for estimating $r$ have been proposed in the literature; many of these methods are based on the assumption that $X_1, \ldots, X_n$ are i.i.d. random variables. See, for example, Watson and Leadbetter (1964 a and b), Ahmad (1976), Singpurwalla and Wong (1983) and Hassani, Sarda and Vieu (1986).

However, in certain cases it is much more realistic to suppose that the random variables $X_1, ..., X_n$ are dependent, although this dependence weakens for groups of r.v.'s far apart. This may be the case, for instance, for the time intervals between occurrence of earthquakes (Rice and Rosenblatt (1976), Udias and Rice (1975)). By this reason, we will assume that the underlying r.v.'s come from a strictly stationary process and satisfy some mode of dependence which is reasonable from a practical point of view. In this context, we also find several nonparametric estimates of hazard function (see Györfi, Härdle, Sarda and Vieu (1990), Sarda and Vieu (1989), Roussas (1990) and their references).

A straightforward way to build a nonparametric estimator of the function $r$, is to take the ratio of nonparametric estimators of density $f$ and survival function $1 - F$. Because the best way to estimates these functions is with kernel type estimators (Silverman (1986) and Sarda (1991)), we will also consider the kernel estimator of the hazard rate function, introduced by Watson and Leadbetter (1964a and b) in a context of independent data. It is defined by

$$r_h(x) = \frac{f_h(x)}{1 - F_h(x)}, \tag{1}$$

where $f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$ is the Parzen-Rosenblatt estimator of $f(x)$, and $F_h(x) = \int_{-\infty}^{x} f_h(t)dt$ is the kernel estimator of $F(x)$. As well as, $K$ is a kernel function, $H(x) = \int_{-\infty}^{x} K(u)du$ is its distribution function and $h = h(n) \in R^+$ is the smoothing parameter, or bandwidth. Optimality properties of the estimator (1) have been proved in Estévez and Quintela (1999).

On the other hand, as in the case of kernel estimators of other unknown function, the choice of the smoothing parameter $h$ is crucial to the effective

performance of the estimator. If the bandwidth is too small, there is too much variance in the sense that features which belong only to the particular data set and not to the true hazard rate may be seen in the estimates. If the bandwidth is too large, there is too much bias in the sense that features of the true hazard rate are smoothed away. The problem of automatic choice of smoothing parameters in kernel nonparametric estimation has been widely studied in the literature. In the whole of proposed methods one must to emphasize the cross-validation procedure. This procedure, introduced by Bowman and Rudemo at the beginning of the eighties (Rudemo (1982), Bowman (1984)), it has been suitably modified depending on the function to estimate (see, e.g., Györfi, Härdle, Sarda and Vieu (1990) and their references), the type of sought window, local or global (Mielniczuk, Sarda and Vieu (1989)), and the independence or dependency conditions between the observations (Hart and Vieu (1990) for density, Chu and Marron (1991) for regression and Estévez and Quintela (1999) for hazard function estimations).

In this presentation, we estimate the hazard function through (1) and we make use a modification of the cross-validation bandwidth studied by Estévez and Quintela (1999). The modified bandwidth cross-validation is also optimal but it make estimations better. Our results will be given in the setting of $\alpha - mixing$ variables, which is the less restrictive among the mixing conditions.

In addition, we will show some simulations that exhibit our results, and we will present an application to a set of real data. We will deal with observations of the time intervals between earthquakes in certain geographical areas.

**Acknowledgements:** This research was financed by the Xunta de Galicia (Spain) under research Project XUGA 10503A98.

## References

Ahmad, I.A. (1976). *Uniform strong convergence of the generalized failure rate estimate.* Bull. Math. Statist., 17, 77-84.

Bowman, A. (1984). *An alternative method of cross-validation for the smoothing of density estimates.* Biometrika, 71, 353-360.

Chu, C.K. and Marron, J.S. (1991). *Comparison of two bandwidth selectors with dependent errors.* Ann. Statist., 19, 1906-1918.

Cox, D.R. and Oakes, D. (1984). *Analysis of survival data.* Chapman an Hall.

Estévez, G. and Quintela, A. (1999). *Nonparametric estimation of the hazard function under dependence conditions.* Communications in Statistics: Theory and Methods, vol. 28, 10, 2297-2331.

Györfi, L., Härdle, W., Sarda, P. and Vieu, P. (1990). *Nonparametric Curve Estimation from Time Series.* Lecture Notes in Statistics, vol. 60, Springer-Verlag.

Hart, J. and Vieu, P. (1990). *Data-driven bandwidth choice for density estimation based on dependent data.* The Annals of Statistics, 18, 873-890.

Hassani, S., Sarda, P. and Vieu, P. (1986). *Approche non paramétrique en théorie de la fiabilité.* Revue de Statistiques Appliquées, vol XXXV, n.4.

Lee, E.T. (1992). *Statistical methods for survival data analysis.* Wiley Series in Probability and Mathematical Statistics.

Mielniczuk, J., Sarda, P. and Vieu, P. (1989). *Local data driven bandwidth choice for density estimation.* Journal of Statistical Planning and Inference, 23, 53-69.

Rice, J. and Rosenblatt, M. (1976) *Estimation of the log survivor function and hazard function.* Sankhyä, A, 38, 60-78.

Roussas, G. (1990). *Asymptotic normality of the kernel estimate under dependence conditions: Application to hazard rate.* Journal of Statistical Planning and Inference, 25, 81-104.

Rudemo, M. (1982). *Empirical choice of histograms and kernel density estimates.* Scand. J. Statistics, 9, 65-78.

Sarda, P. (1991). *Estimating smooth distribution function.* Proceedings of NATO advanced study on Nonparametric Functional Estimation and Related Topics, 271-283. Kluwer Academic Publishers.

Sarda, P. and Vieu, P. (1989). *Empirical distribution function for mixing random variables. Applications in nonparametric hazard estimation..* Statistics, 20, 559-571.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall.

Singpurwalla, N.D. and Wong, M.Y. (1983). *Estimation of the failure rate - A survey of nonparametric methods. Part I: Non-Bayesian methods.* Commun. Statist. Theory Methods, 12, 559-588.

Udias, A. and Rice, J. (1975). *Statistical analysis of microearthquakes activity near San Andreas Geophysical Observatory.* Hollister, California. Bulletin of the Seismological Society of America, 65, 809-828.

Watson, G.S. and Leadbetter M.R. (1964a(b)). *Hazard Analysis I(II).* Biometrika, 51, 175-184 (Sankhyä, A, 26, 110-116).

# Modelling Mortality as a Function of Time in a Clustered Data Bioassay

Sílvia M. de Freitas[1,2], John P. Hinde[1,3] and Clarice G. B. Demétrio[4]

[1] School of Mathematical Sciences, Laver Building, University of Exeter, Exeter, EX4 4QE, UK
[2] email: S.M.D.Freitas@exeter.ac.uk
[3] email: J.P.Hinde@exeter.ac.uk
[4] Departamento de Ciências Exatas - ESALQ/USP, Caixa Postal 9, 13418-900, Piracicaba, SP, Brasil
email: clarice@carpa.ciagri.usp.br

**Abstract:** Here we use the framework of generalized linear models and generalized estimating equations as alternative approaches to the problem of fitting cumulative mortality as a function of time to clustered data. We consider the Dirichlet-multinomial and a random effect model to allow for the extra-multinomial variation (overdispersion) arising from the use of groups (clusters) as experimental units.

**Keywords:** Overdispersion, Dirichlet-Multinomial, Random Effect Model, QL, GEE.

## 1 Introduction

The problem of fitting cumulative mortality as a function of time to clustered data involves modelling the multinomial response over time and possible approaches include survival analysis (Petkau and Sitter, 1989) and ordinal generalized linear models (McCullagh, 1980; Glonek and McCullagh, 1995). An additional aspect here is the possibility of extra-multinomial variation (overdispersion) arising from the use of groups (clusters) as the experimental units. A consequence of failing to take overdispersion into account is not only underestimation of the standard errors of estimated regression coefficients but also those of the lethal time $LT_p$, giving incorrect significance of treatment effects (Hinde and Demétrio, 1998). The use of Liang and Zeger's (1986) generalized estimating equations (GEE) as an extension of Quasi-Likelihood (QL) (McCullagh and Nelder, 1989) has been considered as an alternative approach to the problem of fitting a generalized linear model (GLM) to clustered data.

In this paper we present and analyse data from a biological control assay conducted at ESALQ/USP, in Piracicaba, São Paulo, Brazil. Different

*Beauveria bassiana* (a fungus) isolates ($K = 142$) were used as a microbial control for *H. tenuis* (a termite) to study the pathogenicity and virulence of the fungus isolates. These isolates were applied to groups (clusters) of $n = 30$ termites, with five replications ($r$) per isolate, and the mortality in the groups was measured daily for a period of eight days ($j$), resulting in 710 multinomial observations.

## 2 Alternative Models

The frameworks of GLM and GEE are used to model the cumulative mortality as a function of time. For the systematic part we use a regression model with a logit link and day as an explanatory variable with a linear predictor such as the linear trend $\boldsymbol{\eta}_{ir} = \beta_{1i} + \beta_{2i} t_j$, where $\beta_{1i}$ is the effect of $i^{th}$ isolate ($i = 1, \ldots, K$), $\beta_{2i}$ is the coefficient associated to the linear effect of day on the $i^{th}$ isolate, and $t_j$ is a quantitative variable for day, ($j = 1, \ldots, D$). To model the extra-multinomial variation (overdispersion) we consider two-stage-models for the response: a Dirichlet-multinomial (DM) and a random effect model (REM) (O'Hara Hines and Lawless, 1993).

### Multinomial Model

Suppose that for a single isolate $i$, $Y_{irj}$ is the number of insects dying on day $j$ for the $r^{th}$ replication. Let $R_{irj}$ = cumulative proportion of insects dead by day $j$. If we assume that $Y_{irj} = n_i(R_{irj} - R_{irj-1})$ and $\{Y_{ir1}, \ldots, Y_{irD}\}$ is Multinomial, from the moments of the multinomial distribution we have that the mean vector and covariance matrix of $\boldsymbol{R}_{ir}$ are given by

$$E(\boldsymbol{R}_{ir}) = \boldsymbol{p}_{ir} \qquad \text{and} \qquad Var(\boldsymbol{R}_{ir}) = \boldsymbol{V}(\boldsymbol{p}_{ir}) \tag{1}$$

where the matrix $\boldsymbol{V}(\boldsymbol{p}_{ir})$ has elements $v_{jj'} = v_{j'j} = p_{irj}(1 - p_{irj'})/n_i$, $(1 \leq j \leq j' \leq D)$.

### Dirichlet-multinomial Model (DM)

The DM model assumes that, as a first stage, the multinomial probability vector, $\{\pi_{ir1}, \ldots, \pi_{irD}\}$ has a Dirichlet distribution, and, as a second conditional stage, that $\{\boldsymbol{Y}_{ir} \,|\, \boldsymbol{\pi}_{ir}\}$ has a Multinomial distribution. For the unconditional cell frequencies $\{\boldsymbol{Y}_{ir}\}$ we have a Dirichlet-multinomial distribution. The resulting marginal distribution for the cumulative proportion response vector has mean vector given by

$$E(\boldsymbol{R}_{ir}) = E[E(\boldsymbol{R}_{ir} \,|\, \boldsymbol{p}_{ir})] = E(\boldsymbol{p}_{ir}) = \boldsymbol{\gamma}_{ir}$$

and variance function

$$\begin{aligned} Var(\boldsymbol{R}_{ir}) &= E[Var(\boldsymbol{R}_{ir} \,|\, \boldsymbol{p}_{ir})] + Var[E(\boldsymbol{R}_{ir} \,|\, \boldsymbol{p}_{ir})] \\ &= [1 + \rho_i(n_i - 1)]\, \boldsymbol{V}(\boldsymbol{\gamma}_{ir}), \end{aligned} \tag{2}$$

which is inflated by a factor of $[1 + \rho_i(n_i - 1)]$ compared to that for the standard multinomial model (1).

### Random Effects Model (REM)

Incorporating an additive, normally distributed, random effect in the linear predictor gives a random location shift in the effect of each isolate for each replicate and we have

$$g(p_{irj}) = \eta_{irj} + \varepsilon_{ir} = \beta_{1i} + \beta_{2i} t_j + \varepsilon_{ir},$$

where $\varepsilon_{ir} \sim N(0, \sigma_i^2)$. Writing $p_{irj} = g^{-1}(\eta_{irj} + \varepsilon_{ir}) = h(\eta_{irj} + \varepsilon_{ir})$ and using a second-order Taylor series expansion in the random effect $\varepsilon_{ir}$ for the cumulative multinomial probability vector $\boldsymbol{p}_{ir} = h(\boldsymbol{\eta}_{ir} + \boldsymbol{\epsilon}_{ir})$, around the linear predictor $\boldsymbol{\eta}_{ir}$, we obtain

$$E(\boldsymbol{p}_{ir}) \approx h(\boldsymbol{\eta}_{ir}) = \boldsymbol{\gamma}_{ir} \qquad \text{and} \qquad Var(\boldsymbol{p}_{ir}) \approx h^{'}(\boldsymbol{\eta}_{ir})[h^{'}(\boldsymbol{\eta}_{ir})]^T \sigma_i^2,$$

where $h^{'}(\boldsymbol{\eta}_{ir}) = \partial h(\boldsymbol{\eta}_{ir}) / \partial \boldsymbol{\eta}_{ir}$. The resulting marginal distribution for the cumulative response vector has variance function given by

$$Var(\boldsymbol{R}_{ir}) \approx \boldsymbol{V}(\boldsymbol{\gamma}_{ir}) + (1 - 1/n_i)\, \sigma_i^2\, h^{'}(\boldsymbol{\eta}_{ir})[h^{'}(\boldsymbol{\eta}_{ir})]^T. \tag{3}$$

For each of these we consider using both a single constant overdispersion term and models in which the overdispersion is allowed to vary over isolates. These approaches can be considered as extensions of Williams (1982) methods for the binomial case.

## 3 Estimation

Because of the equal sample size in this bioassay we used a simple constant overdispersion factor with standard quasi-likelihood techniques to estimate the vector of regression coefficients and moment methods to estimate the overdispersion parameter. The quasi-likelihood or generalized estimating equations for the vector of regression coeficients may be written as

$$\sum_{r=1}^{5} \boldsymbol{X}_{ir}^T \left[ \frac{\partial \boldsymbol{\eta}_{ir}}{\partial \boldsymbol{p}_{ir}} \right]^{-1} (\boldsymbol{V}_{ir}^{OD})^{-1} (\boldsymbol{r}_{ir} - E(\boldsymbol{r}_{ir})) = 0.$$

Fixing $\rho_i$ or $\sigma_i^2$ and using the IRLS procedure, the iterative equations for $\boldsymbol{\beta}_i$ are

$$\hat{\boldsymbol{\beta}}_i^{(h+1)} = \left( \sum_{r=1}^{5} \boldsymbol{X}_{ir}^T \boldsymbol{W}_{ir}^{OD} \boldsymbol{X}_{ir} \right)^{-1} \sum_{r=1}^{5} \boldsymbol{X}_{ir}^T \boldsymbol{W}_{ir}^{OD} \boldsymbol{z}_{ir},$$

where $\boldsymbol{W}_{ir}^{OD} = \left[\frac{\partial \boldsymbol{\eta}_{ir}}{\partial \boldsymbol{p}_{ir}} \boldsymbol{V}_{ir}^{OD} \frac{\partial \boldsymbol{\eta}_{ir}}{\partial \boldsymbol{p}_{ir}}\right]^{-1}$, and $\boldsymbol{V}_{ir}^{OD}$ is given by (2) or (3).

## 4 Results and Discussion

The results show that the estimated regression coefficients are similar for the multinomial, Dirichlet-multinomial and the random effects models. This small change in the fixed effects estimates under different overdispersion models is as expected, as these are not very sensitive to the exact form of the variance function. Standard errors are obtained from each of the models and compared with model-robust versions (based on the sandwich estimator). These are shown to differ considerably and those from the DM model seem to be overestimated compared to those from the REM model.

**Acknowledgements:** The authors are grateful to Professor S.A. Batista (ESALQ/USP) for supplying the data. S. M. de Freitas acknowledges financial support from CAPES/Brasil.

## References

Glonek, G. F. V. and McCullagh. P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, Series B*, **57**, 149-192.

Hinde, J.P. and Demétrio, C. G. B. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, **27**, 151-170.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *Journal of the Royal Statistical Society, Series B*, **42**, 109-151.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd. ed. Chapman-Hall, London.

O'Hara Hines, R. J. and Lawless, J. F. (1993). Modelling Overdispersion in Toxicological Mortality Data Grouped over Time. *Biometrics*, **49**, 107-121.

Petkau, A. J. and Sitter, R. R. (1989). Models for Quantal Response Experiments over Time. *Biometrics*, **45**, 1299-1308.

Williams, D. A. (1982). Extra-binomial variation in logistic-linear models. *Applied Statistics*, **31**, 144-148.

# Score Tests for Zero-inflated Poisson Models

Naratip Jansakul[1,2] and John P. Hinde[1,3]

[1] School of Mathematical Sciences, Laver Building, University of Exeter, Exeter, EX4 4QE, UK
[2] email: N.Jansakul@exeter.ac.uk
[3] email: J.P.Hinde@exeter.ac.uk

**Abstract:** A score test comparing Zero-inflated Poisson regression, with a constant proportion of excess zeros, to a Poisson regression model was given by van den Broek (1995). This paper extends the test by allowing the zero probability to depend on covariates. The test performs well and gives similar conclusions to using the log-likelihood ratio test.

**Keywords:** Count data, Score test, Poisson, Zero-inflation.

## 1 Introduction

Poisson regression is a standard model for the analysis of count data. However, in practice, data is frequently more variable than specified by the Poisson model and the count data is described as overdispersed. There are various mechanisms that can lead to overdispersion and a number of different overdispersed models have been proposed (see Hinde and Demétrio 1998). This paper concentrates on situations where the numbers of observed zero counts is larger than expected under a standard Poisson model, giving a very specific type of overdispersion. A simple mixture model, the zero-inflated Poisson (ZIP) model, can be used for data of this form — it is simply a mixture of a Poisson model and a degenerate distribution at zero. The ZIP model can incorporate explanatory variables in both the zero process and the Poisson model.

### Zero-inflated Poisson distribution

If $Y_i, i = 1, \ldots, n$ are counts with mean $\mu_i$, the standard Poisson regression model assumes that $Y_i \sim \text{Pois}(\mu_i)$ with variance function, $\text{Var}(Y_i) = \mu_i$ and canonical link, $\log(\mu_i)$, giving the linear predictor $\log(\mu_i) = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $p\times 1$ vector of unknown parameters and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, is a column vector of an $n \times p$ matrix of covariates, $X$.
For count data with more zeros than expected, a number of models have been proposed. Here we will focus attention on the zero-inflated Poisson model, as given by Lambert (1992). The zeros are now assumed to arise

in two different ways; as an observation from the Poisson model, or else, from a degenerate distribution with a probability mass of one concentrated at zero. This gives a simple two-component mixture distribution with the following probability mass function

$$Pr(Y_i = y) = \begin{cases} \omega_i + (1-\omega_i)\mathrm{e}^{-\lambda_i}, & y_i = 0 \\ (1-\omega_i)\dfrac{\mathrm{e}^{-\lambda_i}\lambda^{y_i}}{y_i!}, & y_i = 1, 2, \ldots, \end{cases} \tag{1}$$

which we denote by $Y_i \sim ZIP(\lambda_i, \omega_i)$.
Simple calculations show that

$$\begin{aligned} \mathbf{E}(Y_i) &= (1-\omega_i)\lambda_i = \mu_i \\ \mathrm{Var}(Y_i) &= \mu_i + \left(\frac{\omega_i}{1-\omega_i}\right)\mu_i^2, \end{aligned} \tag{2}$$

indicating that the marginal distribution of $Y_i$ exhibits overdispersion.
The log-likelihood function is given by

$$\begin{aligned} \ell(\boldsymbol{\lambda}, \boldsymbol{\omega}; \mathbf{y}) &= \sum_i \left\{ \mathrm{I}_{(y_i=0)} \log[\omega_i + (1-\omega_i)\mathrm{e}^{-\lambda_i}] \right. \\ &+ \left. \mathrm{I}_{(y_i>0)} \left[\log(1-\omega_i) - \lambda_i + y_i \log\lambda_i - \log(y_i!)\right] \right\}, \end{aligned} \tag{3}$$

where $\mathrm{I}_{(\cdot)}$ is the indicator function for the specified event, i.e. equal to 1 if the event is true and 0 otherwise.
It is clear that this reduces to the standard Poisson model when $\omega_i = 0$. For positive values of $\omega_i$ we have zero-inflation, however, it is possible for $\omega_i < 0$ and still obtain a valid probability distribution (this corresponds to a deficit of zeros – zero-deflation). An extended mixture model, in which $\omega_i$ is not constrained to be non-negative, is commonly referred to as a zero-modified model.
For applying the zero-inflated Poisson model Lambert (1992) suggested the following joint models for $\boldsymbol{\lambda}$ and $\boldsymbol{\omega}$

$$\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \quad \text{and} \quad \log\left(\frac{\boldsymbol{\omega}}{1-\boldsymbol{\omega}}\right) = G\boldsymbol{\gamma}, \tag{4}$$

where $X$ and $G$ are covariate matrices and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are $p \times 1$ and $q \times 1$ vectors of unknown parameters. The use of the logit link function for $\boldsymbol{\omega}$ constrains $\omega_i$ to lie between 0 and 1 and will be problematic when $\boldsymbol{\omega} = 0$, a case of interest as this corresponds to the standard Poisson model. In view of this and the extension to a zero-modified model it may often be useful to use a different link function for $\boldsymbol{\omega}$. The identity link is one possibility giving the joint models

$$\log(\boldsymbol{\lambda}) = X\boldsymbol{\beta} \quad \text{and} \quad \boldsymbol{\omega} = G\boldsymbol{\gamma}. \tag{5}$$

## 2 Score Tests

Testing if the number of zeros is too large for a standard Poisson model corresponds to testing the hypotheses: $\mathrm{H}_0 : \boldsymbol{\omega} = 0$ against $\mathrm{H}_1 : \boldsymbol{\omega} \neq 0$. van den Broek (1995) proposed a score test for this assuming a log-linear model for $\boldsymbol{\lambda}$, but only a constant model for $\boldsymbol{\omega}$. This corresponds to testing $\gamma = 0$ in (5), where $G$ is taken to be an $n \times 1$ matrix of 1's. Interestingly, having rejected the null hypothesis, he then went on to fit models to $\boldsymbol{\omega}$ using a logit link as in (4).
Here we extend the score test statistic to allow $\boldsymbol{\omega}$ to depend upon covariates through (5) and so again testing $\boldsymbol{\omega} = 0$ is equivalent to testing $\boldsymbol{\gamma} = 0$. Since the score test only requires the maximum likelihood estimates of the parameters under the null hypothesis, the general score test for any covariate model for $\boldsymbol{\omega}$ still only involves fitting the standard Poisson model. Based on the log-likelihood function (3) and the models (5), the gradient vector, $S(\boldsymbol{\beta}, \boldsymbol{\gamma})$, and the information matrix, $\mathcal{I}(\boldsymbol{\beta}, \boldsymbol{\gamma})$ can be calculated. Under the null hypothesis, the general score test statistic is

$$S_{\boldsymbol{\omega}} = S^T(\widehat{\boldsymbol{\beta}_0}, 0) C^{-1} S(\widehat{\boldsymbol{\beta}_0}, 0), \tag{6}$$

where $\widehat{\boldsymbol{\beta}}_0$ is the maximum likelihood estimate under the Poisson model and

$$S(\widehat{\boldsymbol{\beta}}_0, 0) = G^T \left[ \frac{\mathrm{I}_{(y_i=0)} - \mathrm{e}^{-\widehat{\boldsymbol{\lambda}}}}{\mathrm{e}^{-\widehat{\boldsymbol{\lambda}}}} \right],$$
$$C = \mathcal{I}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}_0}, 0) - \mathcal{I}_{\boldsymbol{\beta\gamma}}(\widehat{\boldsymbol{\beta}_0}, 0) \mathcal{I}_{\boldsymbol{\beta}}^{-1}(\widehat{\boldsymbol{\beta}_0}, 0) \mathcal{I}_{\boldsymbol{\beta\gamma}}^T(\widehat{\boldsymbol{\beta}_0}, 0),$$

with

$$\mathcal{I}_{\boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}_0}, 0) = X^T \mathrm{diag}(\widehat{\lambda}_i) X,$$
$$\mathcal{I}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}_0}, 0) = G^T \mathrm{diag} \left( \frac{1 - \exp(-\widehat{\lambda}_i)}{\exp(-\widehat{\lambda}_i)} \right) G,$$
$$\mathcal{I}_{\boldsymbol{\beta\gamma}}(\widehat{\boldsymbol{\beta}_0}, 0) = G^T \mathrm{diag}(-\widehat{\lambda}_i) X.$$

In the case of constant $\boldsymbol{\omega}$ this test reduces to that given by van den Broek (1995).

## 3 Sampling Distribution of $S_{\boldsymbol{\omega}}$

Results from a small-scale simulation study indicate that the $\chi^2_{(q)}$ distribution can be used as a reference distribution in many situations, as suggested by asymptotic theory.

## 4 A Composite Test for $\boldsymbol{\omega}$ Models

As we consider more complex models for $\boldsymbol{\omega}$ we get closer to reproducing the observed zeros in the sample and the model for $\lambda$ approaches that for a truncated Poisson distribution. So a comparison of the fit from a Poisson and a truncated Poisson model can be considered as a test of a general (unspecified) model for $\boldsymbol{\omega}$. An alternative is to follow the approach given in Hart (1999), and to use score test statistics $S_{\boldsymbol{\omega}}$ from increasingly complex models for $\boldsymbol{\omega}$. A model for $\boldsymbol{\omega}$ is selected on the basis of the maximum of $S_{\boldsymbol{\omega}}/q$ over a set of possible models. The performance of this procedure is examined with a small simulation study.

## 5 Applications

The use of the extended score test and the above model selection method for $\boldsymbol{\omega}$ are illustrated on two examples:

- Shoot generation from micro-propagation experiments on the apple variety Trajan (Ridout *et al.* 1998);
- Urinary tract infections for HIV-infected men (van den Broek 1995).

For these two examples, the score test procedure using $S_{\boldsymbol{\omega}}/q$ chooses the same model for $\boldsymbol{\omega}$ as using the log-likelihood ratio test or the Akaike Information Criterion, but requires only the Poisson model to be fitted.

**Acknowledgements:** The authors are grateful to J. van den Broek for providing the HIV data. N. Jansakul acknowledges financial support from The Government of Thailand.

### References

Hart, J. D. (1999). Testing the fit of functions in fully specified likelihood models. In H. Friedl, A. Berghold, and G. Kauermann, Editors, $14^{th}$ *Int$^l$ Workshop on Statistical Modelling*, Graz, Austria, 19–29.

Hinde, J. and Demétrio, C. G. B. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Ridout, M., Demétrio, C. G. B. and Hinde J. (1998). Models for count data with many zeros. In *International Biometric Conference*, Cape Town, 179–190.

van den Broek, J. (1995). A score test for zero-inflation in a Poisson distribution. *Biometrics*, **51**, 738–743.

# Bias Adjusted Pearson estimating functions

S.J. Knudsen[1]

[1] Department of Statistics and Demography, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark; www.statdem.sdu.dk/∼ sjk/

**Abstract:** Generalized estimating equations (GEEs) is a method for regression paramters in longitudinal, and other clustered data models. Often moment estimators are used for estimating the correlation of repeated responses, and simple degrees-of-freedom adjustments are applied to prevent a bias. Instead, we propose Peason estimating functions for covariance parameters, and aply a general bias adjustment of Jørgensen & Knudsen (2000), obtaining REML-like results. We compare with degrees-of-freedom methods, and show that these are deficient.

**Keywords:** Nuisance parameter bias; Profile estimating functions; Mixed Model.

## 1 Pearson Estimating functions

Let $y_i$ denote the $n_i$ response vector, observed for the $i$th of $k$ independent clusters. Let $E(y_i) = \mu_i(\beta)$ denote the mean, $\beta$ being a $p$-vector of regression coefficients, say. For each marginal $y_{it}$, let $\mathrm{Var}(y_{it}) = \sigma^2 v(\mu_{it})$, where $v$ is a known variance function and $\sigma^2$ an unknown dispersion parameter. Assume further that the covariance matrix of $y_i$ is given by

$$\mathrm{Var}(y_i; \beta, \alpha) = V_i^{\frac{1}{2}}(\mu_i) R_i(\alpha) V_i^{\frac{1}{2}}(\mu_i) \ .$$

Here $V_i$ is a diagonal matrix of variance functions, and $R_i$ is a $n_i \times n_i$ regular matrix that depends on an $s$-vector of covariance parameters $\alpha$. We estimate $\beta$ by solving the optimum GEE

$$g_\beta(\beta, \alpha) = \frac{1}{k} \sum D_i^\top \mathrm{Var}(y_i)^{-1} (y_i - \mu_i) = 0 \ ,$$

where $D_i = \nabla_\beta \mu_i$ are $n_i \times p$ gradients all of full rank $p$.

Denote the $i$th *Pearson residual vector* by $r_i = V_i^{-\frac{1}{2}}(y_{it} - \mu_{it})$. A *Pearson estimating function* for the $j$th component of $\alpha$ takes the following form:

$$g_{\alpha j}(\beta, \alpha) = \frac{1}{k} \sum \left\{ r_i^\top A_{ij}^{-1} r_i - \sigma^2 \mathrm{trace}\left(A_{ij}^{-1} R_i\right) \right\} \ , \tag{1}$$

where $A_{ij}$ are regular $n_i \times n_i$ matrices such that $E(\partial g_{\alpha j}/\partial \alpha_j) \neq 0$. The collection $g_\alpha$ of $g_{\alpha j}$s is hence regular for estimating $\alpha$; that is, *unbiased* $E(g_\alpha) = 0$ and the *sensitivity matrix* $E(\nabla_\alpha g_\alpha)$ is regular.

The Pearson estimator for $\sigma^2$, familiar in generalized linear models (GLMs), corresponds to $A_{i1} = R_i = I_i$, in which case $\text{trace}(A_{i1}^{-1} R_i) = n_i$. Included are also normal linear mixed models, where score functions for variance components coincide with (1); see Searle *et al.* (1992, Eq. 6.2.16). In fact, $A_{it}$ may, in general, be inspired by these score functions, arriving at a general $\alpha$-estimating procedure.
Combining $g_\beta$ and $g_\alpha$ gives an unbiased estimating function $g = (g_\beta, g_\alpha)$ for $\theta = (\beta, \alpha)$, say. Thus, under further regularity conditions, $\hat{\theta}$ is consistent and asymptotically normal; see Knudsen (1999, Ch. 2).
In particular, note that $E(\nabla_\alpha g_\beta) = 0$, since $g_\beta(\beta, \alpha)$ depends on $\alpha$ only through $R(\alpha)$. This implies $\alpha$-insensitivity and hence *efficiency-stable* estimation of $\beta$; that is, the asymptotic variance for $\hat{\beta}$ is preserved whether $\alpha$ is treated as known or unknown; see Knudsen (1999, Ch. 2) and Jørgensen & Knudsen (2000).

## 2 Bias adjustment

Often $g_\alpha(\beta, \alpha)$ is sensitive to changes in $\beta$. The *profile estimating function* $\hat{g}_\alpha(\alpha) = g_\alpha(\hat{\beta}_\alpha, \alpha)$ for $\alpha$ along a GEE estimator $\hat{\beta}_\alpha$ at $\alpha$, hence suffer from a bias that may lead to inconsistency and over-optimistic estimates. Under the likelihood set-up, these problems may, in general, be corrected using modified profile likelihoods of Barndorff-Nielsen and Cox & Reid; see McCullagh & Tibshirani (1990) and references therein. Such techniques, however, are not readily available for estimating functions, since these are not gradients of any pseudo-likelihood, say.
A more viable approach is given by Jørgensen & Knudsen (2000), who extend McCullagh & Tibshirani's (1990) bias adjustment for profile score functions. Their adjustment $b$ is defined so that the bias of $\hat{g}_a - b$ is $O\left(k^{-3/2}\right)$. If we assume that all $A_{ij}$ do not depend on $\beta$, then the $j$th component of $b$ is given by $b_j = -\sigma^2 \tilde{p}_j + \eta_j$. Here $\tilde{p}_j$ is easy to compute,

$$\tilde{p}_j = \text{trace}\left\{ \left( \sum D_i^\top V_i^{-\frac{1}{2}} A_{ij}^{-1} V_i^{-\frac{1}{2}} D_i \right) \left( \sum D_i^\top \text{Var}(y_i)^{-1} D_i \right)^{-1} \right\} .$$

Unfortunately, $\eta_j$ is a rather complicated; see Jørgensen & Knudsen (2000). We ignore $\eta_j$ for various reasons: it depends on higher-order moments, and introduces $O(|\alpha|^2)$-terms. Correspondingly, we obtain

$$\tilde{g}_{\alpha j} = \frac{1}{k} \sum \left[ r_i^T A_{ij}^{-1} r_i - \sigma^2 \left\{ \text{trace}\left(A_{ij}^{-1} R_i\right) - \tilde{p} \right\} \right] , \tag{2}$$

for estimating the $j$th component of $\alpha$. We call the solution to $\tilde{g}_\alpha = 0$ for a *BAPE*, an acronym for *Bias Adjusted Pearson Estimator*, and refers to the estimator rather than the estimating function itself.

For GLMs, where $A_{i1} = R_i = I_i$, then $\tilde{p}_1 = p$. Thus the BAPE for $\sigma^2$ is just the Pearson estimator adjusted by degrees of freedoms. Note that if $A_{i1} = I_i$ but $R_i$ represents positive correlation, $\tilde{p}_1 > p$.
For normal linear mixed models, Tibshirani & McCullagh (1990) show that bias adjusting the score function corresponds to REML-estimation, which is known to have nice properties. Their result and $\tilde{g}_\alpha$ coincide, since $\eta = 0$ for normal data models; see Jørgensen & Knudsen (2000).

## 3 Simulation results: one-way Gamma model

Consider iid random effects $u_1, \ldots, u_k$. Assume that all $y_{it}$ are conditionally independent given all $u$s, that the conditional distribution of $y_i$ given $u$s depends only on $u_i$, and that for all $i$ and $t$

$$y_{it} | \, u \sim \mathrm{Ga}(\mu_{it} u_i, \; \omega^2 / u_i) \qquad \text{and} \qquad u_i \sim \mathrm{Ga}\left(1, \tau^2\right) \; ,$$

where $\mathrm{Ga}\left(\mu, \sigma^2\right)$ denotes a Gamma model with mean $\mu$ and variance $\sigma^2 \mu^2$. Thus, by calculus of conditional and unconditional moments, $E\left(y_{it}\right) = \mu_{it}$,

$$\mathrm{Var}\left(y_{it}\right) = \left(\omega^2 + \tau^2\right) \mu_{it}^2 \qquad \text{and} \qquad \mathrm{Cov}\left(y_{it}; y_{iz}\right) = \tau^2 \mu_{it} \mu_{it} \; ,$$

between all $t \neq z$. The correlation structure is hence *exchangeable*, and $R_i = \omega^2 I_i + \tau^2 J_i$, where $J_i$ are $n_i \times n_i$ with every element unity. The variance function is $v\left(\mu\right) = \mu^2$, and the dispersion parameter is $\sigma^2 = \omega^2 + \tau^2$. We interpret $\omega^2$ and $\tau^2$ by calling them *dispersion components*.
The simulation study is based on assuming $\tau^2 = \omega^2 = 0.05$, 0.1 and 0.25. For each value, 1000 data sets where generated, assuming a simple log-polynomial regression,

$$\log\left(\mu_{it}\right) = \beta_1 + \beta_2 t + \beta_3 t^2 \qquad \text{for all } i, t \; ,$$

with true regression coefficients $\beta_1 = 0.01$, $\beta_2 = -0.10$, and $\beta_3 = 0.02$; lack of space prevents us from studying estimates here.
Three methods for estimating $\alpha = \left(\omega^2, \tau^2\right)$ are compared: L-Z, BLUP and BAPE. The L-Z corresponds to the GEE library for SPLUS, and is based on the following moment estimators adjusted by degrees of freedoms:

$$\hat{\omega}^2 = \frac{1}{n-p} \sum \hat{r}_i^\top \hat{r}_i - \hat{\sigma}_u^2 \qquad \text{and} \qquad \hat{\tau}^2 = \frac{1}{m-p} \sum_i \sum_{t>z} \hat{r}_{it} \hat{r}_{iz} \; .$$

with $n = \sum n_i$ and $m = \frac{1}{2} \sum n_i \left(n_i - 1\right)$; see Liang & Zeger (1986, p.17–18).
In the BLUP-approach, proposed by Ma (1999), estimation on $\omega^2$ and $\tau^2$ is based on regular Pearson estimating functions, defined by

$$g_\omega = \sum \left(r_i^\top r_i - n_i \sigma^2\right) \qquad \text{and} \qquad g_\tau = \sum \left(r_i^\top A_{i\tau} r_i - \tau^2 n_i w_i\right) \; , \quad (3)$$

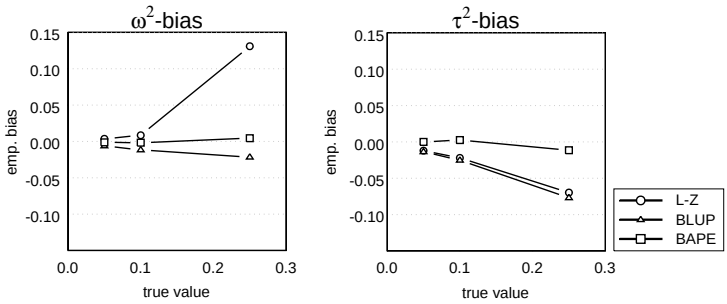



FIGURE 1. Empirical bias for $\hat{\omega}^2$ and $\hat{\tau}^2$.

respectively. Here $A_{i\tau} = w_i^2 J_i$ and $w_i = \tau^2 / \left(\omega^2 + n_i \tau^2\right)$ for all $i$.
The BAPE method corresponds to $\tilde{g}_\omega$ and $\tilde{g}_\tau$, where, in (2),

$$\tilde{p}_j = \text{trace}\left\{ \left(\sum X_i^\top A_{ij}^{-1} X_i\right) \left(\sum X_i^\top R_i X_i\right)^{-1} \right\} .$$

Here $X_i$ denotes $n_i \times p$ design matrices, and $A_{i\omega} = I_i$ for $\omega^2$.
The magnitude of bias is illustrated in Figure 1. For true $\omega^2$ large, both BLUP and BAPE are superior to L-Z, which is seriously biased. Only BAPEs for $\tau^2$ are nearly unbiased compared with BLUP and L-Z. In particular, note that, for L-Z, the bias of the Pearson estimator adjusted by scanty degrees of freedoms$\hat{\sigma}^2 = \hat{\omega}^2 + \hat{\tau}^2$ does not cancel out.

## References

Jørgensen, B. and Knudsen, S.J. (2000). Parameter orthogonality and bias adjustment for estimating functions. To appear.

Knudsen, S.J. (1999). *Estimating Functions and Separate Inference.* Odense University, Dep. of Stat. and Dem., Monographs, **1.**

Liang, K.E. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Ma, R. (1999) *An Orthodox BLUP Approach to generalized Linear Mixed Models*, Odense University. Dep. of Stat. and Dem., Monographs,**2.**

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J.R.Statist.Soc. B*,**52**, 352–344.

Searle S.R., Casella G., and McCulloch C.E. (1992). *Variance Components.* Wiley-Interscience, USA

# Optimization of Experimental Designs

Miguel A.P.M. Lejeune[1]

## 1 Introduction

The idea underlying the construction of optimal experimental designs is to choose a design that optimizes some inference criterion. The optimization problem is difficult to solve analytically. The recourse to experimental design theory has become widespread with the increasing possibilities offered by computers. But, as for the new computers generation, attempting for constructing experimental designs remains very complicated and time-consuming. The difficulty increases exponentially with the number of predictor variables.

## 2 Goals

Taking into account the time consideration, we thus aim at implementing an algorithmic process with which we are able to get highly efficient designs for very large models (up to twenty predictor variables). This algorithmic process presents the following characteristics:
Most exchange algorithms developed, so far, for the construction of optimal designs, performs an optimization operation at each iteration. They consider all points (or experiments) of the candidate space and select the point which maximizes the selected optimality criterion. As opposed to this, we neither construct nor store the candidate space. Our algorithm proposes randomly a new point of the candidate space and does not require such maximization operations. The process is less time-consuming and does not involve any trade-off between the time saved and the level of efficiency reached.

1. We implement a generalized simulated annealing algorithm. The main motive is that simulated annealing is characterized by the possibility to accept a detrimental move. This reduces the risk being trapped in a local minima and facilitates the search of the global optimum.

2. We propose a method generating non-random starting designs. This methods proves to be very indicated for complex first- and second-order models.

3. We customized the algorithmic process to the type of the considered designs (saturated or non-saturated). Time-saving and efficiency considerations lead us to differentiate the algorithmic process for saturated and non-saturated designs.

## 3 Exchange Procedure

In this paper, we propose a new exchange algorithm which can be decomposed into four main steps:

1. Steps 1 and 2 substitute an experiment or point $x_l$ (among the N experiments) for a point $x_k$. The choice of the point $x_k$ to replace depends on the variance of the predicted response at this point. We replace the point $x_k$ which has the lowest generalized variance, where the generalized variance of the predicted response at point $x_i$ is defined as follows: var $((xi)) = f'(x_i)\ (X'X)^{-1} f(xi)$ (1)

   For saturated designs (designs for which the number of experiments N is equal to the number of parameters P), point substitution cannot be performed in the same manner as for non-saturated designs. In fact, by considering a saturated design, the variance computed with formula(1) would be equal to 1 for each point of the N points $xi$. Thus, the variance would lose its discriminatory power and would not allow to select a point $x_k$ to be removed as in (2).

   var $((x_k)) = f'(x_k)\ (X'X)^{-1}\ f\ (x_k) = f'(xi)\ (X'X)^{-1}\ f(xi)$ (2)

   We remove the point $x_k$ and we add a point $x_l$.

2. In step 3, we decide whether the tentative design obtained through the above substitution will become the new current solution in the iterative process.

3. Step 4 defines the stopping criterion of the iterative process.

- Comments on the point substitution Our point substitution procedure is non-sequential. The addition and the removal of points take place during a single iteration. The number of points included in the design remains fixed, equal to N.

  One of the main differences between our algorithm and the K-exchange algorithm developed by Johnson and Nachtsheim [1983] or the K-Coordinate-Exchange algorithm developed by Meyer and Nachtsheim [1995] lies in the fact that their methods are greedy heuristics. They are iterative procedures searching a new solution that maximizes, at each step, the marginal benefit. At each iteration, they consider all candidates in the design space and opt for an exchange, that maximizes the increase of the determinant of (X'X). Our algorithm generates a new neighbouring point or experiment. This does not involve any maximization operation and so reduces the computation time.

A second difference is that we decide to modify only one point (K=1) per iteration. The first reason is that, as noticed by Johnson, this option provides good results whatever the model considered. This is not always the case for larger K. Indeed, the influence of the type of model, the number of factors, parameters and experiments on the value of K is difficult to model. Secondly, the substitution of one point per iteration is less time-consuming. Finally, setting K equal to 1 also allows to stabilize the structure of the design. This stability property is advisable or even required when we use a procedure generating non-random starting designs (see section 4).

- Comments on the Simulated Annealing approach Greedy or local-neighbourhood-search methods have a well-defined termination rule and often stop at a local minimum. Metaheuristics, among them simulated annealing, are more complex algorithms and search for better solutions until an arbitrary stopping point is reached. Simulated annealing, by allowing the search patterns to move away from a path of strict descent and by allowing the random oscillations to be large enough, tends to avoid local minima [Haines, 1987]. This is illustrated by the following figure:

## 4 Non-Random Starting Design

Some authors have implemented a non-sequential method to generate non-random starting designs. They claimed that this method has the advantage of being able to add m experiments simultaneously; this set of m experiments depends not only on the already incorporated experiments but also takes into account the interrelationship of the m experiments to each other. The sensitivity analysis conducted by Johnson and Nachtsheim does not corroborate this argument and shows that sequential methods are at least as efficient and are less time-consuming. That is why we have opted for a sequential method.

So that our non-random starting procedure be effective, we do not have to turn around the structure of the design at each iteration. To have developed a point substitution algorithm for which K is equal to 1 is consistent with this need. We use this method especially for very large problems. It aims at converging more rapidly to the optimum. Starting with an initial non-random design is more time-consuming than starting with an initial random design. This time increase at the begin of the computation process can, for larger models, be compensated in the course of the iterative process. In fact, the procedure generating non-random starting designs helps converging more rapidly and thus decreases the number of iterations.

## 5 Conclusion

In the literature, we find that the difficulty to construct D-optimal increases exponentially with the number of predictor variables. As a result, we do not find in the literature D-optimal designs for first order models including more than 13 factors. Most exchange algorithms implemented for the construction of D-optimal designs store all candidate points and perform an optimization operation for each of these $R^Q$ points at each iteration (R is the number of levels that each factor can take). The exchange algorithm described in this paper does not require such an optimization sequence. This was not compatible with our desire to apply our algorithm to the elaboration of optimal designs for complex models.
We do not perform the algorithmic process in the same way for saturated and non-saturated designs. The process implemented for non-saturated designs is more rapid but cannot be applied for saturated designs because of the properties of the information matrix. We have thus adapted the iterative process for saturated designs. This customized process induces the computing and updating of an additional matrix.
We have implemented a simulated annealing algorithm to construct D-optimal designs for linear models in an attempt to avoid the problems of premature convergence. We have opted for a generalized simulated annealing. We also propose a procedure for generating non-random starting designs. The idea is that the additional time needed at the start of the process can be compensated through the iterative process by a more rapid convergence. This is interesting for larger models. The results we obtain demonstrate that our algorithmic procedure is able to create highly efficient designs, almost 50% of them have a greater efficiency than those found in the literature. Especially for large models, our method turns out to be very effective. This has led us to apply it to models more complex that those handled in the literature. We reach a 90 % D-efficiency level for models with up to 20 predictors.
We also construct an experimental design for a butter producing company. We so apply our algorithm to a second-order model, for which we had to take into account 10 predictor variables and 7 interactions. We obtain a design that reaches a D-efficiency of 95%, that is a greater level of efficiency that the company could reach by using a professional software.

## References

ATKINSON A. & DONEV A. (1997). Experimental Designs optimally balanced for Trends. *Technometrics*, **38**, 333-341.

MEYER R. & NACHTSHEIM C. (1995). The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs. *Technometrics*, **37**, 60-68.

# An algorithm for the construction of cost-efficient and trend-resistant experiments

Lieven Tack[1], Martina Vandebroek[1]

[1] Department of Applied Economics, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium

**Abstract:** This paper presents an algorithm for the construction of cost-efficient designs that are optimally balanced for time trends. The algorithm provides the experimenter with a general method for solving a wide range of practical problems. The results show that the constructed run orders have outstanding performance both in terms of $\mathcal{D}$-optimality and costs.

**Keywords:** Trend-Resistance, Cost, $\mathcal{D}$-optimality, Design Algorithm

## 1 Introduction

In practice, costs often limit the usefulness of optimum experimental designs. A first cost approach deals with measurement costs, i.e. the costs associated with particular factor level combinations. Examples are the equipment cost, the cost of material, the cost of personnel and the cost for spending time during the measurement. Another approach takes into account transition costs or the costs for changing the factor levels from one observation to another. Consider as an example the cost for changing oven temperature. In order to minimize the total cost of the experiment, the number of factor level changes should be kept as low as possible.
Furthermore, when performing the observations in a time sequence, the experimenter often has reason to believe that the observed responses will be influenced by a temporal trend. Examples include equipment wear-out, learning, fatigue and warm-up effects. Minimizing the number of factor level changes is then no longer the only design issue. The objective is to strike a balance between cost-efficiency and the degree of trend-resistance. The existing literature suffers from the fact that it ignores cost considerations and attention is mainly restricted to low-order time trends, factors with only two or three levels, equally spaced time points and regular design spaces. This paper will present an algorithm for dealing with arbitrary design problems.

## 2 Trend-resistant and cost-efficient run orders

Tack and Vandebroek (1999) define the total measurement cost $C^m$ of an experiment as

$$C^m = \sum_{i=1}^{d} n_i c^m(\mathbf{x}_i), \tag{1}$$

where $n_i$ represents the number of replicates at design point $\mathbf{x}_i$ and $c^m(\mathbf{x}_i)$ is the measurement cost at design point $\mathbf{x}_i$. Note that $\sum_{i=1}^{d} n_i$ equals the number of observations $n$. Similarly, the total transition cost $C^t$ is

$$C^t = \sum_{i=1,j=1}^{d} n_{i,j} c^t(\mathbf{x}_i, \mathbf{x}_j), \tag{2}$$

where $n_{i,j}$ denotes the number of transitions from design point $\mathbf{x}_i$ to design point $\mathbf{x}_j$ and $c^t(\mathbf{x}_i, \mathbf{x}_j)$ represents the transition cost from design point $\mathbf{x}_i$ to design point $\mathbf{x}_j$. The total cost $C$ of a run order then equals the sum of the total measurement cost (1) and the total transition cost (2).
If block effects are present, the $n$ runs are assumed to be arranged in $b$ blocks of specified sizes $k_1, \ldots, k_b$ with $n = \sum_{i=1}^{b} k_i$. With $\boldsymbol{\alpha}$ the $(p \times 1)$ vector of important parameters, $\boldsymbol{\beta}$ the $(q \times 1)$ vector of parameters of the polynomial time trend and $\boldsymbol{\gamma}$ the $(b \times 1)$ vector of block effects, the model for the $n$ responses becomes

$$\mathbf{y} = \mathbf{F}\boldsymbol{\alpha} + \mathbf{G}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{F}$ and $\mathbf{G}$ represent the $(n \times p)$ and the $(n \times q)$ extended design matrices for the response function and the time trend respectively. Finally, $\mathbf{Z}$ equals $\mathrm{diag}(\mathbf{1}_{k_1}, \ldots, \mathbf{1}_{k_b})$, $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}_n$. If the block effects are random, we assume that $\mathrm{E}(\boldsymbol{\gamma}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\gamma}) = \sigma_\gamma^2 \mathbf{I}_b$ and $\mathrm{cov}(\boldsymbol{\gamma}, \boldsymbol{\varepsilon}) = \mathbf{0}$. Remark that when no blocking is used, $\mathbf{Z}$ and $\boldsymbol{\gamma}$ are left out from (3).
In the absence of trend effects $\boldsymbol{\beta}$, the $\mathcal{D}$-optimal design $\delta_{\mathcal{D}}$ that maximizes the information on $\boldsymbol{\alpha}$ is found by maximizing

$$\mathcal{D} = \frac{1}{\sigma_\varepsilon^{2p}} \left| \mathbf{F}'\mathbf{F} - \sum_{i=1}^{b} \frac{\sigma_\gamma^2}{\sigma_\varepsilon^2 + k_i \sigma_\gamma^2} \left(\mathbf{F}_i' \mathbf{1}_{k_i}\right) \left(\mathbf{F}_i' \mathbf{1}_{k_i}\right)' \right|, \tag{4}$$

where $\mathbf{F}_i$ refers to that part of $\mathbf{F}$ that corresponds to the $i$th block. When no blocking is used, the second term in (4) is omitted.
Taking into account cost information, a run order $\delta_{(\mathcal{D},C)}$ is said to be $(\mathcal{D}, C)$-optimal if it maximizes the amount of information per unit cost. The appropriate optimality criterion becomes $(\mathcal{D}, C) = \mathcal{D}^{\frac{1}{p}}/C$.
In order to evaluate the influence of trend effects on the variance of the parameter estimates, the $\mathcal{D}_t$-optimality criterion maximizes the information

on the important parameters $\boldsymbol{\alpha}$, whereas the $q$ parameters modeling the time dependence are treated as nuisance parameters. The associated $\mathcal{D}_t$-optimal design $\delta_{\mathcal{D}_t}$ is found by maximizing

$$\mathcal{D}_t = \frac{1}{\sigma_\varepsilon^{2p}} \frac{\left| \begin{bmatrix} \mathbf{F}' \\ \mathbf{G}' \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{G} \end{bmatrix} - \sum_{i=1}^{b} \frac{\sigma_\gamma^2}{\sigma_\varepsilon^2 + k_i \sigma_\gamma^2} \left( \begin{bmatrix} \mathbf{F}' \\ \mathbf{G}' \end{bmatrix} \mathbf{1}_{k_i} \right) \left( \begin{bmatrix} \mathbf{F}' \\ \mathbf{G}' \end{bmatrix} \mathbf{1}_{k_i} \right)' \right|}{\left| \mathbf{G}'\mathbf{G} - \sum_{i=1}^{b} \frac{\sigma_\gamma^2}{\sigma_\varepsilon^2 + k_i \sigma_\gamma^2} (\mathbf{G}_i' \mathbf{1}_{k_i})(\mathbf{G}_i' \mathbf{1}_{k_i})' \right|}. \tag{5}$$

Again, the second term in numerator and denominator of (5) vanishes if no blocking is used. The degree of trend-resistance of the $\delta_{\mathcal{D}_t}$-optimal design is measured by means of

$$\left( \frac{\mathcal{D}_t(\delta_{\mathcal{D}_t})}{\mathcal{D}(\delta_{\mathcal{D}})} \right)^{\frac{1}{p}}. \tag{6}$$

Finally, when also costs are taken into account, the $(\mathcal{D}_t, C)$-optimal run order $\delta_{(\mathcal{D}_t, C)}$ maximizes ${\mathcal{D}_t}^{\frac{1}{p}}/C$. In the next section, we present our algorithm to compute trend-resistant and cost-efficient run orders.

## 3 The design algorithm

In the first phase of the algorithm, a starting run order is constructed by adding $n_r$ randomly chosen design points to arbitrarily chosen time points from a list of user-specified design points and time points. In the second phase, the resulting run order is augmented to $n$ trials by optimizing a user-specified optimality criterion. Finally, the trials are subject to iterative improvement in the third phase. This improvement of the run order consists of alternate exchanges of design points and time points. An exchange leading to the largest improvement of the optimality criterion will be performed. The process continues as long as a design change increases the value of the optimality criterion.

## 4 The flame spectroscopy experiment

This case is based on an example mentioned by Joiner and Campbell (1976) who describe an experiment in which the sensitivity of a spectrophotometer is evaluated (Table 1). The measurements are believed to drift linearly with time due to carbon build-up. We assume the total measurement cost to be fixed and the transition costs are proportional to the times needed to change the factor levels. Our proposed algorithm is used to construct optimal run orders for four different response models (Table 2). Model (2) adds quadratic terms to the first order model (1). Model (3) contains all

linear terms and 2-factor interactions, whereas model (4) contains all linear terms, quadratic terms and 2-factor interactions. Table 2 shows that allowing for cost considerations implies large reductions in the total cost of an experiment. However, this reduction goes at the cost of the degree of trend-resistance of the computed $(\mathcal{D}_t, C)$-optimal run orders. The decease in trend-resistance especially comes true for the first two models. But roughly speaking, the $(\mathcal{D}_t, C)$-optimal run orders outperform the $\mathcal{D}_t$-optimal ones in terms of the cost of information. The reductions in the cost of information range from 61% to 79%.

| | factor | number of levels | time to change (sec) |
|---|---|---|---|
| $x_1$ | lamp position | 2 | 1 |
| $x_2$ | burner position | 2 | 60 |
| $x_3$ | burner height | 3 | 1 |
| $x_4$ | type of flame | 3 | 60 |
| $x_5$ | flow rate | 3 | 120 |

TABLE 1. The Flame Spectroscopy Experiment

| | transition cost | | trend-resistance (%) | | cost per unit information | |
|---|---|---|---|---|---|---|
| model | $\delta_{\mathcal{D}_t}$ | $\delta_{(\mathcal{D}_t,C)}$ | $\delta_{\mathcal{D}_t}$ | $\delta_{(\mathcal{D}_t,C)}$ | $\delta_{\mathcal{D}_t}$ | $\delta_{(\mathcal{D}_t,C)}$ |
| (1) | $\geq$ 2669 | 394 | 100 | 45.11 | $\geq$ 134 | 46 |
| (2) | $\geq$ 2537 | 623 | 99.99 | 64.94 | $\geq$ 245 | 95 |
| (3) | 4170 | 694 | 82.70 | 66.03 | 266 | 57 |
| (4) | 3864 | 1357 | 77.81 | 72.51 | 422 | 161 |

TABLE 2. Comparison of Optimal Run Orders for Different Response Models

## 5 Conclusion

The proposed algorithm enables one to construct cost-efficient and trend-resistant run orders. It is intended to demonstrate something of the wide range of important problems that can be solved. Industrial examples illustrate practical utility and the results show that incorporating cost information highly diminishes the cost of information. The resulting loss in balance for time trends is rather negligible.

## References

Joiner, B. L. and Campbell, C. (1976). Designing Experiments When Run Order is Important. *Technometrics*, **18**, 249-259.

Tack, L. and Vandebroek, M. (1999). $(\mathcal{D}_t, C)$-Optimal Run Orders. *Research Report* 9957, Katholieke Universiteit Leuven.

# Controlling Qualitative Confounders in Nonrandomized Experiments: A Method and its Implementation in SPSS.

Olivia Wuethrich-Martone[1], Marc Müller, and Rolf Steyer

[1] Inst. of Psychology, Friedrich Schiller University, Am Steiger 3/Haus 1, D-07743 Jena, Germany. Tel.: +49-3641-945236. E-mail: s6wuol@rz.uni-jena.de

**Abstract:** In nonrandomized experiments — or not fully controlled studies — conclusions drawn fr om the outcome of an experiment may not always be regarded as a direct consequence of the treatment itself because a third variable, or confounder, may have affected the results. Since a fully randomized experiment is often not feasible or even totally inappropriate because of ethical or costs reasons, an alternative is to outline procedures for controlling potential confounders in nonrandomized experiments. This paper suggests a procedure — and its implementation as a routine in SPSS — for controlling qualitative confounders when qualitative treatments are compared. The procedure is based on the idea of testing for causal effects.

**Keywords:** Nonrandomized Experiments; Qualitative Confounders; Treatments; Causal Effects.

## 1 Introduction

The causal model underlying a randomized experiment is straightforward: once that treatments are accurately manipulated, randomization of units to treatments ensures that the outcome is a direct consequence of the treatments and not of some other spurious influences. One of the goals of causal analysis is to state causal models even for nonrandomized experiments, i.e. for example to individuate causal relationships between the considered treatments when potential confounders are not completely controlled.
The approach to causal modeling we use is due to Neyman (1923/1990), Rubin (1974), Holland (1986) and others. The formalism we adopt is based on Probability Theory and has been presented by Steyer et al. (1995). This approach focalizes on the so-called individual causal effects which are defined as differences between the conditional expected values of the observed variable given a person and a treatment and on their averages over the considered population (average causal effects).
The idea is that even nonrandomized experiments may lead to useful estimates of causal effects once we accurately plan the study itself, for example

by listing and recording the major potential confounding variables and by finding appropriate methods for controlling the bias that may induce.
This paper presents a method for testing causal effects in a design with qualitative treatment variables, qualitative confounders and a continuous response variable. We actually extend the results obtained by Wuethrich-Martone et al. (1999) for a 2-factors unbalanced Analysis of Variance to the most general case of an Analysis of Variance with $p$-treatment variables and $q$-confounders.
Since to date no statistical software offers a straightforward procedure for testing causal effects, we implemented our method in SPSS. A short description of our implementation and of its usage is included.

## 2 Method

Let $Y$ be a continuous response variable and let $X_1, \ldots, X_p$ be $p$ treatment variables. The purpose is to investigate the effects of the treatment variables on $Y$ when $q$ potential confounders $W_1, \ldots, W_q$ apply and a randomized experiment is not feasible. We assume that both the treatment variables and the potential confounders are qualitative and have a joint distribution. Denote with $U$ the unit variable, i.e. the qualitative variable indicating which unit $u$ is observed. Each single unit $u$ undergoes $p$ different treatments represented respectively by a specific value of $X_1$, a specific value of $X_2$ and so on. We may therefore associate with every single unit $u$ a $p$-tuple of treatments which is the combination of the original $p$ treatments. More precisely, the $p$-tuple associated with a unit $u$ is a specific value of a 'new' treatment variable, the so-called *vector treatment*, defined as $\mathbf{X} := (X_1, \ldots, X_p)$. Similarly we define a 'new' potential confounder, the so-called *vector potential confounder*, as $\mathbf{W} := (W_1, \ldots, W_q)$.
Referring for the theoretical background and for the notation to the article of Steyer et al. (1995), the *Individual Causal Effect* of a treatment $\mathbf{x}_i$ vs. a treatment $\mathbf{x}_{i'}$ on (the expectation of) $Y$ for a specific unit $u$ is defined as $ICE_{ii'}(u) := E(Y|\mathbf{X} = \mathbf{x}_i, U = u) - E(Y|\mathbf{X} = \mathbf{x}_{i'}, U = u)$.
Since a unit $u$ can usually be observed under just one condition we concentrate on the *Average Causal Effect* $ACE_{ii'} := \sum_u ICE_{ii'}(u)P(U = u)$, where $P(U = u)$ is the probability of unit $u$ to be observed. In a study we would normally estimate the mean differences $E(Y|\mathbf{X} = \mathbf{x}_i) - E(Y|\mathbf{X} = \mathbf{x}_{i'})$ which are also called *Prima Facie Effects* of a treatment $\mathbf{x}_i$ vs. another treatment $\mathbf{x}_{i'}$. These effects may be causally interpretated only if they are *causally unbiased*, i.e. if they are equal to the corresponding $ACE_{ii'}$ (see for example the Simpson's paradox in Steyer, 1992, Section 3.2).
In randomized experiments the *PFE* are always causally unbiased, but this is not the case in a non-randomized experiment. The goal is therefore to outline methods for estimating and testing causal effects without the benefit of randomization in particular when potential confounders apply. We represent the influence of a potential confounder by an appropriate

adjustment of the treatment means using as weights the marginal probabilities corresponding to the different values of the confounder. We define $E_{adj_{\mathbf{W}}}(Y|\mathbf{X}=\mathbf{x}) := \sum_{\mathbf{w}} E(Y|\mathbf{X}=\mathbf{x},\mathbf{W}=\mathbf{w})P(\mathbf{W}=\mathbf{w})$ conditional expected value of $Y$ *adjusted* for potential confounding w.r.t. $\mathbf{W}$, where $P(\mathbf{W}=\mathbf{w}) := P(W_1=w_1, W_2=w_2, \ldots, W_q=w_q)$.
If we make sure that no other important potential confounder has been ignored, for example assigning units to treatments by conditional randomization (randomly for each value $\mathbf{w}$ of $\mathbf{W}$), then we may extend to our design the results obtained by Wuethrich-Martone et al. (1999) for a 2-factors unbalanced Analysis of Variance. First, once the distribution of $\mathbf{W}$ is known the average causal effects may be computed as $ACE_{ii'} = E_{adj_{\mathbf{W}}}(Y|\mathbf{X}=\mathbf{x}_i) - E_{adj_{\mathbf{W}}}(Y|\mathbf{X}=\mathbf{x}_{i'})\ \ \forall i\,\forall i'$. The influence of the potential confounders results then 'included' in the causal effects.
A second result is that the hypothesis $ACE_{ii'} = 0\,\forall i\,\forall i'$ may be tested in form of a General Linear Hypothesis $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, where the elements of $\boldsymbol{\beta}$ are the true cell means and the rows of $\mathbf{C}\boldsymbol{\beta}$ represent the different average causal effects. To build $\mathbf{C}$ we just need to know the distribution $P(\mathbf{W}=\mathbf{w})$ of the vector confounder, i.e. the joint distribution of the original $q$ potential confounders. The corresponding F-test is $F = [Q_H/df(Q_H)]/[Q_E/df(Q_E)]$, where $Q_H = (\mathbf{C}\hat{\boldsymbol{\beta}})'(\mathbf{C}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}})$, $Q_E = (\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{Z}\hat{\boldsymbol{\beta}})$, $\mathbf{Z}$ is the design matrix of the cell means model and $\hat{\boldsymbol{\beta}}$ is the usual est imate for the true cell means vector $\boldsymbol{\beta}$.

## 3 Implementation

By means of an example Wuethrich-Martone et al. (1999) have shown that the standard routines for the Analysis of Variance do not test causal effects in the nonorthogonal case. To date the only way to test causal effects using a statistical software is to use a routine for testing user-defined General Linear Hypotheses. Such routines usually require that the user specifies the coefficients matrix (for testing causal effects it would be matrix $\mathbf{C}$) and this may be a time-consuming procedure. To get a more versatile tool for the purposes of causal modeling, we implemented our method in SPSS as SCRIPT. With a simple menu-command it is now possible to examine a design with $p$ qualitative treatment variables, $q$ qualitative potential confounders and a continuous dependent variable from the perspective of causal modeling and hence to test for causal effects.
After reading the data (treatment variables, confounders and dependent variable) from an SPSS data file, the SCRIPT builds the vector $\mathbf{X}$ of treatments and the vector $\mathbf{W}$ of potential confounders. Without any loss of generality, the original $p \times q$ design is collapsed into a simpler one-treatment-one-confounder design and the corresponding data may be saved in a new SPSS data file.
The distribution of $\mathbf{W}$ may be computed from the cell frequencies or given a priori as a variable in the data file. The conditional expected values of

$Y$ adjusted for potential confounding w.r.t. $\mathbf{W}$ are then computed and the null hypothesis $ACE_{ii'} = 0 \; \forall i \forall i'$ is tested following the procedure in the former section. The corresponding F-value and p-value are computed.

## 4 Discussion

This paper examined the problem of controlling potential confounders in non-randomized experiments from the perspective of causal modeling. The basic idea is to test for average causal effects and to adjust for confounding. This is done adjusting the treatment means using the marginal probabilities of the confounders as weights. We presented a method for testing causal effects when qualitative treatments and qualitative confounders apply. This procedure is actually an extension of the methods presented by Wuethrich-Martone et al. (1999) for a 2-factors Analysis of Variance. We also implemented this procedure in SPSS as a practical tool for testing causal effects.

## References

Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association*, **81**, 945-970.

Neyman, J.S. (1923/1990). On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. *Statistical Science*, **4**, 465-480.

Rubin, D.B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, **66**, 688-701.

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle.* Stuttgart: Fisher Verlag

Steyer, R., Gabler, S., and Rucai, A.A. (1995). Individual Causal Effects, Average Causal Effects, and Unconfoundedness in Regression Models. In: *SoftStat '95*, 203-210.

Wuethrich-Martone, O., Steyer, R., Nachtigall, C., and Suhl, U. (1999). Causality, Confounding and Unbalanced Analysis of Variance. In: *Statistical Modelling. Proceedings of the 14th International Workshop on Statistical Modelling*, 719-722.

# Predicting Financial Risk by Qualitative Reasoning Techniques

Nuria Agell, Xari Rovira, Carmen Ansotegui[1], Monica Sanchez, and Francesc Prats[2]

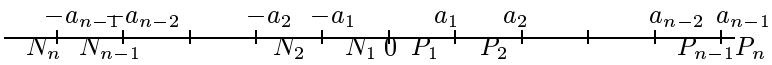
[1] Esade.Avda de Pedralbes,Barcelona
[2] Universitat Politecnica de Catalunya

**Abstract:** The present work is one step in a wider project in which AI techniques will be used to evaluate the probability of issuer default. The aim is to show how Qualitative Reasoning techniques, and particularly orders of magnitude calculus, can be useful in classifying the companies in accordance with their Moody's rating.

**Keywords:** Qualitative Reasoning, Orders of Magnitude, Qualitative Operators, Credit Risk.

## 1 Qualitative Operators defined on the Absolute Orders of Magnitude Model

The absolute orders of magnitude model (Piera 1995; Travé-Massuyès 1997) works with a finite set of qualitative labels obtained via a discretisation of the real line. The number of labels we choose for describing our reality depends on the characteristics of each problem (Agell 1998).The absolute orders of magnitude model with granularity $n$, $OM(n)$, is built via a real line symmetric partition in $2n+1$ classes:




Each class is named basic description and is represented by a label in the set $S_1$.

$$S_1 = \{N_n, N_{n-1}, N_{n-2}, \cdots, N_2, N_1, 0, P_1, P_2, \cdots, P_{n-2}, P_{n-1}, P_n\}$$

Note that all the variables are defined in the same orders of magnitude space (same granularity), although each one can have its own discretisation.

## 2 An application to Credit Risk Prediction: Rating

Moody's and Standard & Poor's classify firms according to their level of credit risk, using both quantitative and qualitative information to assign

ratings to debt. Moody's ratings are labelled Aaa, Aa, A, Baa, Ba, B, Caa, Ca, ranging from high to low credit quality.
The model presented is especially adequate when the goal is to measure the magnitude of a result, based on the qualitative descriptions of the variables that participate, and it is susceptible to adapt to qualitative variables as the industry. The qualitative descriptions appear when either numerical values are unknown or the experts use only their orders of magnitude. There are three main reasons that brought us to a qualitative approach:

- There are qualitative variables that determine the quality of risk.
- The quantitative variables involved have a proper description in qualitative terms. Often the orders of magnitude and tendencies of the variables are more relevant than their exact numeric values.
- The variables involved in credit risk have different relevance or strength in the global calculation. Qualitative operators are able to take into account different degrees of influence.
- The final classification has to be given in a qualitative set of labels.

### 2.1 Prediction Strategy

Let $V_1, \cdots, V_k$ be the qualitative variables defining the characteristics of a firm, and $V_1(C), \cdots, V_k(C)$ the qualitative values taken by these variables for a given firm C. Each one of these variables is qualitatively described via a set of labels, which are intervals of the real line, with an odd number of landmarks; provided by the experts.
It is first necessary to consider the values of $V_1, \cdots, V_k$ described in an $OM(n)$ with the same granularity, and homogenise them into a new common reference with landmarks in a set:

$$B = \{-b_{n-1}, \cdots, -b_1, 0, b_1, \cdots, b_{n-1}\}$$

Then, in order to obtain a synthesis value that reflects the credit quality of firm C, the operator $\Phi$ is applied to the former variables:

$$\Phi : OM(n) \times \cdots \times OM(n) \rightarrow OM(m)$$
$$\Phi(V_1 \cdots V_n) = \{\sum_{i=1}^{K} \alpha_i x_i / x_i \in V_i, \forall i = 1, \cdots, K\}$$

where, if $A \subset \Re$, $[A]$, denotes the minimum convex union of elements of $S_1$ that contains $A$.
The landmarks of the space $OM(m)$ will be $\{\sum_{i=1}^{K} \alpha_i b_i / b_i \in B\}$.
The last step involves taking a new discretisation of the real line in order to express the firm's rating in a qualitative space with eight basic labels, and

so emulate Moody's levels of credit risk. By applying a qualitative function from $OM(m)$ into a space $OM(4)$, which is the qualitative expression of the real identity, the final evaluation of the firm's credit risk is obtained.
The proposed method is currently being implemented. Our software tool accepts quantitative and qualitative data defined in an orders of magnitude structure. First the user must specify the set of landmarks for each variable. Because of the possibility of varying the granularity, every data structure in the program is dynamic. The result supplied by our algorithm is twofold; on the one hand, it provides the classification of the firms according to their credit quality, and, on the other hand, it also allows performance simulations to be carried out by modifying the values of the variables.

### 2.2 Example

This example shows the classification process in a first approach already implemented. As a first step, the approach will be applied using just five variables, given by accounting ratios, and the strength of their influences on the credit risk. The following table shows the ratios and the initial landmarks, which are provided by the experts and will be used to define the orders of magnitude.

| *Variable* | *Initial landmarks* |
|---|---|
| V1 = Leverage | 25, 50, 75 |
| V2 = Return on Capital | I, 2I, 5I |
| V3 = Interest Coverage | 0, 1, 3, |
| V4 = Cash-Flow to Total Debt | 0, 10, 60 |
| V5 = Market Value | 0.5, 1, 1.5 |

Each one of the variables the problem involves has its own description in qualitative terms. Starting from the previously given experts' qualitative descriptions, and in order to be able to apply a qualitative operator consistent with the real line to compose these variables, three streps will be taken: a tranlation to transform the central landmark $l_i$ to 0, a symetrization with respect to 0, and the extention of the sets to obtain the same cardinal for all of the variables. In this case, the five variables are all referenced in an $OM(3)$ with the same landmarks.
The homogenisation process steps applied to the variables give the successive landmarks. Finally, the operator must take into account the influence or strength of each ratio on credit risk. The final qualitative values for the credit quality will be given in an orders of magnitude space $OM(4)$ with granularity 4.
The software tool has been used for this example. We ran the test on eighty well-known American firms whose Moody's classification is available to the public.

## 3 Conclusion

This paper presents an on-going work, which provides the concepts and strategies for synthesising qualitative information from variables. The system is applied in the financial domain to evaluate and simulate credit risk. This approach may also be applicable to problems in other areas where the involved variables are described in terms of orders of magnitude, and the results depend on the given variables and a set of strengths. Some of the on-going tasks consist in:

1. Discovering alternative methods for building a homogenised reference.
2. Adjusting the weights of the qualitative weighted sum $\Phi$, using an historical set of data.

Our final words are to note that this work is only the first experiment with a new and simple idea, though one we are convinced is promising: the idea of adapting qualitative references to the operators involved in a real problem.

## References

Agell, N. (1998). *Estructures matemàtiques per al model qualitatiu de órdenes de magnitud absoluts.* Ph. D. Thesis Universitat Politècnica de Catalunya, Barcelona.

Piera, N. (1995). *Current Trends in Qualitative Reasoning and Applications.* Monografía CIMNE, 33. International Center for Numerical Methods in Engineering, Barcelona.

Travé-Massuyès, L. Dague, Ph and Guerrin, F. (1997). Le Raisonnement Qualitatif pour les Sciences de l'Ingenieur, Ed. Hermès, Paris.

# Selection Criteria for Non Nested Binary Choice Models: A Comparative Study

Teresa Aparicio and Inmaculada Villanúa[1]

[1] Dep. of Economic Analysis,
University of Zaragoza,
Gran Vía, 2, Zaragoza, Spain

**Abstract:** This paper focuses on various model selection criteria in a framework of binary choice econometric models, specifically that of probit and logit models. Basically, we try to analyze the performance of the criteria which we have available to discriminate from among alternative specifications. Such study is developed both theoretically and through a simulation experiment.

**Keywords:** binary choice model, model selection procedures, hypothesis testing procedure, Monte Carlo exercise, consistency.

## 1 Introduction

The objective of model selection is to choose the best specification, according to some given criterion, among a set of models that have been tested in the so-called validation stage. This stage tries to guarantee that valid inferences can be obtained. In the assumed framework of non-nested binary choice models, we consider two situations, namely strictly non-nested models, and overlapping models, following the division of Vuong (1989).
Two models are strictly non nested when there is no common conditional distribution. This condition is satisfied in two cases: (i) if we compare a logit with a probit model, even if all explanatory variables of each model are common, and (ii) if we compare two logit or two probit models where the variables of each model are all specific. Two competing models are overlapping if both are logit(probit), with some common explanatory variables.
We adopt the traditional classification which divides the criteria into "model selection procedures" and "hypothesis testing procedures".
In the first category we consider the following criteria: $AIC$ (Akaike, 1973), $SBIC$ (Schwarz, 1978) and two supplementary criteria, which we denote as $C_2$ and $C_3$. All these criteria are derived in the context of Decision Theory, by using the discrepancy concept (see Linhart and Zucchini, 1986, and Lavergne, 1998), or a loss function in a more general approach.
The second category corresponds to the techniques based on hypothesis testing strategy, and in this class we include the works of Vuong (1989), Pesaran and Pesaran (1993) and Santos Silva (1996).

## 2 The behaviour of the selection criteria

Let us consider $M_1$ and $M_2$, two competing models that we define as $M_1 : p_i = F_1(X_i^T\beta)$ and $M_2 : p_i = F_2(Z_i^T\alpha)$.
When $F_1 = F_2 = F$ (that is to say, both logit or both probit models), and the vectors $X_i$ and $Z_i$ contain some common variables, the models are overlapping. However, if $F_1 = F_2 = F$ and there are only specific variables in each model, or $F_1 \neq F_2$ whatever the kind of explanatory variables (common or specific), then the models are non-nested.
The $AIC(M_j)$ criterion adopts the form:

$$-\frac{1}{N}\sum_{i=1}^{N}[y_i \ln \hat{F}_{ji} + (1-y_i)\ln(1-\hat{F}_{ji})] + \frac{k_j}{N} \tag{1}$$

where $y_i$ is the binary response variable, $k_j$ is the number of parameters of the $M_j$ model, for j=1,2, and $\hat{F}_{ji}$ denotes the distribution function corresponding to $M_j$ model evaluated at $X_i^T\hat{\beta}$ or $Z_i^T\hat{\alpha}$, being $\hat{\beta}$ and $\hat{\alpha}$ the corresponding maximum likelihood estimates. The first term of the $SBIC(M_j)$ criterion is the same first term as that of (1), and the second term is $\frac{k_j lnN}{2N}$. Denoting $\hat{u}_{ji} = (y_i - \hat{F}_{ji})$, we can write $C_2(M_j) = \sum_{i=1}^{N}\hat{u}_{ji}^2(\frac{1}{N} + \frac{2k_j}{N^2})$ and $C_3(M_j) = \frac{1}{m}\sum_{i=N_1+1}^{N_1+m}(\hat{u}_{ji}^{(i-1)})^2$, with $\hat{u}_{ji}^{(i-1)}$ indicating that the estimation is obtained with a sample size $(i-1)$, being $N_1$ the initial sample size. We consider that $N_1$ and $m$ are fixed proportion of N. The AIC and SBIC criteria are explicitly derived in Linhart and Zucchini (1986), and the derivation of $C_2$ and $C_3$ can be found in Villanúa (1997).
We want to analyze the asymptotic behaviour of $P(M_1) - P(M_2)$ where $P(\cdot)$ denotes a selection criterion, evaluated for each particular model. Assuming $M_1$ or $M_2$ as the data generating process (DGP), the sign of $plim(P(M_1) - P(M_2))$ establishes the consistency or inconsistency of each criterion. Therefore, we understand that consistency is the requirement of a criterion in order for it to be considered as adequate. Without loss of generality, we assume $M_2$ the DGP, and then we obtain the specific form of $P(M_1) - P(M_2)$ for the four criteria considered, using Taylor expansions around the true parameter vector $(\alpha)$. For the AIC criterion, such expression $(AIC(M_1) - AIC(M_2))$ is given by:

$$\frac{1}{N}\sum_{i=1}^{N}\{y_i(\ln\frac{F_1(Z_i^T\alpha)}{F_2(Z_i^T\alpha)} + \frac{f_1(Z_i^T\alpha)}{F_1(Z_i^T\alpha)}(X_i^T\hat{\beta} - Z_i^T\alpha) - \frac{f_2(Z_i^T\alpha)}{F_2(Z_i^T\alpha)}(Z_i^T\hat{\alpha} - Z_i^T\alpha)) +$$

$$(1-y_i)(\ln\frac{1-F_1(Z_i^T\alpha)}{1-F_2(Z_i^T\alpha)} + \frac{f_2(Z_i^T\alpha)}{1-F_2(Z_i^T\alpha)}(Z_i^T\hat{\alpha} - Z_i^T\alpha))$$

$$-(1-y_i)(\frac{f_1(Z_i^T\alpha)}{1-F_1(Z_i^T\alpha)}(X_i^T\hat{\beta} - Z_i^T\alpha))\} + A \tag{2}$$

where $f_j$ denotes the correspondig density function, and $A = \frac{k_1 - k_2}{N}$.
The form of $P(M1) - P(M2)$ expression for the SBIC criterion is the same as that of AIC criterion given in (4), adding $\frac{\ln N}{2}$ in the last term. For the $C_2$ and $C_3$ criteria, such expression can be written, respectively, as:

$$\frac{1}{N}\sum_{i=1}^{N}\{(u_{1i}^2 - u_{2i}^2) - 2[u_{1i}f_{1i}(X_i^T\hat{\beta} - Z_i^T\alpha) - u_{2i}f_{2i}(Z_i^T\hat{\alpha} - Z_i^T\alpha)]$$

$$+\frac{2}{N}\left(\frac{k1\sum_{i=1}^{N}\hat{u}_{1i}^2 - k_2\sum_{i=1}^{N}\hat{u}_{2i}^2}{N}\right) \tag{3}$$

and

$$\frac{1}{m}\sum_{i=N_1+1}^{N_1+m}\{(u_{1i}^2 - u_{2i}^2) - 2[u_{1i}f_{1i}(X_i^T\hat{\beta}^{(i-1)} - Z_i^T\alpha) - u_{2i}f_{2i}(Z_i^T\hat{\alpha}^{(i-1)} - Z_i^T\alpha)]\} \tag{4}$$

being $u_{ji} = (y_i - F_j(Z_i^T\alpha)$, and $f_{ji}$ the corresponding density functions evaluated at $Z_i^T\alpha$.
To study the asymptotic behaviour of these four criteria, the following results hold in both overlapping and non nested models: (i) $plim(\hat{\beta}) = \beta^*$; (ii) $plim(\hat{\alpha}) = \alpha$; (iii) $\beta^* \neq \alpha$, being $\alpha$ and $\beta^*$ the true and pseudo-true parameter vectors, respectively.
When we have overlapping models or non-nested models with $F_1 = F_2 = F$ and all the explanatory variables being specific, we conclude that the probability limit of $P(M_1) - P(M_2))$ is possitive, which means that all the criteria considered are consistent. To obtain this conclusion, we use the relations: 1) $y_i = 1 \Rightarrow F_1(X_i^T\beta^*) < F_2(Z_i^T\alpha)$, and 2) $y_i = 0 \Rightarrow F_1(X_i^T\beta^*) > F_2(Z_i^T\alpha)$.
In the same framework of non-nested models, but with $F_1 \neq F_2$, we cannot obtain a conclusion, because we can not assure the maintenance, for every observation, of relations 1) and 2) previously defined, and therefore $plim(P(M_1) - P(M_2))$ contains various terms with opposite signs, being difficult to determine their specific weights. So in this situation we can not derive the theoretical behaviour of the criteria.
We have carried-out a Monte Carlo experiment, that allows us both to solve the indeterminate result obtained in non nested models with $F_1 \neq F_2$, and to compare, assuming several sample sizes, the selection procedures with the hypothesis testing procedures mentioned in section 1. The main results obtained from this exercise are:

a) Overlapping models and non nested models with $F_1 = F_2$. Whatever the $F(\cdot)$ function (logit/probit), all the criteria tend to choose the DGP even with the smallest sample size considered (N=200). The

Santos Silva criteria (based on a LM statistic) is the only one which presents a quite different pattern.

b) Non-nested models with $F_1 \neq F_2$. According to the theoretical indetermined result obtained for this kind of models, the experiment reveals that the specific " model selection procedures"(AIC, SBIC, $C_2$ and $C_3$) and the Vuong statistic, tend to choose the DGP unless $Z \subset X$. However, in this last situation the SBIC criteria behaves better than the others. Whatever the relation between $X$ and $Z$, the criteria of Santos Silva and Pesaran and Pesaran behaves poorly.

## 3 Conclusions

The set of results obtained both empirical and theoretically, appears to show evidence that in the case a) of the last section, the " model selection procedures" behave better than the " hypothesis testing procedures". Nevertheless, one of these criteria (Vuong (1989)) presents an appropriate pattern. In the situations contained in b), all the " model selection procedures" and the Vuong criterion exhibit a different pattern depending on the relation between the variables included in $X$ and $Z$; may be the SBIC criteria presents a less marked differencies. Therefore, the SBIC criteria seems to be the most adequate to select among a set of non-nested (or overlapping) binary models.

## References

Akaike, Y. (1973). Information theory and an Extension of the Likelihood Ratio Principle. In: *Proceedings of the second international Symposium of Information Theory*, 189-204, Budapest: Akademiai Kiado.

Linhart, H. and Zucchini, W. (1986). *Model Selection.* New York: John Wiley and Sons .

Pesaran, M.H. and Pesaran, B. (1993). A Simulation Approach to the Problem of Computing Cox's Statistic for Testing Non Nested Models. *Journal of Econometrics*, **57**, 377-392.

Santos Silva, J.M.C. (1996). A Score Test For Non-Nested Hypotheses with Applications to Discrete Data Models. In: *Econometric Society European Meeting*, Istanbul, Turkey.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464.

Villanúa, I. (1997). Proceso de Validación y Selección en Modelos de Elección Discreta: el Caso Dicotómico. *Tesis doctoral. Universidad de Zaragoza.*

Vuong, Q.H. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, **57**, 307-333.

# A note on Successive Sampling using Auxiliary Information

Eva Artés Rodríguez[1] and Amelia V. García Luengo[2]

[1] Departament of Statistics and Applied Mathematics, University of Almería, Spain e-mail: eartes@ualm.es
[2] amgarcia@ualm.es

**Abstract:** The problem of estimation of a finite population mean for the current occasion based on the samples selected over two occasions has been considered. For the case when two auxiliary variables are negatively correlated with the main variable, a double–sampling multivariate product estimate from the matched portion of the sample is presented. Expressions for optimum estimator and its variance have been derived. The gain in efficiency of the combined estimate over the direct estimate using no information gathered on the first occasion is computed.

**Keywords:** Successive Sampling; Bivariate Product Estimator; Gain in Efficiency; Matching fraction.

## 1 Introduction

Some of the reasons which explain that partial replacement of sample units should be used are:

1. It reduces costs (using totally new samples at each time can be unduly expensive).
2. It increases the estimators' accuracy.
3. The indefinite presence of the same units in the sample can result in failures and efficiency reduction of the estimators.

In order to cover a wide range of practical situations, this paper focuses on the development of the theory on successive sampling, aiming to build the optimum estimator of the mean at the second occasion, by using a double sampling multivariate product estimator for the matched part of the sampling, and a simple mean based on the unmatched part of the sample on the second occasion. We have used two auxiliary variables, $x_1$ and $x_2$ (negatively related to the study variable), as they are the most frequently applied.

## 2 Development of the Method.

Suppose that the samples are of size $n$ on both occasions, we use a simple random sampling and the size of the population $N$ is sufficiently great for the factor of correction be ignored.
Let a simple random sample of size $n$ be selected on the first occasion from a universe of size $N$. When selecting the second sample, we assume that $n - u = m$ of the units of the selected sample on the first occasion are retained for the second occasion (matched sample) and the remaining $u$ units are replaced by a new selection from the universe $N - m$ left after omitting the $m$ units.
Information about both auxiliary variables $x_1$ and $x_2$ is available for the first occasion, whose means are denoted $\bar{x}_1$ and $\bar{x}_2$, respectively. Let $y$ be the variable under study on the second occasion, and we suppose that is negatively correlated with $x_1$ and $x_2$. Let:, $\bar{x}_1^m$,$\bar{x}_2^m(\bar{y}_m)$ = matched sample mean on the first (second) occasion estimating $\bar{X}_1$,$\bar{X}_2(\bar{Y})$; $p = \frac{m}{n}$, the matching fraction. , $\bar{y}_u$ = unmatched sample mean on the second occasion estimating $\bar{Y}$, $C_0 = \frac{S_y}{\bar{Y}}$ $C_i = \frac{S_{x_i}}{\bar{X}_i}$ $\Delta_i = \frac{C_i}{C_0}$ $i = 1, 2$, $\rho_{01}$ = *Pearson* correlation coefficient is between $x_1$ y $y$ , $\rho_{02}$ = *Pearson* correlation coefficient is between $x_2$ y $y$ , $\rho_{12}$ = *Pearson* correlation coefficient is between $x_1$ y $x_2$ , The unmatched ($u$ units) and matched ($m$ units) portions of the second occasion sample provide independent estimates ($\bar{y}_m$ and $\bar{y}_u$) of the population mean on the second occasion $\bar{Y}$. For the matched portion an estimate improved of $\bar{Y}$ may be obtained using a double sampling multivariate product estimate

$$\bar{y}'_m = \omega_1 \frac{\bar{x}_1^m}{\bar{x}_1} \bar{y}_m + \omega_2 \frac{\bar{x}_2^m}{\bar{x}_2} \bar{y}_m$$

If $W = (\omega_1, \omega_2)$ is defined, we obtain

$$V\left(\bar{y}'_m\right) = \bar{Y}^2 W D W' \tag{1}$$

where $D = (d_{ij})$ , is the matrix defined by

$$d_{ij} = \frac{1}{m} C_0^2 + \left(\frac{1}{m} - \frac{1}{n}\right)\left(C_i C_j \rho_{ij} + C_0 C_i \rho_{0i} + C_0 C_j \rho_{0j}\right) \quad i, j = 1, 2$$

Obtain the minimum variance of the estimator

$$V\left(\bar{y}'_m\right) = \bar{Y}^2 \hat{W} D \hat{W}'$$

where, the optimum weighting vector given by

$$\hat{W} = \frac{eD^{-1}}{eD^{-1}e'}$$

with $e = (1,1)$ and $D^{-1}$ is the inverse matrix of $D$.
Hence, we construct an estimate of the mean of the population on the second occasion, $\bar{Y}$, by combining the two independent estimates, $\bar{y}'_m$ and $\bar{y}_u$ with weights $\omega$ and $(1-\omega)$. Thus

$$\bar{y}_{2PM} = \omega \bar{y}'_m + (1-\omega)\bar{y}_u$$

$$V_{min}\left(\bar{y}_{2PM}\right) = \frac{S_y^2}{n}\frac{1+qZ}{1+q^2Z} \tag{2}$$

where $Z = \Delta\left(2\rho_0 + \Delta\frac{1+\rho}{2}\right)$.
Minimizing in (2), we have the optimum matching fraction 0

$$p_{opt} = \frac{1+Z-\sqrt{1+Z}}{Z}$$

The gain in precision $G$ of the combined estimate $\bar{y}_{2PM}$, obtained by using a double-sampling multivariate product estimate from the matched portion of the sample on the second occasion, over the direct estimate, is given by

$$G = \frac{V\left(\bar{y}\right) - V\left(\bar{y}_{2PM}\right)}{V\left(\bar{y}_{2PM}\right)} = \frac{-Zp\left(1-p\right)}{1+\left(1-p\right)Z}$$

where $Z = \Delta\left(2\rho_0 + \Delta\frac{1+\rho}{2}\right)$ y $V\left(\bar{y}\right) = \frac{S_y^2}{n}$.
Necessarily $p \leq 1$. If $p = 1$ (perfect matching) or $p = 0$ (no matching), the gain is zero. For other $p$ $(0 < p < 1)$, there will be positive gain if

$$\frac{2\rho_0}{1+\rho} < -\frac{1}{2}\left(\frac{C}{C_0}\right)$$

Further, we conclude that the gain in precision of the combined estimate, $\bar{y}_{2PM}$, over the direct estimate, $\bar{y}$, increase with increasing $\rho_0$ absolute value (larger dependence between the auxiliary variables $x_1$ and $x_2$ with the variable under study $y$), and decreasing $\rho$ (smaller correlation between $x_1$ and $x_2$).

## 3 Empirical Study.

We have used the data collected in a survey on healthy habits and fitness level to assess the optimal operation of the proposed method. This study was carried out over a population of fourteen-year-old schoolchildren in Almeria schools during April and June, 1998. We have intended to develop a sampling scheme that provides us with more accurate estimators of the studied variables.

In order to achieve the targets of the study, we have considered the estimation of the endomorphic component ($\bar{y}$, one of the multiple variables which affect the survey) at the second occasion, taking as auxiliary variables the arm maintained flexion ($\bar{x}_1$) and the maximum volume of oxygen ($\bar{x}_2$) from the first occasion. The estimation procedure was performed by combining the estimators for the two independent samples of schoolchildren: $\bar{y}'_m$ and $\bar{y_u}$.

Moreover, we have compared the accuracy of the proposed estimator with other indirect estimators. Table 1 shows the results. As we can see, the ratio method is not efficient when the auxiliary variables are negatively correlated to the principal variable $y$, as the gain in accuracy over $\bar{y}$ is negative (G = -18.03%). However, the combined estimator based upon a bivariant product estimator for the matched part of the sample and a simple sample mean of the unmatched part, $\bar{y}_{2PM}$, is more accurate than the correspondent estimator which makes use of a univariant product estimator for the matched sample, $\bar{y}_{2p}$, and it even improves the accuracy of the one which makes use of regression estimator for the matched part, $\bar{y}_{2reg}$. The last column shows the efficiency gained from the different estimators, regarding $\bar{y}$.

**TABLE 1**

| Estimators | Auxiliary Variables | Variance | % Gain in precision |
|---|---|---|---|
| 1. Direct $\bar{y}$ | none | $\frac{S_y^2}{n}$ | |
| 2. Univariate Product $\bar{y}_{2p}$ | $x_1$ | $0.94\frac{S_y^2}{n}$ | 6.38% |
| 3. Bivariate Ratio $\bar{y}_{2RM}$ | $x_1$ and $x_2$ | $1.22\frac{S_y^2}{n}$ | $-18.03\%$ |
| 4. Bivariate Regression $\bar{y}_{2reg}$ | $x_1$ and $x_2$ | $0.88\frac{S_y^2}{n}$ | 13.62% |
| 5. Bivariate Product $\bar{y}_{2PM}$ | $x_1$ and $x_2$ | $0.87\frac{S_y^2}{n}$ | 14.53% |

**Acknowledgements:** Special Thanks to Antonio Jesús Casimiro Andújar for providing the data of the empirical study.

## References

Artés, E., Rueda, M. y Arcos, A. (1998). Successive Sampling using a Product Estimate, *Applied Sciences and the Environment, Computational Mechanics Publications*, 85–90.

Singh, M.P. (1967). Multivariate Product Method of Estimation for Finite Populations, *Journal of the Indian Society Agricultural Statistics*, **19** (2), 1–10.

# A Choice of Software for Measuring and Correcting for Unobserved Heterogeneity.

S A. Ayis

[1] Departament of Epidemiology and Public Health,
University of Newcastle Upon Tyne, Medical School,
NE2 4HH, Newcastle Upon Tyne; e-mail S.A.M.Ayis@newcastle.ac.uk

**Abstract:** Variables that influence the reponse by any degree but not accounted for using Standard Generalised Linear Regression Models may lead to parameter estimates that are seriously biased. Estimated standard errors may also be affected. A similar effect may result form the correlation due to repeated observations. Corrective methods that deal with such situations are available. The estimation provided by some of these is here investigated using simulations. The software "SABRE" is used for computation. Situations where the estimations is affected are identified and alternative methods and/or software are used.

**Keywords:** Unobserved Heterogeneity (U-H), Logistic Model, Logistic Normal Model, ML3.

## 1 Introduction

The standard logistic model is widely used in the analisys of contingency tables when the outcome of interest is dichotomus and one or more categorical variables are used as explanatory variables. The model may be defined as;

$$Logit(p_j) = Log\left[\frac{p_j}{1-p_j}\right] = x_j'\beta$$

where $p_j = p(y_j = 1|x_j)$, $x_j$ and $y_j$ are vectors of explanatory variables and outcome respectively for the j'th subject, $j = 1, 2, \ldots, n$. $\beta$ is the vector of parameters to be estimated. The model is however, invariabily inadequately specified. In contingency tables where, subjects are clasiffied into cells according to an observed category for each of the explanatory variables, omission of variables or variable levels for any reason affects the identical independent distribution (i.i.d) assumption, that is because, the probability of succes $p_{ij}$ within a cell is not contant but varies systematically with other factors which are omitted, this is unobserved heterogeneity, hereafter will be refered to as U-H. , Using standard models designed for observations that are(i.i.d) in the presence of U-H leads to estimates that

are biased and confidence intervals coverage properties that are misleading, however, models that deal with such
situations are available. In this study the effect of U-H on the performance of the standard logistic model is investigated. In addition the performance of appropriate models that correct for U-H is explored systematically under various conditions using simulation. Two software are employed to carry out the investigations. The focus of the study is on the following issues;

1- to identify circumstances where the bias is more pronounced.

2- to investigate corrective solutions, are any better than others?.

3- to explore circumstances where the solutions are more effective.

To address these issues, The standard logistic model is first used to fit simulated data with known effect of U-H. The logistic normal model of SABRE is then used as an appropriate corrective approach. A variety of situations are covered, these include: (1) different values of the probabilities $(p_0, p_1)$, (2) different sizes of U-H variance.and (3) different combinations of clusters per sample J and replications per cluster K., All three conditions were investigated one at a time by keeping two conditions constant and allowing the third one to vary. The aim is generally to explore situations where the U-H has the biggest impact on estimates and their confidence intervals.,

ML3(Multilevel Models 3) is used as an alternative approach when estimates of SABRE are affected. The results are very limited as only one dichotomous explanatory variable is used. These however, allow us to infer more general conclusions as the extension to contingency table in general ought to be straight forward. Tables 1 highlights some of the results.

***Table 1. A comparison between two models for the estimation of two level model for two structures of probabilities and different designs of samples***

| *Method* | *SABRE* | | *ML3* | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{\beta}$ | $\hat{\sigma}$ | $\hat{\beta}$ | $\hat{\sigma}$ | $(p_0, p_1)$ | K | J | $\sigma$ |
| estimate | 1.02 | 0.97 | 1.03 | 0.99 | | | | |
| se | 0.27 | 0.15 | 0.27 | 0.25 | (0.62,0.82) | 8 | 96 | 1.0 |
| sd | 0.26 | 0.15 | 0.26 | 0.32 | | | | |
| | | | | | | | | |
| estimate | 1.03 | 0.97 | 1.00 | 1.04 | | 32 | 24 | |
| se | 0.35 | 0.17 | 0.46 | 0.38 | | | | |
| sd | 0.53 | 0.22 | 0.44 | 0.40 | | | | |
| | | | | | | | | |
| estimate | 2.01 | 0.98 | 1.94 | 0.97 | (0.27,0.73) | 8 | 96 | |
| se | 0.28 | 0.15 | 0.26 | 0.24 | | | | |
| sd | 0.28 | 0.15 | 0.27 | 0.26 | | | | |
| | | | | | | | | |
| estimate | 2.10 | 0.96 | 2.10 | 1.02 | | | | |
| se | 0.38 | 0.17 | 0.45 | 0.36 | | | | |
| sd | 0.57 | 0.22 | 0.44 | 0.34 | | 32 | 24 | |

## 2 Conclusions

1- The Logistic-Normal model of SABRE;

(i) the estimation of the parameter $\beta$ is good for all the range of structures of probabilities investigated, the range of variance of U-H considered and for all combinations of j and k. An exceptional case however, is when the variance of U-H is large, the two probabilities lie further apart on the two sides of the logistic curve, the number of replications k is large and the number of clusters j is small. The cause is likely to be the break down of the asymptotic theory.

(ii) The model performs fairly well to estimate $\beta$ whether the covariates are time varying or time invariant.

(iii) the estimation of $\sigma$ is good when the sample size is large, a downward bias is however, reported for small k and or small j. The estimation of s in general is slightly better when the covariates are time varying.

(iv) at situations where there is effectively no unobserved heterogeneity the model over corrects $\sigma$.

(v) the estimation of the standard error of $\beta$ and $\sigma$ is good except where the number of replications is very large, number of clusters relatively small and the two probabilities lies further apart. At such situations the standard error is under estimated.

2- ML3 in comparison with SABRE method: Comparisons cover two structures of probabilities and two values of U-H variance these includes Situations where SABRE is Appropriate as well as where it is affected, results suggested that at situations where SABRE method underestimates the standard errors (v) above, the ML3(1995) two level logistict model provides a solution.

3- Suggested Solutions Using SABRE: Results of SABRE suggests that, the increase in precision achieved by increasing the number of replications is not particularly large. For such situations by cutting down the number of replications, a better estimation of the parameters especially the variances can be achieved without much loss of precision to the estimates of the fixed parameters themselves. Thus if SABRE is to be used, this approach may overcome the estimation problems for variances and random terms without sacrificing much efficiency for the estimation of the fixed parameters . To improve the estimation further and to use all the information available the data could be split into replicates satisfying the above conditions and then pooled estimates obtained from the replicates. From the other results it appears that we can safely appeal to the modified two level logistic

model of ML3(1994) to provide correct estimation of the parameters for small as well as large number of replications and for the range of probabilities considered.

## Acknowledgment

I am thankful to professor D'Holt who supervised my PhD on Modelling Unobserved Heterogeneity from which this paper is a part, and to Dr. Marie South for her help in programming with FORTRAN.

## References

Barry, J., B. Francis and R.Davies (1990). SABRE :Software for the Analysis of Binary Recurrent Events. A guide for Users. Centre for Applied Statistics, Lancaster University.

Goldstein, H. (1994). Improved Estimation for Logit and Loglinear Multilevel Models. Multilevel Modelling Newsletter. Vol., 6, No.1, pp 2.

# $L_2$-Tests with Fixed Kernel for Specification of Parametric Models

Knut Bartels[1]

[1] University of Potsdam, Faculty of Economics and Social Sciences, Chair for statistics and econometrics, August-Bebel-Str. 89, D-14482 Potsdam, Germany, phone: +49-331-977-3812, fax: +49-331-977-3210, bartels@rz.uni-potsdam.de

**Abstract:** In contrary to many approaches to specification testing of parametric models here test are considered with a fixed kernel, respectively bandwidth. This leads to nonnormal and case dependent limiting distributions under the null hypothesis, but critical values can be derived by resampling methods. A simplification of the wild bootstrap is constructed that substantially reduces the amount of computer time needed for testing nonlinear models.
The consideration of fixed kernels provides some insights that are disguised if the test statistics are subject to an asymptotically vanishing bandwidth. As a key result the consistency of the tests depends on the kernel function.

**Keywords:** model specification; nonlinear models; resampling methods; simulation studies.

## 1 Introductory Remark

This short article is a summary of the results in Bartels(1999). Due to the limited space, the presentation is very concise and no proofs can be given.

## 2 Test Problem

The problem of testing the specification of a parametric model

$$\mathsf{E}[Y|X=x] = f(x,\theta_0) \text{ for } (Y,X) \sim \mathsf{D}$$

and a known function $f : I\!R^d \times \Theta_0 \to I\!R$ is considered. For a measurable function $g : I\!R^d \to I\!R$ let $\mathsf{D}\{g\}$ be the set of all distributions $\mathsf{D}$ on $I\!R \times I\!R^d$ such that for $Z = (Y,X) \sim \mathsf{D}$ the variance $\mathsf{Var}[Y]$ exists and $\mathsf{P}\big\{\mathsf{E}[Y|X] = g(X)\big\} = 1$ holds, where the probability is taken with respect to the marginal distribution $\mathsf{D}_X$. The test problem then is

$$\mathbf{H_0}: \quad \mathsf{D} \in \mathcal{D}_0 := \bigcup_{\theta\in\Theta_0} \mathsf{D}\{f(\cdot,\theta)\}$$

versus

$$\mathbf{H_1}: \quad \mathsf{D} \in \mathcal{D}_1 := \bigcup_{g \in \mathcal{B}(\boldsymbol{R}^d, \boldsymbol{R})} \mathsf{D}\{g\} \setminus \mathcal{D}_0 \quad ,$$

where the union is taken over all Borel-measurable functions $g : \mathbb{R}^d \to \mathbb{R}$. For a sample $Z_1, \ldots, Z_n$ the tests statistics

$$\hat{T}_n = n^{-1} \sum_{1 \le i < j \le n} \hat{U}_i \hat{U}_j K_{ijn} \quad \text{and} \quad \hat{T}_n^{(v)} = n^{-1} \sum_{1 \le i,j \le n} \hat{U}_i \hat{U}_j K_{ijn} \qquad (1)$$

are employed, where $\hat{U}_i := u(Y_i, X_i, \hat{\theta}_n) = Y_i - f(X_i, \hat{\theta}_n)$ denote the parametrically estimated residuals with respect to $\hat{\theta}_n = \hat{\theta}_n(Z_1, \ldots, Z_n)$. $\hat{K}_{ijn} := k_n(X_i, X_j)$ are weights with a symmetric kernel $k_n = k_{\{Z_1, \ldots, Z_n\}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, that may also depend on the sample.
The $U$- and $V$-type statistics (1) are a general form for $L_2$-tests. They were considered explicitly by Zheng(1996), but many other tests statistics found in the literature can be transformed to one of the statistics in (1), e.g. those of Bierens(1990), Härdle and Mammen(1993) or Stute(1997).

## 3 Theoretical results

Here the special case of a *fixed kernel* is considered, that is $k_n(x_1, x_2) = \kappa(x_1, x_2)$ for some bounded function $\kappa$, such that $k_n$ does not depend on the sample. This means especially that the test statistic must not depend on a variable bandwidth $h = h_n \to 0$ for $n \to \infty$.
The latter is often assumed, since it leads to a normal limiting distribution under some moment condition. But it is known that this normal limit is approached very slowly, such that it is almost of no practical use. Instead bootstrap methods are used, that are shown to work (if at all) using the same slow convergence to a normal limit; e.g. Härdle and Mammen(1993). Considering the tests with fixed kernel leads to a better explanation, why the bootstrap methods are superior to the normal approximation.
For fixed kernels under $\mathbf{H}_0$ and certain weak regularity conditions the $U$-type statistic in (Refe:Tn) has an asymptotic distribution as

$$\mathcal{L}\Big( \gamma + \sum_{j=1}^{\infty} \lambda_j (\chi_{1j}^2 - 1) \Big)$$

where $\gamma \in \mathbb{R}$ is a constant, $\chi_{11}^2, \chi_{12}^2, \ldots$ are independent $\chi_1^2$-distributed random variables and $\lambda_j$ are the eigenvalues of some linear operator $\mathcal{Q} = \mathcal{Q}_{\kappa, \mathsf{D}, \theta_0} : g(\cdot) \mapsto \int_{\boldsymbol{R}^{d+1}} q(\cdot, t, \theta_0) g(t) \mathrm{d}\mathsf{D}(t)$. This can be proven using a limit law for degenerate $U$-statistics by Gregory(1977); details are found in Bartels(1999).

Under $\mathbf{H}_1$ the statistics (Refe:Tn) are not degenerate and thus asymptotically distributed normal with respect to a higher order and an expectation that is determined by the deviation $\Delta(x,\theta) = \mathsf{E}[Y|X=x] - f(x,\theta)$ from the model. If this deviation lies in the sum of the eigenspaces belonging to positive eigenvalues of the kernel operator $\mathcal{K}_\theta : g(\cdot) \mapsto \int_{\mathbb{R}^d} \kappa(\cdot,t,\theta) g(t) \mathrm{d}\mathsf{D}_X(t)$, then the tests are consistent against this alternative.
Thus the tests are generally consistent, if the operator induced by $\kappa$ is positive definite. A symmetric kernel $\kappa$ is positive definite, if his Fourier transform $\bar{\kappa}$ is strictly positive. For example the Gaussian kernel $\kappa(x_1,x_2) = \exp(-(x_1-x_2)^2)$ is thus shown to be positive definite, while kernels with compact support are at best positive semidefinite - and do not lead to consistent tests with fixed kernels.
Since this limiting distribution is case dependent and cannot be tabulated, the critical values $\tau_{\alpha n}$ for the tests must be determined by resampling methods. Using a theorem of Dehling and Mikosch(1994) on the bootstrap of degenerate $U$-statistics, resampling methods for approximating the critical values of the tests can be derived. These methods include the wild bootstrap, but also suggest a simpler method, a Monte-Carlo-approximation that can be viewed a linear approximation to the wild bootstrap. Thus for testing linear models and using the least squares estimator the Monte-Carlo-Approximation and the wild bootstrap produce identical results. For nonlinear models the Monte-Carlo-approximation substantially reduces the amount of computer time needed, since the calculation of an estimator for each resample is avoided.
Although considering the tests asymptotically with a fixed kernel, the behavior of tests with fixed kernels belonging to a family $\{\kappa_h \,|\, \kappa_h(x_1,x_2) = h^{-d}\kappa(\frac{x_1-x_2}{h}), h \in \mathbb{R}_{>0}\}$ that is derived by applying different bandwidths to a general kernel $\kappa$ is necessary. The change of bandwidth has only the effect of a change of scale in the Fourier analysis and therefore is irrelevant for asymptotic properties.
For finite samples, of course, the choice of the fixed bandwidth has an effect. This is illustrated in an approximative formula for the power $\Gamma$ in the case $f : [0,1] \times \Theta_0 \to \mathbb{R}$:

$$\Gamma_n(\mathsf{D}_j) \cong \mathsf{P}\left\{ nc^2\sqrt{\mathbf{h}} \frac{\bar{\kappa}(j\mathbf{h}\pi)}{\sqrt{2\sum_{j=0}^{\infty} \bar{\kappa}(j\pi)^2}} + O_p(n^{\frac{1}{2}}) > \hat{\tau}^*_{\alpha n}(1,1) \right\}.$$

Here $\mathsf{D}_j$ denotes a distribution belonging to a deviation $\Delta(x,\theta_0) = c \cdot \cos(jx)$ from the model $f$, and $\hat{\tau}^*_{\alpha n}(1,1)$ denotes the critical value derived by resampling methods for bandwidth $h = 1$ and standardized errors. It thus can be seen how very large and very small bandwidths lead to tests with low power.
All theory can easily be extended to multidimensional ($Y \in \mathbb{R}^c, c \geq 2$) and index models. This, for example, makes it possible to test the specification of multinomial logit models. An application to marketing data of product

choice is presented in Bartels et al.(1999).

## 4 Simulation studies

The same models as in Rodrigues-Campos et al.(1998) and Stute et al. (1998) have been considered for a comparison of the different tests. Results in general coincide, while the kernel based tests provide more flexibility. Further a nonlinear growth model has been studied in simulations. This was achievable in reasonable time only by the new Monte-Carlo-approximation of the critical values. Comparative lists of results, graphics and extensive interpretations are found in Bartels(1999). An implementation of the test procedures is available in XploRe (http://www.xplore-stat.de).

## References

Bartels, K. (1999). Tests zur Modellspezifikation in der nichtlinearen Regression. *Dissertation Universität Potsdam.*

Bartels, K., Boztug, Y., and Müller, M. (1999). Testing the multinomial logit model. *Discussion paper* **19**, *SFB 373, Humboldt Universität zu Berlin.*

Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica* **58**(6), 1443–1458.

Dehling, H. and Mikosch, T. (1994). Random quadratic forms and bootstrap for U-statistics. *Journal of Multivariate Analysis* **51**, 392–413.

Gregory, G. G. (1977). Large sample theory for U-statistics and tests of fit. *The Annals of Statistics* **5**(1), 110–123.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics* **21**(4), 1926–1947.

Rodrigues-Campos, M. C., Gonzales Manteiga, W., and Cao, R. (1998). Testing the hypothesis of a generalized linear regression model using nonparametric regression estimation. *Journal of Statistical Planning and Inference* **67**, 99–122.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**(2), 613–641.

Stute, W., Gonzales Manteiga, W., and Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association* **93**(441), 141–149.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* **75**, 263–289.

**Acknowledgements:** This work was supported by Sonderforschungsbereich 373, hosted by the Humboldt-University Berlin.

# An Application of Nonlinear Regression for Correlated Data in Chemical Kinetics

R. Bellio [1], C. U. Carlsen [2], M. V. Kröger-Ohlsen [2] and L. H. Skibsted [2]

[1] Department of Statistics, University of Udine, Udine, Italy
[2] Food Chemistry, Department of Dairy and Food Science, The Royal Veterinary and Agricultural University, Frederiksberg, Denmark

**Abstract:** We present the statistical aspects of a study aimed to investigate a chemical reaction between a hypervalent myoglobin species and the antioxidant chlorogenate. As often in chemical kinetics, the individual estimates of reaction parameters should satisfy a linear dependence on the reciprocal temperature. This study shows the importance of taking into account the correlation among subsets of observations that may be present in kinetic data.

**Keywords:** Chemical Kinetics; Correlation Structures; Nonlinear Regression.

## 1 Introduction

In this paper we describe the analysis of a chemical experiment aiming to investigate the kinetics of the reduction of iron in ferrylmyoglobin by chlorogenate. Here we concentrate on the statistical aspects of the analysis, which are potentially interesting for similar studies, like analysis of compartment models (see for example Bates and Watts, 1988). The chemical aspects can be found in Carlsen *et al.* (2000).

The experiment was aimed to assess the effect of the concentration of chlorogenate on the rate constant of the reaction with myoglobin. Four different sets of experiments were performed at four different temperatures (5°, 12°, 19° and 25°). At each temperature, several reactions mixtures with the same chlorogenate concentration and different pH values were prepared, and 7-15 measurements of the rate constant were carried out.

A nonlinear model describing how the rate constant varies as a function of the chlorogenate concentration and the pH, based on theoretical grounds and previous studies (Kröger-Ohlsen and Skibsted, 1997), was formulated. The model was the following

$$\begin{aligned} f(A,H;\mathbf{K},\mathbf{k}) &= \left(k_1 \frac{A}{1+K_p\,A} + k_2 \frac{K_p\,A^2}{1+K_p\,A}\right) \frac{H}{H+K_a} \\ &+ \left(k_3 \frac{A}{1+K_i\,A} + k_4 \frac{K_i\,A^2}{1+K_i\,A}\right) \frac{K_a}{H+K_a}\,, \end{aligned} \tag{1}$$

where $A$ is the chlorogenate concentration, $H = 10^{-pH}$, $\mathbf{K} = (K_i, K_p, K_a)$ are equilibrium constants and $\mathbf{k} = (k_1, k_2, k_3, k_4)$ are rate constants. Each of the equilibrium and rate constants depends on temperature. A common form of dependence for equilibrium constants is the van't Hoff relation

$$K = \exp\left(\frac{\Delta S^{\ominus}}{8.31441}\right) \exp\left(-\frac{\Delta H^{\ominus}}{8.31441\, T}\right), \qquad K \in \mathbf{K}, \tag{2}$$

where $\Delta S^{\ominus}$ and $\Delta H^{\ominus}$ are the reaction entropy and the reaction enthalpy respectively, and $T$ is the temperature in Kelvins. For the rate constants, a common form of dependence upon temperature is the Arrhenius relation

$$k = 6.20927 \times 10^{12} \exp\left(\frac{\Delta S^{\#}}{8.31441}\right) \exp\left(-\frac{\Delta H^{\#}}{8.31441\, T}\right), \qquad k \in \mathbf{k} \tag{3}$$

where $\Delta S^{\#}$ and $\Delta H^{\#}$ are the activation entropy and the activation enthalpy respectively. The aim of the study was the estimation of the parameters of the individual model (1) at each temperature and to verify the van't Hoff and Arrhenius relations for the various parameters.

## 2 Individual estimation

Nonlinear regression was used for fitting model (1) to the data available at each temperature. The logarithm transformation was applied to both sides of the model equation to stabilize the variance. The model equation eventually was defined by

$$\log k_{obs}^{(T)} = \log f(A, H; \mathbf{K}^{(\mathbf{T})}, \mathbf{k}^{(\mathbf{T})}) + \varepsilon, \qquad \mathrm{T} = 278K, 285K, 292K, 298K,$$

with $\varepsilon$ being an independent error, which we assumed to be normally distributed with variance $(\sigma^2)^{(T)}$. The model does not hold when $pH$ is above a certain value, which was estimated to be around 7. The total sample sizes at each temperature were respectively $n = (176, 246, 202, 206)$ for $T = (278K, 285K, 292K, 298K)$. There were theoretical reasons suggesting that model (1) was only approximately correct; a better model for the reaction, however, was known to be much more complicated and non-estimable from the data. Moreover, we were aware of error in measurement of pH values, and the measurements coming from the same stock solution were likely to be serially correlated. For these reasons, we considered a clustered structure for the data, assuming an intra-cluster correlation structure. Although clustering was partially due to slight misspecification of the functional form of the model, we could not overlook the induced correlation. In fact, the assumption of independent observations would have led to overconsideration of the accuracy of the model. Hence, we assumed that observations made from the same stock solution and with the same pH value were correlated.

| Structure | number of parameters | $p$-value |
|---|---|---|
| Constant (no correlation) | $4 \times 8$ | $< 0.001$ |
| $AR1$ | $4 \times 9$ | 0.11 |
| $CS$ | $4 \times 9$ | 0.38 |
| $AR1 + CS$ | $4 \times 10$ | 0.37 |

TABLE 1. *Global test of linearity for all the equilibrium and rate constants with different correlation structures.*

To limitate the dependence on the assumed correlation, different *working correlation structures* were fitted. More precisely, we supposed that, if $Y = \log k_{obs}$, observations within the same cluster were correlated

$$\text{corr}\,(Y_{ij}, Y_{ij'}) \neq 0\,, \qquad j \neq j'\,,$$

with $i$ denoting the cluster and $j$ the replication. At each temperature, there were approximately from 20 to 26 clusters. We tried three different correlation structures, namely, following Diggle, Liang and Zeger (1994), *Compound symmetry* (CS), *First-order autoregression* (AR1), and *Compound symmetry+First-order autoregression* (CS+AR1).
Maximum likelihood estimates were computed in `S-Plus`, and the fit was judged satisfactory at all temperatures. We did not notice any substantial difference in the values of the fixed-effects estimates (the equilibrium constants and the rate constants) with different correlation structures, but substantial differences for their standard errors.

## 3 Aggregate estimation

The linear dependence of the logarithm of the parameters on the reciprocal temperature expressed by (2) and (3) were adapted directly into the model equation. The results are shown in Figure 1, visualizing the van't Hoff and Arrhenius plots. The full lines in the plots show the the ordinary linear regression line between the rate or equilibrium constant estimated for each of the four temperatures with CS+AR1 correlation structure, and the vertical bars show 95% confidence intervals for each parameter. The fitted lines from aggregate estimation are shown as dotted lines. We note a good agreement, with small deviations in some cases. A formal likelihood ratio test on the linearity of the parameter with respect to the reciprocal temperature was performed. The test was repeated for each correlation structure, and the related $p$-values are given in the Table 1. The assumption of correlated observations seems to have a strong effect on the inferential conclusions. It is reassuring that the hypothesis of global linearity is not rejected for all the correlation structures.
In this example the correlation among observations is somehow a nuisance aspect of the problem. In order to alleviate its effect on the accuracy of

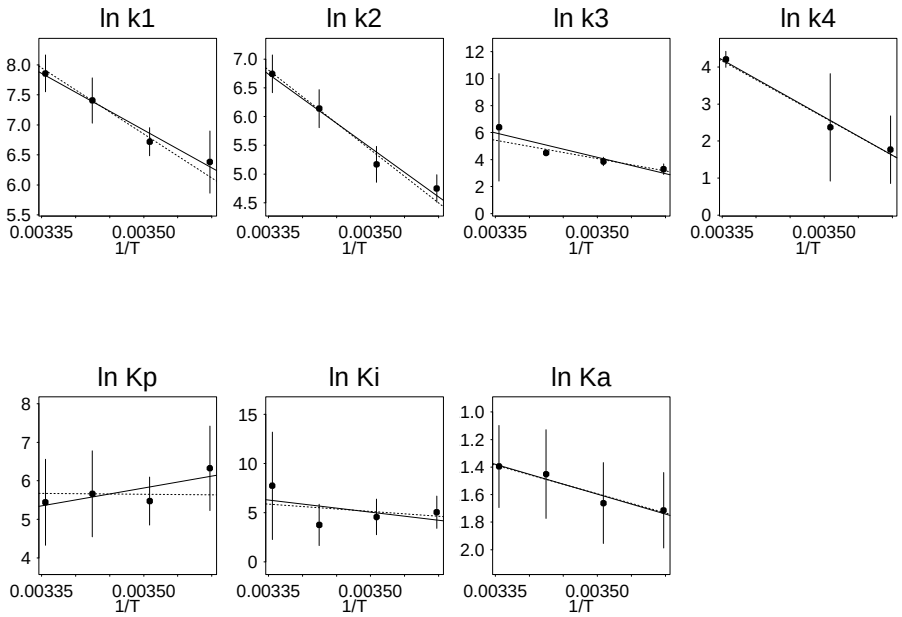



FIGURE 1. *Van't Hoff and Arrhenius plots according to model (1), comparing ordinary regression lines from individual estimates (solid) and aggregate estimates (dotted). (Modified from Carlsen et al., 2000).*

the final estimates (slopes and intercepts for all the parameters), standard errors for the final estimates were computed by a robust method. To this end, we adapted a formula by Diggle, Liang and Zeger (1994).
We concluded that the model (1) was able to describe the reaction. The final parameters of the reaction were judged satisfactory and closely connected to previously published data.

## References

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications.* Wiley, New York.

Carlsen, C. U., Kröger-Ohlsen, M. V., Bellio, R. and Skibsted, L. H. (2000). Protein binding in deactivation of ferrylmyoglobin by chlorogenate and ascorbate. *J. Agric. Food Chem.*, **48**, 204–212. .

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Kröger-Ohlsen, M. V. and Skibsted, L. H. (1997). Kinetics and mechanism of reduction of ferrylmyoglobin by ascorbate and D-isoascorbate. *J. Agric. Food Chem.*, **45**, 668–676.

# Taguchi Methods as a Powerful Engineering Tool

Monica Bernabe Fernandez

[1] Department of Business Organization, University of Basque Country, Faculty of Engineering, c/Alda. de Urquijo s/n 48013 Bilbao, Bizkaia, Spain; oebefem@ehu.es

**Abstract:** Taguchi methods were developed by Genichi Taguchi to improve the implementation of quality control in Japan. They are based on the design of experiments to provide near optimal quality characteristics. A controversy does exist concerning these methods. Many in academia complain that Dr.Taguchi does not faithfully follow all applicable statistical rules. Many engineers on the other hand prefer Dr.Taguchi's approach, which uses statistics as a foundation but emphasizes engineering judgment. This paper review his most notable contributions to quality improvement and the parameter design method for reducing cost and improving quality.

**Keywords:** Taguchi methods ; Robust design ;Design of experiments; Parameter design ; Quality engineering

## 1 Contributions to Quality

### 1.1 The Loss Function

Taguchi's philosophies and methodologies are directed towards achieving highest quality at minimum cost. Quality, defined as satisfying or exceeding customer expectations or a failure to deliver it, is expressed in monetary terms as quality loss.
Taguchi defines quality loss via his loss function. He unites the financial loss with the functional specification through a quadratic relationship that comes from a Taylor series expansion. The loss is a quadratic function of the deviation of the quality of interest from its target value. If the target value is a nominal value, the loss function takes the form of a parabola.
The quality loss relates performance, in terms of deviations from target, to financial loss, demostrating that the key to high quality is to minimize performance variability, which in turn minimizes quality loss.

### 1.2 Robustness

A second major contribution is the concept of robustness. We can define robustness from both a product and a process related standpoint as follows:

Product: The ability of the product to perform consistently as designed with minimal effect from changes in uncontrollable operating influences.
Processs: The ability of the process to produce consistently good product with minimal effect from changes in uncontrolable manufacturing influences.
To minimise loss, one is faced with the task of producing the product at optimal levels with minimal variation in its functional characteristics. The factors which affect the product's functional characteristics are of two types: Controllable and Noise factors. Controllable factors are those factors which can easily be controlled such as choice of material, cycle time or cool time in an injection moulding process. Noise factors, on the other hand, are those nuisance variables which are either difficult, or impossible, or expensive to control. Examples of noise variables include typical manufacturing variations such as non-uniformity in raw materials, deviation in the customer's environment, and deteroration or wear-out in component parts over time.

### 1.3 Parameter Design

A third major contribution is the parameter design, as the most important part for reducing cost involved in the quality engineering optimisation.
Product quality must be *engineereed in*. This is the thrust of Dr. Taguchi's off-line quality control activities which involve both product design and process design stages. There are three sequential stages for optimization :system design, parameter design, and tolerance design.
System design is the process of applying scientific and engineering knowledge to produce a basic functional prototype design. The prototype model defines the configuration and attributes of the product undergoing analysis or development. The initial design may be functional but it may be far from the optimum in terms of quality and cost.
The next stage, parameter design, is an investigation conducted to identify the settings of design parameters that optimize the performance characteristics and reduce the sensitivity of engineering designs to the sources of variation. By determining the best settings or values of those factors that are inexpensive to change or control, quality can be improved without adding cost.
Finally, tolerance design is employed if the reduced variation obtained through parameter design is not sufficient. It involves tightening tolerances on product parameters or process factors whose variations impart large influence on the output variation. In other words, tolerance design means spending money buying better grade materials, components or machinery.

## 2 The Experiment in Parameter Design

### 2.1 Planning the Experiment

The plan for a robust design experiment has two parts, a external array (control factors) and a external array (noise factors). For each experimental setting based on the external array, an experiment based on the internal array experimental setting is performed. This results in crossed experimental array.
The parameter design uses orthogonal arrays ($OA$) for internal and external array. Based on design of experiment theory, Taguchi 's orthogonal arrays provide a method for selecting an intelligent subset of the parameter space and significantly reduce the number of experimental configurations. Ortogonal arrays are not unique to Taguchi. All regular $m^{k-p}$ fractional factorial experiment are orthogonal arrays. For example, $OA_8$ is a $2^{7-4}$ fractional factorial plan, $OA_{16}$ is a $2^{15-11}$ fractional factorial plan and $OA_{12}$ is a Plackett-Burman plan. However , Taguchi has simplified their use by providing tabulated sets of standard orthogonal arrays and corresponding linear graphs to fit a specific projet. In the array, the columns are mutually orthogonal. That is, for any pair of columns, all combinations of factor levels occur, and they occur an equal number of times .
The symbology for an orthogonal array is $L_a(b^c)$ where : L = Latin square a = The number of test trials b = The number of levels for each column c = The number of columns in the array
The number of columns of an OA represents the maximum number of factors that can be studied using that array. For example the $L_9$ design reduce 81 ($3^4$ ) configurations to 9 . Using a noise array to introduce noise in a experiment, we can provide estimates of the interaction between every control parameter or interactions between them and every noise parameter.

### 2.2 The Analysis of Data

The traditional analysis which is performed with data from a designed experiment is the regular analysis. Taguchi stresses the importance of additionally studying the variation of the response and has introduce the use of the $S/N$ ratio to facilitate such investigation. In its simplest form , the $S/N$ ratio is the ratio of the mean (signal) to the standard deviation (noise) and is directly tied into the loss function. While there are many different $S/N$ formulae, three of them are considered standard and are generally applicable when the quality characteristic (response) can be classified as "higher the better", "lower the better" or "nominal the best". $S/N$ ratios are computed for each of the control experimental conditions and whatever the type of quality characteristic, the transformations are such that the $S/N$ ratio is always interpreted in the same way: the larger the $S/N$ ratio is, the better.

There are several ways to aprroach this analysis. One common way is to do a statistical analysis of variance (ANOVA) and perform F test to see which factors are statistically significant but Taguchi recommends use the level average analysis. This approach involves plotting the effects and visually identifying the factors which appear to be significant.
The analysis has two steps. First, finding the parameter setting that maximizes the $S/N$, and only then finding the set of factors that has a significant effect on the mean but does not influence the $S/N$ ratio, and use these factors to bring the mean on target.
Finally, Taguchi ' s methods require a confirmation run to check for both the validity of experimentation and, also, the reproducibility of the experimental conclusions.

## 3 Discussion

A controversy does exist concerning these methods. Many in academia complain that Dr.Taguchi does not faithfully follow all applicable statistical rules. Many engineers on the other hand prefer Dr.Taguchi's approach, which uses statistics but emphasizes engineering judgment. If the engineers who better know the product or procces are the same who plan the experiments, the risk that unexpected interactions or unexpected curvature might be present decrease, and the Taguchi ' experiment planning methodology we have described is an effective tool for reducing cost and improving quality. It provides a single approach that can be followed when the problem includes qualitative parameters, quantitative parameters, or a mixture of both.

## References

American Supplier Institute (1998).*Robust Design Using Taguchi Methods.* Workshop Manual ,ASI Livonia.

Box ,G.E.P (1993). Quality Improvement-The New Industrial Revolution. *International Statistical Review*, **61**, 1, 3-19.

Fowlkes, W.I. and Creveling ,C.M. (1995).*Engineering Methods for Robust Product Design : Using Taguchi Methods in Technology and Product Development.* Addison-Wesley Publishing Company.

Koyama ,K. and Nakano ,K. (1995).Aplication of Quality Engineering on Filter Circuit.*Journal of Quality Engineering Forum*, **3**, 3.

Logothetis, N. and Wynn ,H.P. (1989).*Quality through design : Experimental Design , Off-Line Quality Control and Taguchi's Contributions.* Oxford University Press , Oxford.

Wu , Y. and Wu, A. (1997).*Diseño Robusto Utilizando los Métodos Taguchi.*Ediciones Díaz de Santos. Madrid

# Comparing and Predicting between Several Methods of Measurement

Bendix Carstensen[1], Jaana Lindström[2], Jaakko Tuomiletho[2] and Knut Borch-Johnsen[1]

[1] Steno Diabetes Center, Niels Steensens Vej 2, DK-2820 Gentofte, bxc@novo.dk
[2] National Public Health Institute, Helsinki, Finland

**Abstract:** A model for comparing measurements by several methods is proposed. The model is applied for construction of predictions between 11 different methods for measuring blood glucose.

**Keywords:** Method comparison, mixed models, variance components, prediction

## 1 Problem and material

In order to provide a basis for comparing different analytical methods for blood glucose and make predictions from one method to another, a number of blood samples have been analysed by 11 different methods.
74 persons from 5 centres were chosen to participate, and blood was sampled from these persons at 0, 30, 60 and 120 minutes after an oral glucose tolerance test (75g) was taken, so 296 blood samples were available.
None of the blood-samples have been measured by all 11 methods, some were measured with 8 methods, some with 4 and some with 3. All samples from the same centre were measured by the same set of methods. All samples were measured by the methods `N.PLAS1` and `N.PLAS2`. Measurements are all in mmol/l, so they are all on the same scale, but this is no prerequisite for the methods developed here.

## 2 The model

For each method we assumed a linear relation of the measurement $y_{mit}$ to the unknown "true" value for the blood sample, $\mu_{it}$:

$$y_{mit} = \alpha_m + \beta_m \mu_{it} + a_{mi} + e_{mit} \tag{1}$$

where $y_{mit}$ is the outcome of method $m$ on sample $t$ from individual $i$. The random method×individual interaction, $a_{mi}$, and the combined measurement error and method×sample interaction $e_{mit}$ were assumed normally distributed with variances $\tau_m^2$ and $\sigma_m^2$ respectively.

The "true" values, $\mu_{it}$, are taken as parameters and hence no distributional assumptions are made about them. This is because subjects chosen for method comparison studies cannot reasonably be assumed to be representative of populations on which the derived prediction rules are to be applied. Seen from a prediction point of view, the $\mu$s must be regarded as nuisance parameters of no interest *per se*.
Clearly, the $\mu_{it}$s are only determined up to an affine transformation, and that can be remedied by putting $\alpha_1 = 0$ and $\beta_1 = 1$, but for symmetry reasons we shall keep the model in the form (1)
Many authors have used models similar to (1), where the $\mu$s are taken to be normally distributed, leading to factor-analysis types of models (Carter, 1981; Dunn & Roberts, 1999). If it is possibile to choose subjects for a method comparison study one would try to get the values to be evenly spread over the relevant range. Choosing subjects normally distributed would be to deliberately decrease power to detect deviations from the linearity at the extremes of the range.

## 3 Estimation algorithm

For fixed values of $\mu_{it}$ the model (1) is a linear mixed model with separate regressions for each $m$ on $\mu_{it}$ and a random effect of method×individual (and of course a residual variance). The variances of the random effects and the residual are specific for each method. Note that because the methods are assumed uncorrelated given the $\mu$s, the model falls apart in separate models for each method.
On the other hand, for fixed values of $\alpha_m$, $\beta_m$ and $a_{mi}$ the model (1) may be formulated as:

$$y_{mit} - \alpha_m - a_{mi} = \mu_{it}\beta_m + e_{mit}$$

i.e. regression of $y_{mit} - \alpha_m - a_{mi}$ on $\beta_m$ through the origin with separate slopes for each $(i, t)$, allowing for separate variances between methods.
Thus, estimation could be performed by alternating between the two formulations, fixing either set of parameters in turn.

## 4 Prediction

Predictions based on the model (1) should include both the measurement variation $\sigma_m$ and the method×individual variation $\tau_m$, but also take the uncertainty in $y_{20}$ into account.
For a (new) observed value of $y_2$, $y_{20}$, we get by inserting the naive estimate of $\mu_0$ based on $y_{20}$:

$$y_{10} = \alpha_1 + \beta_1\mu_0 + a_{10} + e_{10} = \alpha_1 + \beta_1\frac{y_{20} - \alpha_2 - a_{20} - e_{20}}{\beta_2} + a_{10} + e_{10}$$

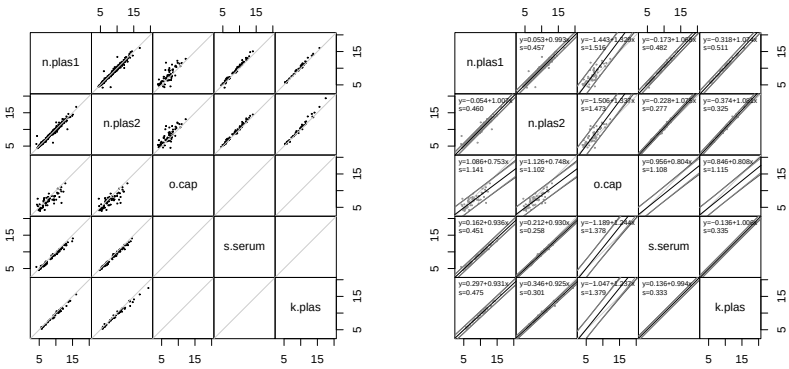



FIGURE 1. **Left:** *The original data for 5 of the 11 measurement methods.* **Right:** *The estimated connections between methods of measurement, with 95% prediction limits.*

From which we find the mean and variance of $y_{10}$:

$$\mathrm{E}(y_{10}) = \alpha_1 + \frac{\beta_1}{\beta_2}(y_{20} - \alpha_2), \qquad \mathrm{V}(y_{10}) = \left(\frac{\beta_1}{\beta_2}\right)^2 (\tau_2^2 + \sigma_2^2) + (\tau_1^2 + \sigma_1^2)$$

so the prediction variance depends both on the variance on the scale of the predictee as well as on the scale of the predictor. Note the fiducial character of the argument — when computing the variance of $y_{20} - \alpha_2 - a_{20} - e_{20}$, only the variance contributions from $a_{20}$ and $e_{20}$ are used, the variance of $y_{20}$ is ignored. This is because the computations is done in the *conditional* distribution of $\mu_0$, i.e. of $a_{20}$ and $e_{20}$ *given* $y_{20}$.
Note also that this kind of prediction interval has the property that it will produce a set of prediction bounds in a $(y_1, y_2)$-plot which is the same regardless of whether $y_1$ is predicted from $y_2$ or vice versa — the slope of the line linking $y_1$ with $y_2$ is $\beta_1/\beta_2$ so the vertical distance between two lines with this slope is $\beta_1/\beta_2$ times the horizontal, which is exactly the ratio of the standard deviations used in prediction in the two directions.

## 5 Results for the glucose data

The algorithm was implemented in SAS, using proc mixed to produce the estimated BLUPs in each iteration.
Figure 1 shows the estimated relation between methods with 95% prediction limits. Note that even though there are no samples measured by both `o.cap` and `s.serum` it is still possible under the model to produce an estimate of the connection between the methods with prediction intervals.

These relationships of course hinges quite strongly on the assumption of the mutual linear relationship between the models.
Clearly, the method `o.cap` has a very large variance, where as the other four methods in this display seems to perform quite well.

## 6 Discussion

A somewhat simpler variant of model (1) is that for simple replicates within each method, where $t$ is just the indicator of replicate, and where one would replace $\mu_{it}$ with $\mu_i$.
The method proposed here will only work if there is sufficient replication to determine the $\mu$s, either by replicates of meaurements by each method or through measurement by several methods. Formally, the model would be identifiable if there were either two replicates on one method or three methods, but for practical purposes one would not feel comfortable by much less than 5 datapoints per nuisamce parameter ($\mu_{it}$) — in this example we had 1302 datapoints and 296 nuisance parameters (!).
The method proposed here is a method where the "true" values of the samples are taken as parameters and not as random effects or latent variables with some distribution. This results in a model with a lot of nuisance parameters, but they have more or less the same role as the nuisance parameters that one conditions out (or omits) when doing a paired t-test. In fact, if the method$\times$individual-effects were 0 and the $\beta$s were 1, the proposed model would correspond to that underlying the paired t-test. Therefore the estimation of the nuisance parameters will probably only have marginal effect on the effeciency in the estimation of the parameters of interest.

### References

Bland, J.M. and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet,* **i**, 307–310.

Bland, J.M. and Altman, D.G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research,* **8**, 135–160.

Carter, R.L (1981). Restricted maximum likelihood estimation of bias and reliability in the comparison of several measuring methods. *Biometrics,* **37**, 733–741.

Dunn, G. and Roberts, C (1999). Modelling method comparison data. *Statistical Methods in Medical Research,* **8**, 161–179.

# Cox Proportional Hazard Model For Altitude Decompression Sickness

Raj S. Chhikara[1], Floyd M. Spears[1] and Thomas English[1]

[1] University of Houston-Clear Lake, 2700 Bay Area Blvd., Houston, TX 77058

**Abstract:** An application of Cox proportional hazard model is made to investigate its suitability to fit the hypobaric decompression sickness data of NASA. It is shown that a stratified model based on this application provides a suitable model for the analysis of these data.

**Keywords:** Censoring; Goodness of Fit; Stratified Model; Residual Analysis.

## 1 Introduction

Aviators and astronauts may experience altitude decompression sickness (DCS) as a result of reduced environment pressure. When astronauts have to perform extra-vehicular activities or when there is a damage or loss of the cabin or space suit pressure, they may be exposed to acute environment pressure reduction and thus run the risk of DCS. Scientists at NASA/Johnson Space Center have conducted, over several years, experimental tests using hypobaric chambers and simulated extra-vehicular activities to determine the DCS incidence and its onset time for human subjects. During a test, each subject is monitored for Doppler detectable bubbles, and the test is terminated either upon incidence of a DCS onset or when the test period is over. The test duration is recorded in each case, and so an observation is either the DCS onset time or the test termination (censored) time for a test subject. Empirical studies of decompression sickness have been made by NASA using the experimental data collected from the hypobaric chamber tests. Conkin et al. (1996) modeled the probability of DCS occurrence by relating it to decompression stress as measured by the relative change in atmosphere that the human body can sustain without incurring DCS. This decompression stress measure is called the tissue ratio (TR) index. A number of researchers based their developments on certain bubble dynamic models to characterize the DCS occurrence rate. The analyses reported in their and other more recent studies undertaken by the researchers from NASA and US Air Force focus on the selection of a probabilistic model that statistically provides the best possible fit for the experimental DCS data.

The statistical analyses have mostly been limited to parameter estimation for a model and do not provide the significance level or reliability measure

for the data-fitted model. The maximum likelihood method is invariably used in estimating the model parameters and the probability density or distribution function, whereas the Kaplan-Meier method is used for the empirical estimate of the underlying distribution function. The goodness-of-fit is judged either by comparing the graphs of the two estimates of the distribution function, or by the chi-square test based on the maximum likelihood value resulting from a model fit. It has been shown that both the lognormal and the log logistic probability models provide a reasonable model fit to the NASA DCS data.

Besides the DCS onset time observations, data are available on a number of physiological and related variables (Conkin, et al. (1992)). Previous data analysis studies conducted by NASA and US Air Force have made use of these variables to explain the variability in DCS onset time. In modeling the DCS onset time response, the important variables are ambient pressure (P2) at test altitude, nitrogen (N2) pressure, particularly that determined in the theoretical 360 minute half-time compartment (PN2360), exercise, and preoxygenation, among others. A log logistic model involving some of these explanatory variables has been fitted to the DCS data. Conkin, et al. (1996) and Kannan, et al. (1998), among others, have empirically shown that a log logistic model provides a reasonable model-fit and that several of these variables contribute significantly to the occurrence of DCS.

In the present study we apply the Cox proportional hazard model to analyze the DCS data obtained from NASA. These data consist of 1321 test duration times of which 1154 are censored and 167 are DCS onset time observations. The following covariates are used in this modeling: P2, PN2360, ALTTIME (altitude time), TR360, and EXER (exercise). The choice of these variables is partly based on the fact that their measurements are also available for the test subjects for which the DCS data are analyzed.

## 2 Statistical Analysis

We first discuss the application of Cox proportional hazard model and then determine the significance of each of these variables and interactions among them. The results in Table 1 show the significant variables.

**Table 1: Model Fit**

| Variable | $\beta$ | $\exp(\beta)$ | $se(\beta)$ | z | p-value |
|---|---|---|---|---|---|
| PN2 | -2.3629 | 0.0941 | 0.3677 | -6.43 | 1.3e-010 |
| PN2360 | 1.2220 | 3.3940 | 0.2192 | 5.57 | 2.5e-008 |
| TR360 | -2.6822 | 0.0684 | 0.9184 | -2.92 | 3.5e-003 |
| PN2360*EXER | 0.0824 | 1.0859 | 0.0273 | 3.02 | 2.5e-003 |

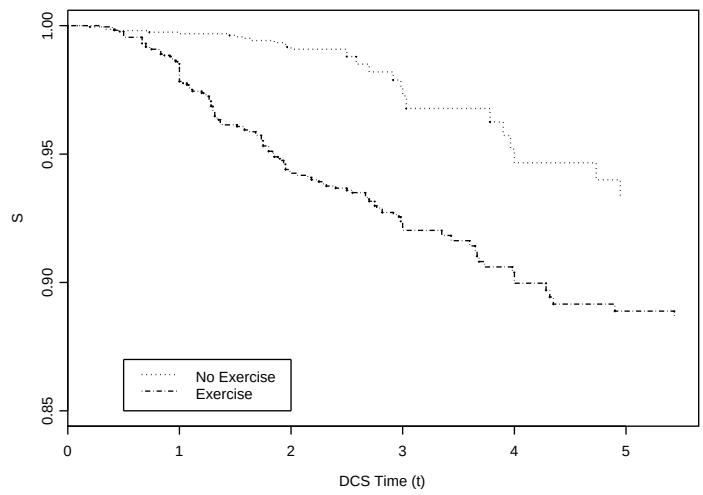
The model with interaction has a likelihood value of 281, while the model without interaction has a likelihood value of 279. This shows there is no

significant advantage to the interaction term model over the non-interaction term model.

## 3 A Stratified Model

The covariate EXER is an exercise indicator variable, and hence it takes value as either 0 or 1. This presented computational difficulty in obtaining residuals associated with EXER as a covariate in the model and thus, examining of the model assumption. Instead if EXER is used as a stratified variable, and not as a covariate in the model, this will not present computational difficulty in determining the residuals using the S-Plus routine resid() and carrying out the residual analysis needed to evaluate the model suitability.

Figure 1: Cox Model Fitted Survival Functions




Wei (1984) proposed an omnibus test of proportional hazard between two groups using data containing censored observations. We consider here an application of Wei's test to verify the assumption of proportional hazards between the two groups stratified by EXER. The Wei test statistics is a measure of goodness of fit and its computed value is 1.44. Using the non-exercise group, we compute a truncation value of $27/263 = 0.1$ ( number observed/number in sample) needed to look up the p-value from the applicable table of cumulative probabilities as given in Koziol and Byar (1975). This yields a p-value less than 0.001. This allows us to reject the hypothesis

of proportional hazards between the exercise and non-exercise groups and thus provides a justification for the use of above stratification.
Finally, we assessed the assumption of proportional hazards for the stratified model by examining the rescaled Schoenfeld residuals and found that proportional hazards is a reasonable assumption. Figure 1 shows the fitted stratified model for exercise and non-exercise cases depicted separately.
The estimated model coefficients and their standard errors (se) obtained for the two cases of the variable EXER are listed below for the stratified model.

**Table 2: Stratified Model Fit**

**Exercise = 0**

| Variable | $\beta$ | exp($\beta$) | se($\beta$) | z | p-value |
|---|---|---|---|---|---|
| PN2 | -3.16 | 0.0422 | 1.263 | -2.51 | 0.0120 |
| PN2360 | 1.41 | 4.1147 | 0.522 | 2.71 | 0.0067 |
| TR360 | -3.91 | 0.0201 | 2.071 | -1.89 | 0.0590 |

**Exercise = 1**

| Variable | $\beta$ | exp($\beta$) | se($\beta$) | z | p-value |
|---|---|---|---|---|---|
| PN2 | -1.778 | 0.167 | 0.527 | -3.395 | 0.00069 |
| PN2360 | 0.973 | 2.645 | 0.313 | 3.107 | 0.00190 |
| TR360 | -0.849 | 0.428 | 1.522 | -0.558 | 0.58000 |

## References

Conkin, J., Bedhl, S. R. and Van Liew, H. D. (1992). A Computerized Databank of Decompression Sickness Incidence in Altitude Chambers. *Aviation, Space, and Environmental Medicine*, 63, 819-824.

Conkin, J., Kumar, K. V., Powell, M. R., Foster, P. P. and Waligora, J. M. (1996). A Probabilistic Model of Hypobaric Decompression Sickness Based on 66 Chamber Tests. *Aviation, Space, and Environmental Medicine*, 67, No. 2, 176-183.

Kannan, N., Raychandhuri, A., and Pilmanis, A. A. (1998). A loglogistic model for altitude decompression sickness. *Aviation, Space, and Environmental Medicine*, 69.

Koziol, J. A., Byar, D. P. (1975). Percentage Points of the Asymptotic Distributions of One and Two Sample K-S Statistics for Truncated or Censored Data. *Technometrics*, Vol 17 No 4:507-510

Wei, L.J. (1984). Testing Goodness of Fit for Proportional Hazards Model With Censored Observations. *JASA*, 79:649-652.

# Adjusted profile score: some applications

Iain Currie[1] and Maria Durbán[2]

[1] Department of Actuarial Mathematics and Statistics, Heriot-Watt University, Edinburgh, EH14 4AS, Scotland
[2] Biomathematics & Statistics Scotland, The King's Buildings, Edinburgh, EH9 3JZ, Scotland

**Abstract:** McCullagh & Tibshirani (1990) proposed an adjusted profile score as a way of solving the twin problems of bias and optimistic standard errors associated with the use of the profile likelihood. Durban and Currie (2000) derived the exact adjustments for a general non-linear normal regression model. In this paper we apply these results to three examples, and show that adjusted profile likelihoods can be obtained in all three cases.

**Keywords:** Adjusted profile score; adjusted profile likelihood; non-linear regression.

## 1 Introduction

The score function computed from the full log-likelihood has zero expectation (unbiased) and variance equal to the negative of the expectation of its derivative matrix (information unbiased). However, the score function computed from the profile likelihood is, in general, neither unbiased nor information unbiased. McCullagh & Tibshirani (1990) suggested that the properties of estimates based on the profile likelihood would be improved if the profile score function was centred and scaled so that it too was unbiased and information unbiased. McCullagh & Tibshirani (1990) concentrated on giving asymptotic formulae for their corrections in a very general setting. Durban and Currie (2000) gave exact expressions for the corrections for the non-linear normal regression model

$$Y \sim \mathcal{N}(X(\psi)\lambda, \Sigma(\psi)), \tag{1}$$

where $\psi$ is regarded as the parameter of interest and $\lambda$ as the nuisance parameter. This model has a number of special cases. In particular, if $X(\psi) = X$, i.e., $X$ does not depend on the parameter of interest, then (1) reduces to the familiar model for variance components, in which case our results show that the usual residual maximum likelihood (REML) is both unbiased and information unbiased.

We give some notation. We denote the profile score function by $U(\psi) = \partial\ell_{\mathrm{p}}(\psi)/\partial\psi$ where $\ell_{\mathrm{p}}(\psi)$ is the profile log-likelihood in $\psi$. The adjusted profile score is then

$$\widetilde{U}(\psi) = W(\psi)(U(\psi) - m(\psi)) \tag{2}$$

where the centring term $m = m(\psi)$ and the scaling term $W = W(\psi)$ for model (1) are given in Durban and Currie (2000).
In this paper we apply these results to three examples: the ratio of normal means problem, a non-linear regression problem studied by Ratkowsky (1983), and a response surface model due to Draper and Guttman (1980).

## 2 Examples

**Example 1:** We consider the ratio of normal means problem (McCullagh and Tibshirani's example 10). We observe $y_1 \sim \mathcal{N}(\lambda, 1/n)$ and $y_2 \sim \mathcal{N}(\psi\lambda, 1/n)$. This model is in the family (1) with $\Sigma = I_2/n$ and $X' = (1, \psi)$. In this example, $\psi$ is a scalar, so $m$ and $W = w$, say, are both scalars. We find the centring term $m = 0$, and the scaling factor $w$ is

$$w^* = \frac{1}{1 + \dfrac{1+\psi^2}{n(y_1+\psi y_2)^2}} \approx 1 - \frac{1+\psi^2}{n(y_1+\psi y_2)^2} \tag{3}$$

where $w^*$ indicates that we have replaced the nuisance parameter $\lambda$ in $w$ by $\hat{\lambda}_\psi$. The profile log-likelihood for this problem is

$$\ell_{\mathrm{p}} = -\frac{n}{2(1+\psi^2)}(y_2 - \psi y_1)^2, \tag{4}$$

and the adjusted profile log-likelihood can be found since

$$\ell_{\mathrm{ap}} = \int^{\psi} w^*(t)\frac{d\ell_{\mathrm{p}}}{dt}\,dt = \ell_{\mathrm{p}} - \tfrac{1}{2}\log\left(1 + \frac{2\ell_{\mathrm{p}}}{1+n(y_1^2+y_2^2)}\right) \tag{5}$$

where we have arranged that both $\ell_{\mathrm{p}}$ and $\ell_{\mathrm{ap}}$ have the same maximum value (of zero). Fig. 1 gives a plot of $\ell_{\mathrm{p}}$ and $\ell_{\mathrm{ap}}$ for $n = 5$ and $y_1 = y_2 = 1$.

**Example 2:** We consider first models of the form

$$Y \sim \mathcal{N}(X(\psi)\lambda, \sigma^2 I) \tag{6}$$

where the parameter of interest is taken as $(\psi', \sigma^2)' = (\psi_1, \ldots, \psi_{r-1}, \sigma^2)'$. We find $m_1 = m_2 = \ldots = m_{r-1} = 0$ and $m_r = -p/(2\sigma^2)$ where $p$ is the rank of $X$ with resulting bias adjusted profile log-likelihood

$$\ell_{\mathrm{ap}} = -\frac{n-p}{2}\log\sigma^2 - \frac{1}{2\sigma^2}y'(I-P)y \tag{7}$$

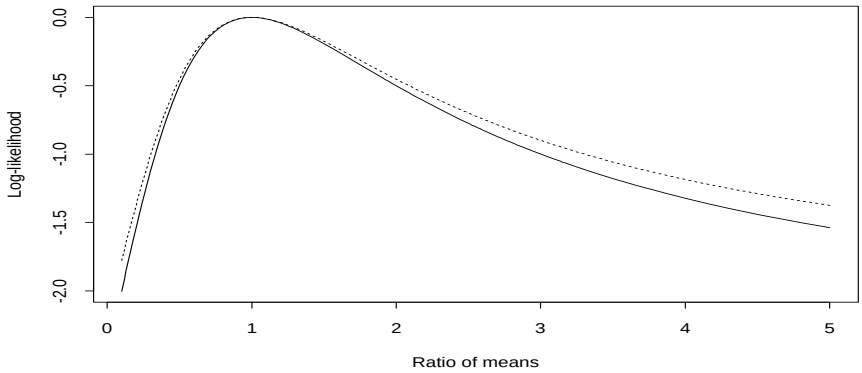



FIGURE 1. Profile likelihood —— and adjusted profile likelihood - - - for ratio of normal means example: $n = 5$, $y_1 = y_2 = 1$

where $P = X(X'X)^{-1}X'$. It follows that the estimates of $(\psi_1, \ldots, \psi_{r-1})$ based on the bias adjusted profile log-likelihood are equal to the maximum likelihood estimates based on the full log-likelihood. The estimate of $\sigma^2$ is $\hat{\sigma}^2 = y'(I-P)y/(n-p)$, a REML style estimate.
One example of this sub-class of models is a non-linear regression model discussed in Ratkowsky (1983, p93), and fitted by him to data on leaf production. We write his model as

$$y_i = \alpha + \beta \exp(-\rho z_i) + \epsilon_i \tag{8}$$

and take $\rho$ and $\sigma^2$ as parameters of interest. Note that $\rho$ is the parameter of physical interest. The model is of the form (6) with $\lambda' = (\alpha, \beta)$ and $X = [1, x]$ where 1 is a vector of 1's and $x$ is the vector with $x_i = \exp(-\rho z_i)$. The bias adjustment is dealt with in (7), and the scale adjustment is

$$W = \begin{pmatrix} \left(1 + \dfrac{\sigma^2}{S_{xx}\,\beta^2}\right)^{-1} & 0 \\ 0 & 1 \end{pmatrix} \tag{9}$$

where $S_{xx} = \sum(x_i - \bar{x})^2$. We find $W(1,1)$ has the value 0.9968 at the estimates obtained from the adjusted score, and so the 'plug-in' value of $W \approx I_2$. Thus, the bias adjusted profile log-likelihood (7) is very close to being information unbiased also.

**Example 3:** Draper & Guttman (1980) introduced a response surface model for variety trials in which the response on a particular plot may be affected by the variety on a neighbouring plot. We write their model as

$$Y = G(Z_1\beta + Z_2\gamma) + \epsilon = GZ\lambda + \epsilon$$

where $Z_1$ and $\beta$ are the design matrix and parameter vector for the variety effects, and $Z_2$ and $\gamma$ are the design matrix and parameter vector for the block effects. We take $Z = [Z_1 : Z_2]$ and $\lambda' = (\beta', \gamma')$. The matrix $G$ specifies the neighbour effects and depends on a parameter $\rho$ ($\alpha$ in Draper and Guttman's paper). Again, we have a model of the form (6) with $(\rho, \sigma^2)$ as the parameter of interest and $\lambda$ as the nuisance parameter. Draper and Guttman present data from an experiment on mildew control which was laid out in a single strip of 38 plots. The model is defined by $G = I + \rho S$ where $S$ specified the neighbour effects. We apply our results with $X = (I+\rho S)Z$ and $\Sigma = \sigma^2 I$. The estimates obtained from the adjusted score are $\hat{\rho} = 0.158$ and $\hat{\sigma}^2 = 0.0262$ . The value of the scale adjustment matrix $W^*$ at $\hat{\rho}$ and $\hat{\sigma}^2$ is $W^*(\hat{\psi}) = \{\{0.982, 0\}, \{0, 1\}\}$ and again the scale adjustment to the score is very small. It is not possible to obtain the bias and scale adjusted log-likelihood for $\rho$ and $\sigma^2$ jointly but, noting that $\rho$ and $\sigma^2$ are orthogonal we can take $\sigma^2 = \hat{\sigma}^2 = 0.0262$, and then scale (7) by $w = 0.982$. This gives an adjusted log-likelihood (with maximum value zero) for $\rho$

$$\ell_{\mathrm{ap}}(\rho) = 0.982 \left( \frac{n-p}{2} - \frac{1}{2\hat{\sigma}^2} y'(I-P)y \right) \tag{10}$$

which is exactly unbiased and approximately information unbiased. A plot shows that there is little difference between the profile likelihood for $\rho$ and this adjusted profile likelihood.

## 3 Concluding remarks

Bias and optimistic standard errors are two common problems associated with the profile likelihood. The results of Durban and Currie (2000) provide a possible solution for a wide class of non-linear regression models. The examples in this paper add to the examples already provided in Durban and Currie (2000) and illustrate the flexibility of model (1).

### References

Draper, N. R. & Guttman, I. (1980). Incorporating overlap effects from neighbouring units into response surface models. *Appl. Statist.* **29**, 128-134.

Durban, M. & Currie, I. D. (2000). Adjustment of the profile likelihood for a class of normal regression models. *Scand. J. Statist.* **27**, *to appear.*

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52**, 325-344.

Ratkowsky, D. A. (1983). *Nonlinear Regression Modelling.* New York: Marcel Dekker.

# Exploratory and Confirmatory Statistical Modelling of a Materialism Scale: US and UK Samples.

Mark A. P. Davies[1], Chris Manolis[2], Melvin Prince[3]

[1] Inst. University of Loughborough, Loughborough, UK,
[2] Quinnipiac College, Hamden, CT., USA
[3] Pace University, New York, NY, USA

**Abstract:** The Richins and Dawson materialism scale used consumer values as the basis of its conceptualization. The original scale research empirically established three value-oriented scale components-acquisition centrality, acquisition in the pursuit of happiness, and possession-defined success., The present study re-evaluates their Material Values Scale from a cross-national perspective, comparing US and UK samples. Scale reliabilities are found to be largely unchanged from the original study. However, in both countries, the model fit for the three value components is somewhat below previous standards. Scale revision strategies are suggested for a more acceptable fit of the values constructs.

## 1 General issues of measurement validation

Measurement error include problems with convergent or discriminant validity. Convergent validity is upheld when all items that supposedly represent a domain of the concept do belong to the domain of the concept, that may be indicated by average correlations of item scores being similar within the domain, but distinct from all other items that aim to represent alternate domains of the concept (i.e., discriminant validity). A large coefficient alpha is a test for internal consistency to ensure that all items share a similar amount with the core concept. A useful way of measuring convergent and discriminant validity is by the multitrait-multimethod matrix.

## 2 Advantages of multi-scale items

Multi-item scales are preferred to single item measures that tend to correlate poorly with attributes they are supposed to measure, and suffer from

poor discriminant validity. When combined into multi-item measures, the specificity can be averaged out, with reliability increasing as the number of items in a combination increases.

## 3 Factor analysis

When the alpha coefficients are too low, the domain may need re-specifying (since perhaps respondents guessed their meaning), or additional items on each subscale might have been ignored. To rule out random factors accounting for the findings, the results should be reproduced with the purified number of items submitted to another sample of subjects. Replication of test results is also required for generalization of knowledge. Test-re-test coefficient of correlation scores examine the stability of the measures over time, with a reliability coefficient of 0.80 or above showing reasonable reliability. Both factor analytical and principal components analyses should be conducted because the magnitudes of factor loadings, the signs attached to the factor loadings, and the items loaded might differ between the two techniques.

## 4 Applications to materialism

Construct validity has been difficult to establish because there is no consensus on how materialism should be conceptualised or measured.
Hence, we discuss three research streams of materialism: as traits; and as values at the individual and societal level, before focussing on values at the individual level. Since the complexity of materialism is recognised to be a multidimensional concept, there is an initial need to verify the factorial validity of the measure, and then to test for convergent and discriminant validity. Repeated tests of factorial, convergent, and discriminant validity are required across time and population groups to ensure the measure is appropriately understood and valid.

## 5 Study Objectives

The primary objective of our study is to test assumptions about the reliability and validity of the original 18-item scale of values proposed by Richins and Dawson(1992) for both US and UK. To accomplish this, it is

necessary to assess (a) the internal scale reliabilities, (b) the unidimensionality of sub-scales, (c) the consistency of performance of sub-scale items with the original scale, (d) the fit of a three factor model to the scale and (e) the equivalence of scale structure cross-nationally.
Hypotheses for the US and UK to be tested in the present study are:

- H1 Reliability levels of the overall Material Values Scale are acceptable
- H2 Each Material Values sub-scale is unidimensional
- H3 The three sub-scale domains of success, centrality and happiness are reproducible
- H4 The Material Values scale items continue to reflect sub-scale domains
- H5 The latent variable model of the Material Values scale has the same structure cross-nationally

A second objective of our study is recommend any necessary revisions of the measurement model of the Material Values scale, i.e., to improve the reliability and validity of the instrument.

## 6 Method

Data collection was undertaken among undergraduate business students at two universities–one in Northeast US(n = 122), and the other in the Midlands UK(n = 138). The measure validation of the scale included use of Alpha Reliability estimates, principal components analysis, exploratory factor analysis(EFA), and confirmatory factor analysis(CFA). For the CFA, several alternative models were tested for model fit, in addition to the Richins and Dawson three factor model.

## 7 Main Findings and Conclusions

The study's analyses were designed to test and evaluate a theoretically-driven measurement instrument– the Materialist Values Scale by Richins and Dawson. Internal validity, factorial validity and the cross-national(US vs UK) structural equivalence of the scale were examined.

The scale exhibited very good internal reliability, exceeding the recommended benchmark Cronbach Alpha level of .80(DeVellis, 1991). The subscales showed acceptable reliabilities. The scale's three dimensional structure was reproduced in this research, although imperfectly.
The original 18-item Richins and Dawson Scale(ORD) needed item refinement since multi-dimensionality was found for subscales where unidimensionality should have been observed. Also, the confirmatory model subsuming constructs of success, centrality and happiness did not fit well.
A revised 14-item scale(RDD) exhibited unidimensionality of construct measurement, displaying an acceptable confirmatory fit. Additionally, the RDD scale shows cross-national structural equivalence of invariance between the US and UK samples. We suggest that future research on materialism beyond a single country should consider modifications of the Richins and Dawson Scale as part of their measurement methodologies.

# A Shifted Binomial Model for Rankings

Angela D'Elia[1]

[1] Dipartimento di Scienze Statistiche, Università di Napoli Federico II
Via G. Sanfelice 47, I-80134 Napoli, Italy; e-mail: angdelia@unina.it

**Abstract:** The paper discusses a model for rankings based on a paired comparisons structure. The statistical model is developed by means of a Shifted Binomial random variable; then, inferential and computational issues concerning model parameters estimation are addressed. Some empirical results show the usefulness of the approach in analysing preferences.

**Keywords:** Ranks Modelling, Paired Comparisons, Shifted Binomial random variable

## 1 Introduction

The analysis of preferences and, more in general, of ratings is a useful tool in several fields, as marketing, political sciences, medicine, psychology, etc. Usually, ratings expressed as rankings have been analysed in a paired comparisons framework, as in the Bradley-Terry model and its extensions (Bradley, 1976). The generalization to an arbitrary number of response categories has, then, led to proportional cumulative odds models (McCullagh, 1980) in a generalized linear models approach. In the same spirit, D'Elia (1999) proposed a model for ranks based on an Inverse Hypergeometric random variable, whose parameters are meaningful in preferences terms. In this paper we discuss an alternative model for ranks, which exploits the logic of paired comparisons in order to study the relation between raters features and rankings. Moreover, this structure generalizes the fitting of a discrete random variable for non-monotonic data.

## 2 A Shifted Binomial Model

Let $r_{ij}$ be the rank given by the $i$-th rater to item $j$ $(i = 1, 2, ..., n;\ j = 1, 2...., m)$ among $m$ alternatives. The complete ranking of all the items represents, for each rater, a permutation of the integers $1, 2, .., m$. This ranking can be thought of as the final product of a choice mechanism based, for each item, on $m - 1$ paired comparisons with the others.
Let's define the following Bernoulli random variable $(\forall\ h \neq j)$:

$$\begin{array}{ll} W_{jh} = 0 & \text{if item } j \text{ is preferred to item } h \\ W_{jh} = 1 & \text{if item } j \text{ is } \textbf{not} \text{ preferred to item } h. \end{array}$$

Assuming that $W_{jh}$ and $W_{jk}$ are mutually independent ($\forall\ h \neq k$), it results:

$$S_j = \sum_{h \neq j} W_{jh} \sim Binomial(m-1, \psi_j),$$

where $\psi_j = Pr(W_{jh} = 1)$. Thus $S_j$ is equal to 0 if the item $j$ is preferred in all the comparisons; it is equal to 1 if it is preferred in all the comparisons but one, and so on; it takes the value $m-1$ if it is never preferred.
Consequently, the observed ranks for a $j$-th item can be thought of as realizations of a Shifted Binomial random variable $R_j = S_j + 1$, where

$$Pr(R_j = r) = \begin{pmatrix} m-1 \\ r-1 \end{pmatrix} \psi_j^{r-1}(1-\psi_j)^{m-r}, \qquad r = 1, 2, ..., m;$$

$$E(R_j) = (m-1)\psi_j + 1; \qquad Var(R_j) = (m-1)\psi_j(1-\psi_j).$$

Notice that when $\psi_j \to 1$, $Pr(R_j = 1) \to 0$ and $E(R_j) \to m$: this means that $\psi_j$ is a *disagreement parameter*, because the greatest is $\psi_j$ the smallest is the probability that item $j$ is located in the first ranking positions.
In order to analyse the relation between the covariates $X_1, X_2, ..., X_p$ measured on each rater and the ranking he expressed, we let

$$log\,[\psi_j/(1-\psi_j)] = \mathbf{X}\beta$$

where $\mathbf{X}$ is the design matrix. Thus

$$E(R_j) = (m-1)\left(\frac{e^{\mathbf{X}\beta}}{1+e^{\mathbf{X}\beta}}\right) + 1.$$

It results that:

$$\begin{array}{ll} \text{if } \mathbf{X}\beta \to -\infty & \text{then } E(R_j) \to 1; \\ \text{if } \mathbf{X}\beta \to 0 & \text{then } E(R_j) \to (m+1)/2; \\ \text{if } \mathbf{X}\beta \to \infty & \text{then } E(R_j) \to m. \end{array}$$

This points out that the expected rank increases (the liking decreases) with the predictor value, and viceversa when $\mathbf{X}\beta$ decreases.

## 3 Inferential and computational issues

Let $\mathbf{r} = (r_1, r_2, ..., r_n)'$ be the sample of obseved ranks for a given item $j$. For the model specified above, the log-likelihood function is (dropping out the $j$ index):

$$log\mathcal{L}(\psi; \mathbf{r}) = \sum_{i=1}^{n} log\mathcal{L}(\psi; r_i) = \sum_{i=1}^{n} \{(r_i - 1)log(\psi) + (m - r_i)log(1-\psi)\}.$$

After some algebra, it can be shown to be equal to:

$$log\mathcal{L}(\beta;\mathbf{r}) = \sum_{i=1}^{n}\{(r_i-1)\mathbf{x}_i\beta - (m-1)log(1+e^{\mathbf{x}_i\beta})\},$$

where $\mathbf{x}_i$ is the $i$-th row of $\mathbf{X}$.
A careful choice of starting parameters values is needed, for maximum likelihood estimation, in order to improve the achievement and the speed of convergenge of numerical optimization algorithms. Then, a useful strategy could be derived from the previous expression of $E(R)$.
If we use $r_i$ as a proxy of $E(R)$, we obtain:

$$\frac{r_i-1}{m-1} = \frac{e^{\mathbf{x}_i\beta}}{1+e^{\mathbf{x}_i\beta}}.$$

Thus, it results that

$$log\left(\frac{r_i-1}{m-r_i}\right) = \mathbf{x}_i\beta,$$

and the starting values of the coefficient estimates can be obtained as the Least Squares solution of

$$\mathbf{z} = \mathbf{X}\beta + \epsilon$$

where the elements of $\mathbf{z}$ are defined as: $z_i = log\left(\frac{r_i-1}{m-r_i}\right), \quad i = 1,2,..,n.$
In the model with only the constant term, instead, the maximum likelihood estimate of $\beta_0$ can be analitically obtained as:

$$\hat{\beta}_0 = log\left(\frac{\bar{r}-1}{m-\bar{r}}\right), \qquad \text{where} \quad \bar{r} = \sum_{i=1}^{n} r_i/n.$$

## 4 An analysis of the attitudes to the professions

A research has been conducted to analyse the professional attitudes of young people: a sample of 183 undergraduates, attending the Faculty of Political Sciences, were asked to rate 14 different professions, giving rank 1 to the most preferred, rank 2 to the second best, and so on. Several covariates were measured on each subject, like age, sex, etc., to study the relation between students' features and the preferences they gave. As an illustrative example, in this section we show just few results.
One of the typical professions for people attending the Faculty of Political Sciences is the activity inside political parties and trade-unions. Fitting the proposed model to the ranks given to this kind of activity, we got the following results:

| *Covariates* | $\hat{\beta}$ | *e.s.* | $\hat{\beta}/e.s.$ | $p-values$ |
|---|---|---|---|---|
| constant | -1.7136 | *0.3679* | -4.6572 | 0.0000 |
| Sex (S) | 1.0137 | *0.0860* | 11.7894 | 0.0000 |
| Age (A) | 0.0654 | *0.0174* | 3.7662 | 0.0001 |
| Father's profession (Fp) | 0.6400 | *0.1465* | 4.3685 | 0.0000 |

It appears that, *ceteris paribus*, for the women (S=1) the disagreement odds-ratios for political activities are nearly 3 times greater than for the men (S=0): $exp(\hat{\beta}_1) = exp(1.0137) = 2.7558$. Moreover, the dislike increases with the age, and when the father of the interviewed is a professional (Fp=1) it nearly doubles: $exp(\hat{\beta}_3) = exp(0.64) = 1.8965$. For instance, comparing the expected rank for four 20 years-old people, who are different for sex and father's profession, we have:

| *Covariates values* | | E(R) |
|---|---|---|
| S=1 | Fp=1 | 11.10 |
| S=1 | Fp=0 | 9.42 |
| S=0 | Fp=1 | 8.26 |
| S=0 | Fp=0 | 6.20 |

## 5 Concluding remarks

In this paper we discuss ranking modelling where the ranks express people preferences, by means of a model based on a Shifted Binomial random variable. In fact, this scheme seems adequate to represent the rating mechanism and the resulting model is a consistent tool for preferences analysis.
Further work will be devoted to study the possible connection with a different model based on an Inverse Hypergeometric random variable (D'Elia, 1999; 2000), whose structure resulted appropriate for ranks, too. Moreover, issues concerning selection model and goodness of fit evaluation in comparing models performance will have to be faced.

**Acknowledgements:** This work has been financially supported by MURST and University of Naples Federico II funds.

### References

Bradley, R. A. (1976). Science, Statistics, and Paired Comparisons (with discussion). *Biometrics*, **32**, 213 - 239.

D'Elia, A. (1999). A Proposal for Ranks Statistical Modelling. *Statistical Modelling* (Friedl, H., Berghold, A., Kauermann, G. eds.), Graz - Austria, 468-471.

D'Elia, A. (2000). Un modello lineare generalizzato per i ranghi: aspetti statistici, problemi computazionali e verifiche empiriche. *Italian Journal of Applied Statistics*, **12**, in press.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, B*, **42**, 109 - 127.

# Combining response categories in ordinal response models

Dee Denteneer[1], Jan Engel[2], Gerard Hollemans[1]

[1] Philips Research Laboratories, Prof. Holstlaan 4, 5656 AA Eindhoven, the Netherlands, e-mail: {Dee.Denteneer, Gerard.Hollemans}@philips.com
[2] CQM, Vonderweg 11, 5600 AK Eindhoven, the Netherlands, e-mail: egl@cqm.nl

**Keywords:** Ordinal response models, usability data, item response theory, EM

## 1 Introduction

Combining response categories in ordinal response models is a recurrent practice in statistical modelling. Often, the primary motivation to combine response categories is 'to make asymptotics work' and is invoked in order to avoid null cells. A general method to combine response categories and analysis of this practice appears to be lacking.
However, such a method is indispensable for the analysis of data with many response categories via a proportional odds (PO) model, as in a data set on the usability of user interfaces to a jukebox with which we were confronted. In it, the response data were captured by moving a slider on a computer screen with a mouse to yield a discrete valued score between 1 and 100. The fact that there are 100 response categories gives rise to complications in the analysis: the response scores are considered to be 'true' score plus noise, so that they are only partially ordered, rather than ordinal. Hence, the standard analysis via ordinal response models is invalidated. The primary question then is for a well founded method to find the number of 'true' response categories and to combine the response scores into these categories. Secondary questions pertain to the accuracy of the parameter estimates: how do biases and/or variances change when response scores are combined.
In this paper, we focus on the first question and propose a method to combine the response scores into a, small, number of categories. The method can be formally justified in the important case that the response scores are partially ordered. The theory is described in Section 2. In section 3, we show via simulation that the method works in a number of artificial situations. Also, this section reveals parameter bias in case techniques for ordinal data are applied to data that are only partially ordered. Section 4 applies the method to the jukebox data and Section 5 presents ideas for further research.

## 2 Combining response categories

We make the following assumptions to combine the response categories. First, there is a discrete valued latent variable, $Y^*$, taking values in an ordered set, whose elements will be denoted by $\{1, \ldots, C\}$. Second, the probability distribution of the random variable depends on the covariates in a way that can be described accurately by one of the ordinal response models. Third, we assume that $Y^*$ cannot be observed directly. Rather, we observe a discrete valued response $Y$, taking values in an ordered set, denoted by $\{1, \ldots, K\}$. There is a stochastic relation between latent variable and response:

$$p(k|c) := \Pr(Y = k|Y^* = c). \tag{1}$$

Informally, we assume that $K$ is large, e.g. the number of values that can be indicated by means of a slider, and that $C$ is small and equal to the intrinsic number of response categories. The response values cannot be considered to be fully ordered with respect to the latent variable $Y^*$, so that it is not correct to assume one of the ordinal response models for $Y$ directly.
Rather, the likelihood to be maximised equals

$$\prod_{i=1}^{n}\prod_{k=1}^{K} p(Y = k|x_i)^{y_{ik}} = \prod_{i=1}^{n}\prod_{k=1}^{K} (\sum_{c=1}^{C} p(Y^* = c|x_i)p(k|c))^{y_{ik}}, \tag{2}$$

with $n$ number of observations, $x_i$ covariate vector, and $y_{ik}$ indicator of $y_i = k$, and $p(Y^* = c|x_i)$ restricted by the ordinal response model. Moreover, in developing (2) it is assumed that

$$\Pr(Y = k|Y^* = c, x_i) = p(k|c), \tag{3}$$

i.e. that the dependence of the response on the covariates is through the latent variable only.
Direct maximisation of (2) is difficult. EM is possible, but the following assumption allows for a computationally simpler approach with intuitive appeal: assume that the latent variable partitions the responses into contiguous groups. Thus there is a map from observed response to latent response, $\mathcal{C} : \{1, \ldots, K\} \to \{1, \ldots, C\}$, with $\mathcal{C}(1) = 1$ and

$$\mathcal{C}(k + 1) - \mathcal{C}(k) \leq 1. \tag{4}$$

It follows that the inner sum in the likelihood (2) reduces to just one term, and the log likelihood reduces to

$$\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}(\log(p(\mathcal{C}(k)|x_i)) + \log(p(k|\mathcal{C}(k)))). \tag{5}$$

Now (5) can be optimised by enumerating (a subset of) all maps $\mathcal{C}$, and optimising conditionally on the map.

| Transformation | Alg. 2 | PO |
|---|---|---|
| $y = y^*$ | 1.01 (0.91, 1.12) | 1.00 (0.90, 1.07) |
| $y = 2y^* + b$ | 1.02 (0.92, 1.14) | 0.81 (0.71, 0.86) |
| $b \sim \text{Bin}(1, .5)$ | | |
| $y = [3y^* + \epsilon]_3^9$ | 0.95 (0.85, 1.08) | 0.76 (0.66, 0.87) |
| $\epsilon \sim N(0, 1/3)$ | | |

TABLE 1. Results obtained in simulation; 'Transformation': map from latent variables to observed responses, here $[\,]_3^9$ denotes rounding to nearest integer in the range 3 to 9; 'Alg. 2': estimates of $\theta$ (median plus first and third quartile) in 100 simulations optimising (5); 'PO': idem, but with a standard PO model for the responses. Note that the correct value for $\theta$ equals 1.

## 3 Simulations

The algorithm was applied to a series of artificial examples with data $(x_i, y_i), i = 1, \ldots, 250$, with $x_i$ a single, known, covariate (generated from a standard normal distribution) and $y_i$ a response score. This response score is obtained by transforming a latent response score, $y_i^*$, generated from a multinomial distribution with the probabilities satisfying a PO model:

$$\text{logit}(Pr(y^* \leq c)) = \alpha_c - x\theta;\ c = 1, \ldots, C-1, \tag{6}$$

with $\theta = 1$, $C = 3$, $\alpha_1 = -1.11$ and $\alpha_2 = -0.11$. The three transformations are given in Table 1, first column. They are such that the number of categories is increased and that the latent variable fully (Models 1 and 2) or approximately (Model 3) partitions the response classes, so that our assumption on this issue is at least approximately satisfied. The goal is then to recover the classes of the latent score and to estimate $\theta$. The simulations showed that the algorithm performed well in both respects. However, assuming a standard PO model for the observed responses leads to biased estimates, in case the response scores are obtained as 'true scores plus noise' (Models 2 and 3). See Table 1 and note that the correct value for $\theta$ equals 1.

## 4 Application to jukebox remote control data

The algorithm was applied to data from a study on advanced remote control of a cd jukebox; the remote controls are described in De Vet and Buil (1999). The effectiveness, among others, of three user interfaces was measured by means of three distinct items. Responses were captured by moving a slider on a computer screen with a mouse, see Hollemans (1999), to yield discrete valued scores between 1 and 100 for each item interface combination for 32 respondents. The model used is an additive PO model; the coefficients

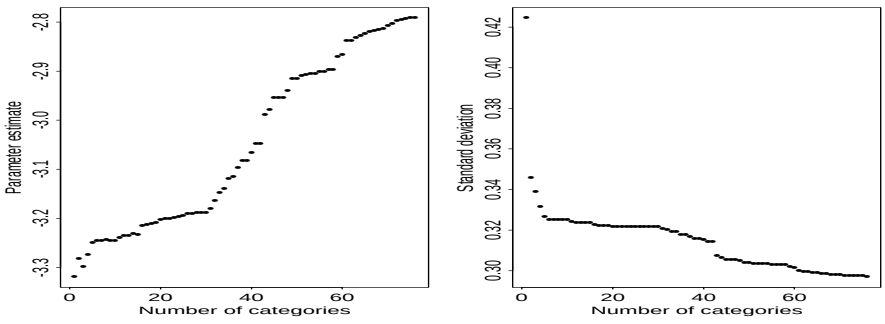



FIGURE 1. Impact of combining categories on parameter estimates and variances

corresponding to the items were found to be significantly different from zero, those corresponding to the interfaces were not.
The algorithm combined the response categories into 8 classes. It had a substantial impact on both estimate and variance of the significant item parameters; see Figure 1 for the first of these parameters. This effect on parameter estimates and variances was also noted in McCullagh and Nelder (1989, Section 5.6). For the insignificant interface parameters, the effect was on the variances only.

## 5 Conclusion

Further research into this area is called for. Of particular importance are the comparison with the more general EM based approach alluded to in Section 2, and a theoretical underpinning of the effect on bias and variance.

### References

Agresti, A. (1984). *Analysis of ordinal categorical data.* Wiley, New York.

De Vet, J. and V. Buil (1999). *A personal digital assistant as advanced remote control for audio/video equipment.*
See: `http://www.dcs.gla.ac.uk/mobile99/`.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models,* 2nd ed. Chapman & Hall, London

Hollemans, G. (1999). User Satisfaction Measurement Methodologies: Extending the User Satisfaction Questionnaire. In H.-J. Bullinger and J. Ziegler (eds.). *Human-Computer Interaction: Ergonomics and User Interfaces*, Volume 1 of the Proceedings of the 8th International Conference on Human-Computer Interaction (pp. 1008-1012). Lawrence Erlbau, Mahwah, NJ.

# Robust and Quantile Smoothing with P-splines and the $L_1$ Norm

Paul H. C. Eilers[1]

[1] Department of Medical Statistics, Leiden University Medical Center, P.O. Box 9604, 2300 RC Leiden, The Netherlands

**Abstract:** P-splines use penalized least squares regression to fit smooth curves to data. Introduction of the $L_1$ norm reduces the sensitivity to outliers. With the same norm in the penalty a linear program results. The interior point algorithm is very effective in solving the linear program. Asymmetric weights allow the estimation of smooth quantile curves.

**Keywords:** Interior point method, $L_1$ norm, linear programming, penalized estimation.

## 1 Introduction

Most smoothing methods estimate expected values. A notable exception is quantile smoothing, see Koenker, Ng and Portnoy (1994), He and Ng (1999). In this paper I study the adaptation of $L_1$ criteria to P-splines, using the sum of absolute values in both the measure of fit and the penalty. A linear program results, which can be solved efficiently with the interior point algorithm. The $L_1$ norm can be made asymmetric with different weights for positive and negative residuals, giving an effective quantile smoother without complicating the linear program.

## 2 Theory: P-splines with $L_2$ and $L_1$

P-splines (Eilers and Marx 1996)use penalized regression. The data are projected on a B-spline basis and a difference penalty on the B-spline coefficients gives continuous control over smoothness. In the least squares setting the objective function is

$$S_2 = \sum_{i=1}^{I}(y_i - \sum_{j=1}^{J} b_{ij}c_j)^2 + \lambda^2 \sum_{j=d+1}^{J} (\Delta^d c_j)^2, \tag{1}$$

where the observations are $(x_i, y_i)$, $i = 1 \ldots I$, the $I$ by $J$ matrix $B = [b_{ij}]$ is the B-splines basis and $\Delta^d$ indicates differences of order $d$. As the $L_2$ norm

is used in the measure of fit and in the penalty, I will call them P2-splines. We get P1-splines if we introduce the $L_1$ norm, the sum of absolute values:

$$S_1 = \sum_{i=1}^{I} |y_i - \sum_{j=1}^{J} b_{ij} c_j| + \lambda \sum_{j=d+1}^{J} |\Delta^d c_j|. \tag{2}$$

The reason for using the $L_1$ norm for the fit is to get better robustness. It is well known that in the one-dimensional case minimization of $\sum |y - c|$ gives the median for $c$, which is very insensitive to outliers. We hope to get the same performance in smoothing. The choice for the $L_1$ norm in the penalty is a pragmatic one: if we did keep the sum of squares, an awkward optimization problem would result, while (2) leads to a linear program, for which efficient solution algorithms are available.
The $L_1$ norm is special case of the asymmetric norm $\rho(u;\alpha)$, with $0 < \alpha < 1$:

$$\rho(u;\alpha) = \begin{cases} \alpha u & \text{if } u > 0, \\ (1-\alpha)|u| & \text{if } u \leq 0. \end{cases}$$

As a shorthand notation I use $|u|_\alpha = \rho(u;\alpha)$. One can easily show that in a one-dimensional sample minimization of $\sum |u|_\alpha = \sum \rho(y;\alpha)$ gives the (approximate) $\alpha$-th quantile. This suggests that a quantile smoother can be constructed by using the objective function

$$S_\alpha = \sum_{i=1}^{I} |y_i - \sum_{j=1}^{J} b_{ij} c_j|_\alpha + \lambda \sum_{j=d+1}^{J} |\Delta^d c_j|. \tag{3}$$

## 3 Computation

Penalized regression can be reduced to standard regression by data augmentation. Let be $D$ be a matrix such that $Dc = \Delta^d c$. Then (2) can be written as

$$S_2 = |y - Bc|^2 + \lambda^2 |Dc|. \tag{4}$$

Minimization of $S_2$ is equivalent to minimization of $|y^+ - B^+ c|^2$ if

$$B^+ = \begin{pmatrix} B \\ \lambda D \end{pmatrix}, \quad y^+ = \begin{pmatrix} y \\ z \end{pmatrix}, \tag{5}$$

with $z$ a vector of zeros of length $J - d$. The same is true for P1-splines if we minimize $|y^+ - B^+ c|$, where $|u|$ indicates the $L_1$ norm of $u$.
S-PLUS has a function `l1fit()` that can be used directly with $B^+$ and $y^+$. Unfortunately, this function can not handle the asymmetric case. But note (Portnoy and Koenker, 1997) that asymmetric $L_1$ regression can be written as a linear program (with $0 \leq u$ and $0 \leq v$):

$$\begin{array}{ll} \text{min!} & \sum_{i=1}^{I} u_i + \sum_{i=1}^{I} v_i, \\ \text{s.t.} & \sum_{j=1}^{J} b_{ij} c_j + u_i - v_i = y_i \quad i = 1 \ldots I. \end{array} \tag{6}$$

The penalty extends this problem to (with $0 \leq s$ and $0 \leq t$)

$$\begin{aligned} \text{min!} \quad & \alpha \textstyle\sum_{i=1}^{I} u_i + (1-\alpha) \sum v_i + \lambda \sum s_i + \lambda \sum_{i=1}^{I} t_i, \\ \text{s.t.} \quad & \textstyle\sum_{j=1}^{J} b_{ij} c_j + u_i - v_i = y_i, \quad i = 1 \ldots I, \\ \text{and} \quad & \textstyle\sum_{k=1}^{J} d_{jk} c_k + s_j - t_j = 0, \quad j = 1 \ldots J. \end{aligned} \tag{7}$$

Standard (simplex) linear programming code can be used to solve this problem. However, Portnoy and Koenker (1997) showed that the interior point algorithm is very effective for the computation of regression quantiles. A program (written in `Ox`) can be found on Koenker's home page at the University of Illinois: `www.econ.uiuc.edu`. I translated it to Matlab and pure S-PLUS (no linked compiled functions), and found it to perform well.

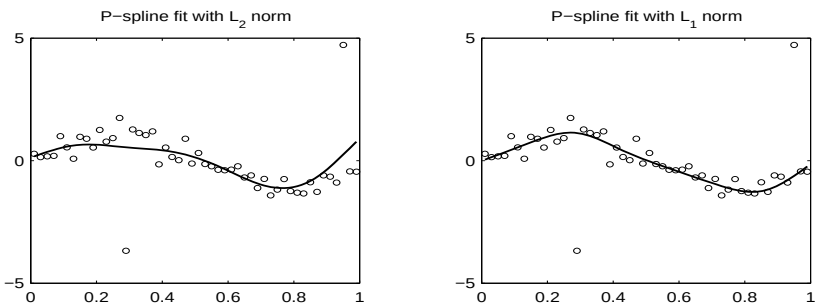



FIGURE 1. Simulated data with two outliers, fitted with 10 cubic B-splines and a second order difference penalty with $\lambda = 0.1$; left with $L_2$ norm, right with $L_1$ norm.

## 4 Results

Figure 1 shows simulated data and fits by P2-splines (left) and P1-splines (right). It is clear that the latter ignores the two outliers.
Figure 2 illustrates quantile smoothing. The data are ozone measurements near Rotterdam in the Netherlands, in 1992. The 50, 80 and 90 percentile curves were estimated.

## 5 Discussion

Compared to smoothing splines (Koenker *et al.*, 1994), P1-splines have the advantage of the reduced number of equations. Also smoothing $L_1$ splines are piecewise linear, whereas the B-splines can have any degree (there is no danger of zero derivatives at low degree). He and Ng (1999) use first or second order derivatives of the B-splines in the penalty, which complicates

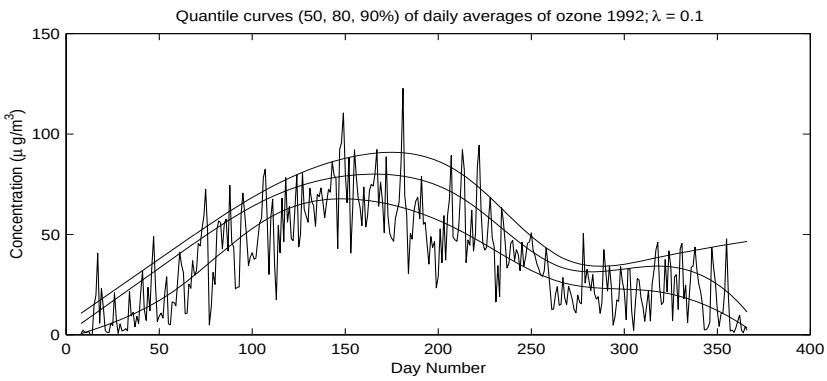



FIGURE 2. Quantile curves for ozone concentration measurements.

the algorithm. The difference of coefficients is easier to handle: extension to higher order is a mechanical process and the degree of the B-splines can be arbitrary. He and Ng use a classical linear programming code, implemented in FORTRAN and linked to S-PLUS. This complicates porting to different computing environments. The use of the interior point algorithm allows implementation in pure S-PLUS (and Matlab) with adequate performance. He and Ng also discuss the incorporation of monotonicity and convexity constraints by extending the linear program. The algorithm of this paper can be extended in the same way. An interesting alternative is to use "soft" constraints, in the form of asymmetric penalties, as was done by Eilers (1994) for difference smoothing with an $L_2$ norm. This amounts to adding terms of the form $\gamma|\Delta^q c|_\alpha$ to the penalty

## References

Eilers P.H.C. (1994) Sign-constrained, monotone and convex nonparametric regression with asymmetric penalties. *Proceedings of the 9th International Workshop on Statistical Modelling.*

Eilers, P.H.C. and Marx, B.D. (1996) Flexible smoothing with splines and penalties (with discussion). *Statistical Science*, **11**, 89–121.

He, X. and Ng, P. (1999) COBS: qualitatively constrained smoothing via linear programming. *Computational Statistics*, **14**, 315–337.

Koenker, R. Ng, P. and Portnoy, S. (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–680.

Portnoy, S. and Koenker, R. (1997) The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators (with discussion). *Statistical Science*, **12**, 279–296.

# Estimation of the exponent of a fractional brownian noise

Consuelo García Tejedor and Frederic Utzet

[1] Departament de Matemátiques, Universitat Autonoma de Barcelona, 08193 Bellaterra (Barcelona), Spain, e-mail: consuelo@mat.uab.es, utzet@mat.uab.es

**Abstract:** A new way to estimate the exponent of a fractional brownian noise is shown. Consistence and assymptotic normality of the estimator are proved and, through simulation studies, a comparison between our estimator and the R/S, maximum likelihood and Bardet estimators is carried.

**Keywords:** Long memory process, fractional brownian motion

## 1 Long memory processes

The typical models for stationary time series, for example the ARMA models, presuppose that the time series has a short memory, that means, $\sum_k |\gamma(k)| < \infty$, where $\gamma(k)$ is the autocovariance function at lag $k$. However, begining by the pioneering work of Hurst (1951) concerning the level of the river Nil (using a very large data set –from 622 to 1281 A.D.) and continuing with the works of Mandelbrot, models with long memory –the series $\sum_h \gamma(h)$ is divergent– have been considered. Roughly speaking, in a short memory time series, two distant variables are almost independent; on the other hand, in a long memory time series very long cycles exist. The long memory time series are becoming more and more important; first because there are, in reality, lots of situations with long cycles (in hydrology, economics or other sciences), and secondly, because the term *long cycle* depends on the time scale, and at present time, the so-called *hight-frequency data*, that is, data which are taken in almost continuous time (for example, the value of a stock recorded each minute) allow for the observation of *long cycle* in a short period of time.

## 2 The fractional brownian noise

The first, main and simplest example of a long memory process is *fractional brownian noise*, $\{X_n,\, n \geq 1\}$: it is a centered stationary gaussian process with covariance function given by

$$\gamma(k) = \frac{1}{2}\,\sigma^2\big(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\big), \quad k \in \mathbf{Z}.$$

The number $H \in (0,1]$ is called the *exponent* of the fractional noise; when applied, the interesting case is $H \in [0.5,1)$. For $H = 1/2$, we find $\gamma(k) = 0, \quad \forall k \neq 0$, and since it is a gaussian series, we obtain that all the variables are independent; that is, $\{X_n,\, n \geq 1\}$ is a sequence of i.i.d. gaussian random variables. In the same way as it is possible to construct a sequence of i.i.d. gaussian random variables from a brownian process, we can obtain a fractional brownian noise from the continuous time process called the *fractional brownian process.*
On the other hand, a fractional brownian noise is the paradigmatic example of a *self-similar process*, which is the stochastic translation of a fractal curve: The series $\{X_t,\, t \geq 1\}$ satisfies

$$X_t \stackrel{\mathcal{L}}{=} t^H X_1.$$

## 3 The estimation of the exponent H

From the statistical point of view, the main problem in dealing with fractional brownian motion is the estimation of the exponent $H$. In his work, Hurst (1951) proposes the so-called $R/S$ method, which is very intuitive and appealing, but there is no easy asymptotic theory for it, which is a drawback for its use for statistical inference. Another possibility is the maximum likelihood estimator of $H$ based in the gaussian character of the series. However, the series which we are studying are very long, and therefore it is difficult to work with the likelihood function, and it is necessary to use the Whitle approximations. Beran (1994) gives a complete survey of this methodology. Recently, Bardet (1998) has proposed a new method using the properties of the autocovariance function. Here, we propose a technique which is similar to the Bardet one but, in our opinion, simpler and computationally easier.

## 4 Estimation of $H$ by non linear regression

The main ideas of our method for estimating $H$ are the following.
Consider $n$ observations $X_1, \ldots, X_n$ from a fractional brownian noise of parameter $H$. We will use the autocorrelation function $\rho(k) = \gamma(k)/\gamma(0)$ instead of the autocovariance function:

$$\rho(k) = \frac{1}{2}\left(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\right). \tag{1}$$

Define the *sample autocorrelation* at lag $k$, $\widehat{\rho}(k)$, by

$$\widehat{\rho}_n(k) = \frac{\sum_{j=1}^{n-k}\left(X_j - \overline{X}_n\right)\left(X_{j+k} - \overline{X}_n\right)}{\sum_{j=1}^{n}\left(X_j - \overline{X}_n\right)^2},$$

where $\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j$. As is well known, for a large collection of time series, $\widehat{\rho}_n(k)$ is a consistent asymptotically normal estimator of $\rho(k)$ . Here, we use a powerful theorem of Arcones (1994) and the delta method to prove that for $H \in [0.5, 0.75)$ the convergence in law of the stochastic process $\Big\{\sqrt{n}\big(\widehat{\rho}_n(k) - \rho(k)\big),\, n \geq 1\Big\}$ to a centered gaussian process $\{\xi_n,\, n \geq 1\}$ with covariance function

$$\mathrm{Cov}(\xi_k, \xi_l) = \sum_{r=1}^{\infty} \big\{\rho(r+k)+\rho(r-k)-2\rho(r)\rho(k)\big\}\big\{\rho(r+l)+\rho(r-l)-2\rho(r)\rho(l)\big\}.$$

Then we write,

$$\left.\begin{array}{ll} \widehat{\rho}_n(k_1) & \approx \rho(k_1) + \dfrac{1}{\sqrt{n}}\,\xi(k_1) \\ \\ \widehat{\rho}_n(k_2) & \approx \rho(k_2) + \dfrac{1}{\sqrt{n}}\,\xi(k_2) \\ \vdots & \qquad\quad \vdots \\ \widehat{\rho}_n(k_\ell) & \approx \rho(k_\ell) + \dfrac{1}{\sqrt{n}}\,\xi(k_\ell) \end{array}\right\}$$

and the nonlinear function $\rho(k)$ is given by (1). Then, we use non linear regression to estimate $H$. The initial value of $H$ is taken from the expression of $\rho(1)$; specifically, we take

$$\widehat{H}^{(0)} = \frac{1}{2}\Big(\log_2\big(1 + \widehat{\rho}_n(1)\big) + 1\Big).$$

There are different points which is necessary to consider carefully: the random variables $\xi(k_1), \ldots, \xi(k_\ell)$ are not independent or the choice of $k_1, \ldots, k_\ell$. To deal with the first point, we use weighted least squares through the Cholesky decomposition of the covariance matrix of the errors.

In table 1 there are the results of a study where 42 fractional brownian motions of sizes ranging from 2.000 to 4.000 and diferent values of the parameter $H$ were simulated; the autocorrelacion at the first 10 lags were used in the non-linear regression; in the column with the label $S.D.$ there are the standard deviations of the estimators $\widehat{H}$ obtained with the different simulations. In addition, see in Table 2 a comparison between our estimator (denoted by $\widehat{H}$) and the $R/S$ (denoted $\widehat{H}_{R/S}$), maximum likelihood ($\widehat{H}_{ML}$) and Bardet ($\widehat{H}_B$) estimators.

TABLE 1. Results of the simulation studies

| $H$ | $\widehat{H}$ | $S.D.$ |
|---|---|---|
| 0.53 | 0.529 | 0.011 |
| 0.57 | 0.566 | 0.006 |
| 0.60 | 0.598 | 0.005 |
| 0.63 | 0.630 | 0.008 |
| 0.69 | 0.681 | 0.005 |
| 0.72 | 0.722 | 0.013 |
| 0.74 | 0.746 | 0.022 |

TABLE 2. Comparison between estimators

| $H$ | $\widehat{H}$ | S.D. | $\widehat{H}_{R/S}$ | S.D. | $\widehat{H}_{ML}$ | S.D. | $\widehat{H}_B$ | S.D. |
|---|---|---|---|---|---|---|---|---|
| 0.53 | 0.529 | 0.011 | 0.575 | 0.028 | 0.529 | 0.012 | 0.510 | 0.018 |
| 0.57 | 0.566 | 0.006 | 0.617 | 0.028 | 0.568 | 0.005 | 0.559 | 0.016 |
| 0.60 | 0.598 | 0.005 | 0.617 | 0.021 | 0.598 | 0.005 | 0.597 | 0.013 |
| 0.63 | 0.630 | 0.008 | 0.651 | 0.015 | 0.631 | 0.008 | 0.628 | 0.018 |
| 0.69 | 0.681 | 0.005 | 0.701 | 0.026 | 0.684 | 0.005 | 0.675 | 0.012 |
| 0.72 | 0.722 | 0.013 | 0.731 | 0.019 | 0.724 | 0.0114 | 0.724 | 0.016 |
| 0.74 | 0.747 | 0.022 | 0.728 | 0.037 | 0.751 | 0.022 | 0.755 | 0.021 |

## References

Arcones, M. A. (1994). Limit theorems for nonlinear functionals of a stationary gaussian sequence of vectors. *Annals of Probability*, **22**, **N. 4**, 2242–2274.

Bardet, J. M. (1999). Un test d´ autosimilarité pour les processus gaussiens á accroissements stationaires. *Probability Theory*, serie I, 521-526.

Beran, J. (1994). *Statistics for Long Memory Processes.* Chapman and Hall: New York.

Hurst, H. E. (1951). Long term storage capacity of reservoirs. *Trans. Am. Soc. Civil Engin.* , **116**, 770–799.

Mandelbrot, B. B. and van Ness, J. W. (1968). Fractional brownian motion, fractional noises and applications. *Siam Rev.*, **10**, **N. 4**, 422–437.

# Sequential Diagnosis of Shocks in Dynamic Linear Models

Pilar Gargallo[1] and Manuel Salvador[1]

[1] Dpto. Métodos Estadísticos , Universidad de Zaragoza
[2] Address of second author

**Abstract:** This paper considers an automatic sequential monitoring and intervening scheme of the one-step forecast errors in a Dynamic Linear Model, generalizing that proposed by West and Harrison (1997). The methodology is illustrated with the well-known Nile data provided by Cobb (1978).

**Keywords:** Dynamic Linear Models; Sequential Tests; Outliers, Interventions, Bayesian Inference

## 1 Introduction

Dynamic Linear Models (DLM), first introduced by Harrison and Stevens (1976), provide an important tool to model forecast systems in a time series context. In this kind of system, it is necessary to be open to possible changes in the current model, because the more vigilant and succesful the forecaster is in anticipating major events, the more effective are the decisions.
This paper discusses automatic sequential monitoring methods of the forecasting activity to detect breakdowns. The basic idea, given in West and Harrison (1997), is very simple: when an observation is made, the consistency of forecast and outcome for the routine model is compared with the consistency of alternative models that represent breakdowns of interest, for example, an outlier, a level increase or a change in the slope of the series. In this paper, the family of alternative models proposed by De Jong and Penzer (1998) is used. This family uses the addition of shocks to model a large range of potential structural changes.

## 2 The Problem

Let the routine model $M_0$ be a standard normal DLM given by:

$$\begin{aligned} Y_t &= F_t'\theta_t + v_t \\ \theta_t &= G_t\theta_{t-1} + w_t \end{aligned}$$

where $v_t \sim N(0, V)$ is the observational error and $w_t \sim N(0, VW_t)$ is the evolution error. It is assumed that for all t and all s with $t \neq s$, $v_t$ and $v_s$ are independent, $w_t$ and $w_s$ are independent, and $v_t$ and $w_t$ are independent. For clarity, it is also assumed that the variances V and $W_t$ are known for each time t. The initial information is given by $\theta_0/D_0, V \sim N(m_0, VC_0)$. The observational and evolution error sequences are independent of $\theta_0/D_0, V$.
With $D_t$ denoting the information set at time t, the one-step forecast and posterior distributions are given, for each t, as follows:
Posterior at t-1: $(\theta_{t-1}/D_{t-1}, V) \sim N(m_{t-1}, VC_{t-1})$
Prior:$(\theta_t/D_{t-1}, V) \sim N(a_t, VR_t)$ with $a_t = G_t\, m_{t-1}$, $R_t = G_t\, C_{t-1}G_t' + W_t$
Forecast:$(Y_t/D_{t-1}, V) \sim N(f_t, VQ_t)$ with $f_t = F_t' a_t$, $Q_t = F_t' R_t F_t + 1$
Posterior at t: $(\theta_t/D_t) \sim N(m_t, VC_t)$ with $m_t = a_t + A_t e_t$, $C_t = R_t - A_t A_t' Q_t$ where $A_t = R_t F_t Q_t^{-1}$ and $e_t = Y_t - f_t$
Assessing the consistency of the observed values $Y_t$ is essentially equivalent to assessing the consistency of the forecast errors. Thus, under $M_0$, the forecast distribution for $e_t$ is simply $(e_t/D_{t-1}, V) \sim N(0, VQ_t)$.
The alternative model $M_1$ is defined as in De Jong and Penzer (1998) and is given by:

$$\begin{aligned} Y_t &= F_t'\theta_t + X_t\delta + v_t \\ \theta_t &= G_t\theta_{t-1} + H_t\delta + w_t \end{aligned}$$

where $X_t$ and $H_t$ are called the shock design matrices and $\delta$ is the shock magnitude. Many departures observed in practice can be represented by a simple intervention, that is to say, by taking all $X_t$ and $H_t$ to be 0 except at a single point t = s where $H_s$ = H and $X_s$ = X . The shock design matrices X and H determine the type of simple intervention resulting from a shock.
In this particular case, it is possible to show that under $M_1$,

$$(e_{s+k}/D_{s+k-1}, \delta, V, M_1) \sim N(x_{s+k}\delta, VQ_{s+k})$$

where $x_{s+k}$ = $r_{s+k}$ - $g_{s+k}$ with

$$r_{s+k} = \begin{cases} F_s'H + X & k = 0 \\ F_{s+k}'d_{s+k} & k \geq 1 \end{cases} \qquad d_{s+k} = \begin{cases} H & k = 0 \\ G_{s+k}d_{s+k-1} & k \geq 1 \end{cases}$$

and

$$g_{s+k} = \begin{cases} 0 & k = 0 \\ F_{s+k}'b_{s+k} & k \geq 1 \end{cases} \qquad b_{s+k} = \begin{cases} 0 & k = 0 \\ G_{s+k}n_{s+k-1} & k \geq 1 \end{cases}$$

$$n_{s+k} = \begin{cases} A_s(F_s'H + X) & k = 0 \\ b_{s+k} + A_{s+k}(r_{s+k} - g_{s+k}) & k \geq 1 \end{cases}$$

When an observation is made, the Bayes Factors for the standard model is computed against each of the alternatives. If any of the Bayes Factors show evidence against the standard model, a signal is generated to alert the forecaster to a potential problem; if change is deemed appropriate, subjective intervention is made at that time. By contrast, if there is no

clear evidence against the routine model, then it is given the benefit of the doubt and no signal is generated.
If the prior distribution of $\delta$ is $(\delta/D_{s-1}, V) \sim N(\lambda_{0s}, q_{0s})$, then the cumulative Bayes Factor is given by:

$$H_s(k) = \frac{P_0\left(e_s, ..., e_{s+k}/D_{s-1}, V\right)}{P_1\left(e_s, ..., e_{s+k}/D_{s-1}, V\right)} = \left(\frac{q_{0s}}{q_{s+k}}\right)^{1/2} \exp\left\{\frac{1}{2V}\left(\frac{\lambda_{0s}^2}{q_{0s}} - \frac{\lambda_{s+k}^2}{q_{s+k}}\right)\right\}$$

where $P_0$ and $P_1$ are the forecast distribution under $M_0$ and $M_1$ respectively and
$\lambda_{s+k} = q_{s+k}\left\{\frac{\lambda_{0s}}{q_{0s}} + \sum_{i=0}^{k} \frac{x_{s+i}e_{s+i}}{Q_{s+i}}\right\}$ $\qquad q_{s+k} = \left\{\frac{1}{q_{0s}} + \sum_{i=0}^{k} \frac{x_{s+i}^2}{Q_{s+i}}\right\}^{-1}$
This factor is used to analyse the forecast behavior of the model in the last k observations.
The monitoring process starts after a learning period in which initial estimations of the parameters of the standard model are obtained. Every time that a new observation, $y_s$, is incorporated, it is analysed whether it is in discrepancy with the model, or not, specifying a lower limit $\tau_1$ for the bayes factor $H_s(1)$. If the new observation is in discrepancy with the model, it becomes a doubtful intervention until new information indicates whether or not it must be incorporated into the model , or whether it must be discarded.
This decision is made from the value of $H_s$(k) with $k \geq k_{min}$, where $k_{min}$ is a minimum number of observations that are considered to be sufficient to make the decision, and using the constants 1 and $\tau_2$ as the upper and lower limit for the bayes factor $H_s$(k), respectively. Finally, the algorithm analyses if it is possible to simplify the model, removing some of the included interventions by the use of a lower limit $\tau_3$ for $H_s$(k).

## 3 Application

To illustrate the monitoring scheme, consideration is given to the Nile River data taken from Cobb (1978). These data consist of readings of the annual flow volume of the Nile River at Aswan from 1871 to 1970. The series has been examined by Pole et al. (1994) who concluded that a permanent decline in volume has taken place from 1899 onwards. Additionally, De Jong and Penzer (1998) indicated outlying values in 1877 and 1913.
In this paper a local model is proposed:

$$\begin{aligned} Y_t &= \mu_t + v_t \quad \text{with} \quad v_t \sim N(0, V) \\ \mu_t &= \mu_{t-1} + w_t \quad \text{with} \quad w_t \sim N(0, VW_t) \end{aligned}$$

The evolution variance is built using the discount factor methodology given by West and Harrison (1997), the observational variance is estimated at each time and a priori vague initial is specified. Outliers and level changes

are monitored by taking $X = 1$, $H = 0$ and $X = 0$, $H = 1$, respectively, in the alternative model.
If $\tau_1 = 0.7$, $\tau_2 = 0.5$ and $\tau_3 = 1$, the proposed algorithm detects a level change in the 29th observation and outlying values in the 7th, 9th, 18th, 43rd and 94th observations.

## References

Cobb, G. W. (1978), "The problem of the Nile: Conditional solution to a change point problem". *Biometrika*, **65**, 243-251.

De Jong, P. and Penzer, J. (1998), "Diagnosing shocks in Time Series". *Journal of the American Statistical Association*, **93**, 796-806.

Harrison, J. and Stevens, C. (1976), "Bayesian Forecasting (with discussion)". *J. R. Statit. Soc. B*, **38**, 205 - 247

Pole, A. West, M. and Harrison, J. (1994),*Applied Bayesian Forecasting and Time Series Analysis*, Chapman&Hall.

West,M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, Springer. $2^a Ed.$

# Estimating functions based on the modified directed likelihood

Federica Giummolè[1], Laura Ventura[2] and Alessandra Salvan[2]

[1] Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino (Italy); e-mail: giummole@calvino.polito.it

[2] Dipartimento di Statistica, Università di Padova, Via S. Francesco 33, I-35121 Padova (Italy); e-mail: ventura@stat.unipd.it

**Abstract:** This paper discusses a simple adjustment of the likelihood equation for a scalar parameter of interest, based on a higher-order modification of the signed square root of the log-likelihood ratio statistic. The estimator obtained is a refinement of the maximum likelihood estimator and is equivariant under reparameterisation. A simple explicit approximate version of it is also derived.

**Keywords:** ancillary statistic; estimating equation; nuisance parameter; modified directed likelihood; parameterisation equivariance.

## 1 Introduction

This paper discusses a practical approach to point estimation when inference about an arbitrary one-dimensional parameter of interest in the presence of nuisance parameters is desired. The approach is based on some recent results that use modified likelihood quantities as estimating functions for a scalar parameter of interest (Pace and Salvan, 1999). In particular, in the context of multiparameter exponential families, Pace and Salvan (1999) show that the estimating function based on the modified directed likelihood $r^*$, introduced by Barndorff-Nielsen (1986), gives higher-order approximations of the optimal median unbiased estimator (Lehmann, 1986, § 5.6). The modified directed likelihood is a higher-order adjustment of the signed square root $r$ of the usual generalised log-likelihood ratio statistic; we refer to $r$ as the directed likelihood. The modified directed likelihood is defined for arbitrary multiparameter models under broad regularity assumptions. The aim of this paper is to study the estimating equation based on $r^*$ outside exponential families, generalising the results in Pace and Salvan (1999). Since the estimating equation based on $r$ gives the maximum likelihood estimator (m.l.e.), what we are considering here is essentially a correction to the m.l.e., which improves its small sample properties, encompassing as far as possible the requirement of parameterisation equivariance. Other commonly used modifications of the m.l.e. are bias corrections (see e.g. Barndroff-Nielsen and Cox, 1994, § 6.4). However, they do not satisfy

the requirement of equivariance under reparameterisation (see Skovgaard, 1989).

## 2 Estimators based on $r^*$

Let us consider a parametric model with density function $p(x;\omega)$ and log-likelihood $l(\omega) = l(\omega;x)$ under random sampling of size $n$. Assume that the $d$-dimensional parameter $\omega$ is of the form $\omega = (\psi, \lambda)$, where $\psi$ is a scalar parameter of interest and $\lambda$ a $(d-1)$-dimensional nuisance parameter. Let $\hat{\omega} = (\hat{\psi}, \hat{\lambda})$ be the m.l.e. of $\omega$ and $\hat{\lambda}_\psi$ be the m.l.e. of $\lambda$ for a given value of $\psi$. The generic components of $\lambda$ and $\hat{\lambda}$ are denoted by $\lambda_a, \lambda_b, \ldots$ and $\hat{\lambda}_a, \hat{\lambda}_b, \ldots$, respectively, with $a, b, \ldots = 1, \ldots, d-1$. The profile log-likelihood for $\psi$ is $l_P(\psi) = l(\psi, \hat{\lambda}_\psi)$. As in Barndorff-Nielsen and Cox (1994), a tilde over a likelihood quantity is used when the quantity is evaluated at $(\psi, \hat{\lambda}_\psi)$, while a hat denotes evaluation at $(\hat{\psi}, \hat{\lambda})$.
Assume that $a$ is an ancillary statistic, either exactly or at least to an appropriate order of approximation, such that $l(\omega;x) = l(\omega;\hat{\omega}, a)$. The modified directed likelihood $r^*$ for $\psi$ (see also Barndorff-Nielsen and Cox, 1994, § 6.6) is

$$r^*(\psi) = r + \frac{1}{r} \log \frac{cu}{r} \ , \tag{1}$$

where $r = r(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)\{2(l_P(\hat{\psi}) - l_P(\psi))\}^{1/2}$, $c = c(\psi) = |\tilde{l}_{\lambda;\hat{\lambda}}|/\{|\tilde{\jmath}_{\lambda\lambda}||\hat{\jmath}_{\lambda\lambda}|\}^{1/2}$ and $u = u(\psi) = \hat{\jmath}_P^{-1/2}\partial(l_P(\hat{\psi}) - l_P(\psi))/\partial\hat{\psi}$. In the definition of $c$, the matrix $l_{\lambda;\hat{\lambda}} = [l_{a;b}]$ has generic element $l_{a;b} = \partial^2 l/(\partial\lambda_a \partial\hat{\lambda}_b)$. In addition, $j(\omega)$ denotes the observed information matrix, its block corresponding to $\lambda$ is $j_{\lambda\lambda}(\omega) = [j_{ab}] = [-l_{ab}]$, with $l_{ab} = \partial^2 l/(\partial\lambda_a \partial\lambda_b)$, and $j_P(\psi)$ is the observed profile information, i.e. $j_P(\psi) = -l_{P2}(\psi) = -\partial^2 l_P(\psi)/\partial\psi^2$.
The modified directed likelihood is a higher-order pivotal quantity, having a null standard normal distribution with error of order $O(n^{-3/2})$. Moreover, following Pace and Salvan (1999), it gives rise to a simple estimating equation of the form

$$r^*(\psi) = 0 \ . \tag{2}$$

Since $r^*$ is invariant under interest respecting reparameterisations, the solution to (2), denoted by $\hat{\psi}^*$, is equivariant. A numerical procedure is usually required in order to solve (2). Observe that if only the leading term of (1) is used in (2), the solution is trivially $\hat{\psi}$. Hence, the estimator $\hat{\psi}^*$ is a refinement of $\hat{\psi}$, with the estimating equation (2) giving implicitly a correction of order $O_p(n^{-1})$, as will be shown in the next section.

## 3 Explicit asymptotic version of $\hat{\psi}^*$

In this section we obtain an asymptotic expansion for $\hat{\psi}^*$, having the form

$$\hat{\psi}^* = \hat{\psi} + \hat{m} + O_p(n^{-3/2}) = \hat{\psi}^*_A + O_p(n^{-3/2}) \ , \tag{3}$$

where $m = m(\psi, \lambda)$ is a correction term of order $O_p(n^{-1})$ and $\hat{\psi}^*_A = \hat{\psi} + \hat{m}$. In the following, it is convenient to use index notation and Einstein summation convention. For the partial derivatives of $l(\psi, \lambda)$, we write $j_{a\psi} = -l_{a\psi} = -\partial^2 l/(\partial\lambda_a \partial\psi)$, $j_{ab\psi} = -l_{ab\psi} = -\partial^3 l/(\partial\lambda_a \partial\lambda_b \partial\psi)$, $j_{abc} = -l_{abc} = -\partial^3 l/(\partial\lambda_a \partial\lambda_b \partial\lambda_c)$, $l_{ab;c} = \partial^3 l/(\partial\lambda_a \partial\lambda_b \partial\hat{\lambda}_c)$ and $l_{a\psi;b} = \partial^3 l/(\partial\lambda_a \partial\psi \partial\hat{\lambda}_b)$. A generic element of $j^{-1}_{\lambda\lambda} = [j_{ab}]^{-1}$ is denoted by $j^{ab}$. In addition, we write $l_{P3} = \partial^3 l_P/\partial\psi^3$ and $l_{P2;1} = \partial^3 l_P/(\partial\psi^2 \partial\hat{\psi})$ for the derivatives of $l_P(\psi)$.
A straightforward Taylor expansion of (2) around $\hat{\psi}$ gives, for $(\hat{\psi} - \psi) = O_p(n^{-1/2})$,

$$r^*(\psi) = r^*(\hat{\psi}) + (\psi - \hat{\psi}) \left. \frac{\partial}{\partial\psi} r^*(\psi) \right|_{\psi=\hat{\psi}} + O_p(n^{-1}) \ . \tag{4}$$

Generalising the calculations of Barndorff-Nielsen (1986) when no nuisance parameters are present and using results in Sartori (1997; unpublished Graduation Thesis, University of Padova, Italy), we find that

$$\begin{aligned} r^*(\hat{\psi}) &= \frac{1}{2\hat{\jmath}_P^{1/2}} \hat{\jmath}^{ab} \left( \hat{\jmath}_{ab\psi} - 2\hat{l}_{a\psi;b} - \hat{\jmath}_{abc}\hat{\jmath}^{cd}\hat{\jmath}_{d\psi} + 2\hat{l}_{ac;b}\hat{\jmath}^{cd}\hat{\jmath}_{d\psi} \right) \\ &- \frac{1}{6\hat{\jmath}_P^{3/2}} [\hat{l}_{P3} + 3\hat{l}_{P2;1}] + O_p(n^{-1}) \end{aligned}$$

and $\partial r^*(\psi)/\partial\psi|_{\psi=\hat{\psi}} = -\hat{\jmath}_P^{1/2} + O_p(n^{-1/2})$. Solving for $\psi$ the resulting linearized estimating equation, we obtain expansion (3) for $\hat{\psi}^*$ with

$$\hat{m} = \frac{\hat{\jmath}^{ab}}{2\hat{\jmath}_P} \left( \hat{\jmath}_{ab\psi} - 2\hat{l}_{a\psi;b} - \hat{\jmath}_{abc}\hat{\jmath}^{cd}\hat{\jmath}_{d\psi} + 2\hat{l}_{ac;b}\hat{\jmath}^{cd}\hat{\jmath}_{d\psi} \right) - \frac{1}{6\hat{\jmath}_P^2} \left[ \hat{l}_{P3} + 3\hat{l}_{P2;1} \right] \ .$$

While the estimator $\hat{\psi}^*$, obtained as the solution of (2), is exactly equivariant under interest respecting reparameterisations, the approximate estimator $\hat{\psi}^*_A$ is not, since the linearized estimating equation (4) depends on the parameterisation. However, in scale and location (or, in general, regression) models it can be easily shown that the approximate estimator $\hat{\psi}^*_A$ is equivariant under scale and location (or regression) transformations.
The calculation of the correction term $m$ is straightforward, since it only involves derivatives of the log-likelihood function evaluated at the m.l.e.. On the contrary, to solve (2), which gives $\hat{\psi}^*$, a numerical procedure is usually required.

## 4 A simulation study and final remarks

In this section we present a simulation study which compares numerically the estimators $\hat{\psi}$ and $\hat{\psi}^*$ of the regression coefficient $\psi = \beta_1$ of the scale and regression model $y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$, $i = 1, \ldots, 10$, where $(\epsilon_1, \ldots, \epsilon_{10})$ is a random sample from the extreme value distribution. The estimators are compared both in terms of the usual centering and dispersion measures, i.e. bias (BI) and mean square error (MSE), and in terms of the probability of understimation (PU), which gives median bias. The results are based on 10000 Monte Carlo trials (using results in Brazzale, 1999) and, as it can be noted, the proposed estimator is particularly accurate and improves on the m.l.e..

| $\hat{\psi}$ | | | $\hat{\psi}^*$ | | |
|---|---|---|---|---|---|
| BI | MSE | PU | BI | MSE | PU |
| 0.014 | 0.038 | 0.480 | 0.001 | 0.038 | 0.500 |

At present, there are several questions that require further study, for instance:
(a) the theoretical properties of the estimators $\hat{\psi}^*$ and $\hat{\psi}^*_A$ should be discussed in more detail;
(b) when the ancillary statistic $a$ is not available, approximations of the sample space derivatives (Skovgaard, 1996) should be considered.

**Notes.** This work has been partially supported by grants from Ministero dell'Università e della Ricerca Scientifica e Tecnologica, Italy.

### References

Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log-likelihood ratio, *Biometrika*, **73**, 307-322.

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics*, Chapman and Hall, London.

Brazzale, A.R. (1999). Conditional simulation for regression-scale models, *Journ. Italian Stat. Soc.*, to appear.

Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd ed., Wiley, New York.

Pace, L. and Salvan, A. (1999). Point estimation based on confidence intervals: exponential families, *J. Statist. Comput. Simul.*, **64**, 1-21.

Skovgaard, I.M. (1989). A rewiew of higher order likelihood inference, *Bull. Int. Statist. Inst.*, **53**, 331-351.

Skovgaard, I.M. (1996). A general large deviation approximation to one-parameter tests, *Bernoulli*, **2**, 145-165.

# Application of Multivariate Techniques to characterize the Structure of the Beechwoods of Burgos Province (Spain)

Concepción González García[1], Angel Martín Fernández[1], Alvaro Sánchez de Medina Garrido[1], Esperanza Ayuga Téllez[1] and Susana Martín Fernández[1]

[1] Unidad docente de Estadística e Investigación Operativa. Departamento de Economía y Gestión de las Explotaciones e Industrias Forestales. E.T.S. de Ingenieros de Montes. Universidad Politécnica de Madrid. Ciudad Universitaria s/n. 28040-Madrid.

**Abstract:** The aims of this work are: to characterise the forest structure of the beechwoods of Burgos province (Spain), by using the least number of variables in this characterization, to group the different vegetal structures of the studied plots in a little number of forest systems. The statistical techniques used were: sampling, exploratory data, correspondence and cluster analysis. These multivariate techniques can be a very useful and practical tool to the forest manager in the expansion areas of beeches in the study area.

**Keywords:** Data analysis, Multivariate techniques, Forest structure, Statistical classification.

## 1 Introduction

The aim of this work is the characterisation of different forest structures in the beechwoods of Burgos province (Spain). Easily measurable variables are used in this study. Multivariate statistical techniques have result very useful to enable this characterization .

## 2 Data

The first data base used came from the $2^{nd}$ National Forestry Inventory ($2^{nd}$ NFI) in Burgos province (Spain) (1986-1995). They were measured in round plots of variable radius (between 5 and 25 m, according to the size of DBH (Diameter at Breast Height, 1.30 m from the ground)). The considered plots were:

- Plots with trees with diameter $\exists$ 7.5 cm ("big trees").

- All the plots with beech regeneration included in the 2nd NFI (the beech regeneration is made up by plants with diameters between 2.5 and 7.5 cm ("small trees")).

The total number of the used plots were 263, all of them with beechwoods presence.They were classified at the first time as follows:

- 201 plots have beeches with diameter $\exists$ 7.5 cm: 166 plots with regeneration and 35 only with big trees without regeneration.
- 62 have only regeneration.
- 228 plots have beechwoods regeneration areas.
- 253 plots have big trees of any other encountered specie besides beechwoods.

The principal variable in the forest inventory is DBH and to the data analysis is grouped like:
CD1: 75 - 124 mm; CD2: 125 - 224 mm; CD3: 225 - 324 mm; CD4: 325 - 424 mm; CD5: 425 - 524 mm; CD6: $\exists$ 525 mm.
Furthermore, to facilitate the modelization, small trees were sistematically assigned to the regeneration group, not lose information.
Moreover, different interesting variables were measured in the plots: number of steams per ha and per diameter class for each specie; basal area (cross section area at 1.30 m from ground) of each specie per ha and per diameter class; and number of trees that belong to the regeneration class per ha of all present species. These variables were selected because of their easy measure in the field and their capability to offer a clear knowledge about structure of the studied forest mass.

## 3 Multivariate Data Analysis and Results

Six non hierarchical cluster analysis were achieved according to different criteria of grouping variable. The variable which best grouped the plots and best information gave about structure of forest mass were the basal area. On the other hand, the variable "number of steams per ha" resulted discriminant to classify the plots by diameter class. These two groups of variables are the starting point to other data analysis to classify the forest mass, like correspondence analysis. Figs. 1 and 2 present results only of beechwood.

## References

Anderberg, M.R. (1973). *Cluster Analysis for Applications.* Academic Press. New York.

Carrasco, J.L. and Hernán, M.A. (1993). *Estadística Multivariante.* Ed. Ciencia 3. Madrid.

Cuadras, C.M. (1981). *Métodos de Análisis Multivariante.* Ed. Universitaria de Barcelona (EUNIBAR).

Oliver, C.D. and Larson, B.C. (1996). *Forest Stand Dynamics.* Wiley. New York.

Radlof, D.L. and Betters, D.R. (1978). Multivariate analysis of physical site data for Wildland classification. *Forest Science.* Vol. **24**, n.1; 2-10.

Weintraub, A.; Saez, G. and Yadlin, M. (1997). Aggregation procedures in Forest Management Planning using Cluster Analysis. *Forest Science.* Vol. **43**, n.2; 274-285.

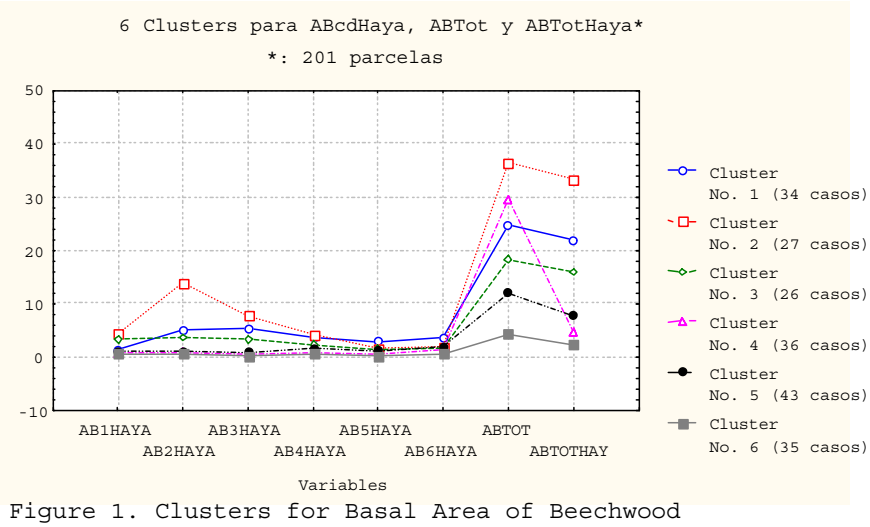



Figure 1. Clusters for Basal Area of Beechwood

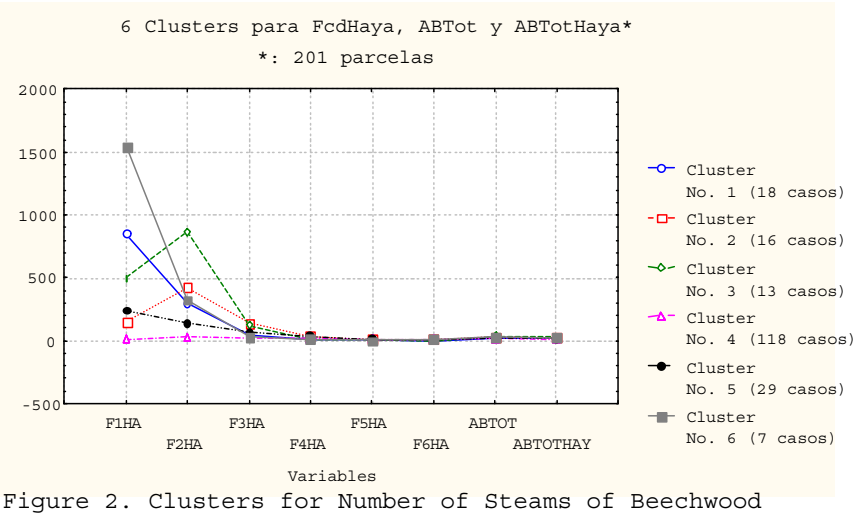



Figure 2. Clusters for Number of Steams of Beechwood

# A Stratified Non-parametric Survival Model for Describing the Distribution of Duration of Studies in Various University Departaments

Aglagia G. Kalamatianou[1] and Sally McClean[2]

[1] Department of Sociology, Panteion University of Athens, Leof.Syngrou 136, Athens 176 71, Greece. E-mail: akalam@panteion.gr
[2] School of Information and Software Engineering, University of Ulster, Coleraine, Northern Ireland. BT52 1SA, UK. E-mail: SI.McClean@ulst.ac.uk

**Abstract:** A stratified non-parametric survival model has been developed and applied to lifetime type data on the duration of studies of students of several university departments called strata. The model incorporates the graduation process at the various strata which is characterised by a threshold for failure and substantial number of cases where failure never occurs. It emerges that the less clearly defined or academically conventional the subject of studies the longer the duration of studies.

**Keywords:** Stratified non-parametric survival models, Survival analysis, Right censoring, Kaplan-Meier estimates, Duration of studies.

## 1 Introduction

It is well known that survival analysis has been used for modelling duration data in many different areas of application such as those of engineering, biomedical and social sciences; see for example Kalbfleisch and Prentice (1980) Cox and Oakes (1984), Blossfeld et al. (1989), Leemis (1995) Klein and Moeschberger (1997). However there are limited references concerning applications to education, particularly with respect to duration of studies. A relevant example was provided recently by Booth and Satchell (1995) and Kalamatianou (1999).

There are universities or other educational institutions where there is no prescribed interval of time within which studies must be completed otherwise students would be dismissed. In other words there is a minimum time for graduation, before that graduation is not possible but students who do not graduate at this time can remain at the university for unlimited time. Greek universities and tertiary educational institutions in general exemplify this phenomenon. In this country it is often observed that a high proportion of students have duration of studies which is longer than the minimum required time for graduation and indeed large numbers of students seem never graduate. This is reflected in a distribution of duration of

studies with a long right tail. An interpretation of this fact can be provided with respect the state of the job market in Greece concerning graduates of different curricula. In addition there seems to be a similar problem in the educational system of Finland and South Africa.
In this paper we develop a stratified non-parametric survival model to describe and compare the distributions of students' stay till graduation in various university departments, by extending the methodology first introduced in Kalamatianou and McClean (1999). The motivation for this paper has been provided by looking at data concerning the duration of undergraduate studies in various departments of a Greek university where the model has been applied. However, our methodology may be applied to any other educational institutions or other types of lifetime data in general with similar characteristics. The main features characterising our analysis are: (1) there are strata that in our application correspond to departments, where the process of duration of stay is similar (2) the data are right censored, (3) there is a threshold for failure in each strata.
The remainder of the paper is set out as follows: In section 2 we give a brief description of the data set. In section 3 we describe the stratified non-parametric model and the procedure for estimation of the distribution of the duration of studies at the various departments and their confidence intervals. The application of the model to the data and the results are given in Section 4. Section 5 is devoted to some conclusions.

## 2 The data

Data were obtained from the official student records of the various departments of an Athenian university which is social and political sciences oriented. The cases examined correspond to 10,313 students who entered all the university departments during ten consecutive academic years starting from 1983-84 and terminating in 1992-93. The follow-up period lasts until the end of February 1997. At the end of the follow up period not all students have obtained a degree; therefore we have right censored data. For each case (student) data are provided concerning students' characteristics. In this paper we use data corresponding to the variables "duration of studies", completed or censored, and "department of enrolment". However a comprehensive description of the graduation process from the Greek university institutions and of the context in which the data set have arisen is given in Kalamatianou and McClean (1999).

## 3 The model

Let $T$, where $T \geq 0$ be a random variable, representing duration of studies, and $t$ any particular value of $T$. A dummy variable $\delta_i$ has been used for indicating censoring so that $\delta_i$ equals 1 if the $i$th observation is complete

and 0 if it is censored. Let $g$, $(g = 1, 2, \ldots, G)$, denote strata, in our case university departments and $S_g(t)$ stand for the survival function that measures the probability that individual $i$ of stratum $g$ , survives longer than $t$. Then:

$$S_g(t) = \begin{cases} 1 & t < c_g \\ a_g + (b_g - a_g)S_{0g}(t - c_g) & t \geq c_g \end{cases} \tag{1}$$

where $0 < a_g < b_g < 1$. $c_g$ is a location parameter indicating the minimum lifetime at stratum $g$. In our application $c_g$ corresponds to the minimum required time for graduation from the department $g$. The parameter $a_g$ accounts for individuals in stratum $g$ that will never fail; in our application this parameter describes students at $g$ department that will never graduate. The parameter $b_g$ is used to account for an individual's lifetime in stratum $g$ who fails immediately after the required minimum time $c_g$. The fraction $1 - b_g$ is the proportion of individuals in stratum $g$ failing at $t = c_g$. In our case it is the proportion of students of the department $g$ who graduate at the first possible opportunity. The survivor function $S_{0g}(t)$ then corresponds to the lifetime distribution of those individuals in stratum $g$ that do not fail immediately and are not of the type that never fail. We note the distribution is degenerate since

$$\lim_{t \to \infty} S_g(t) = a_g \qquad \text{for each } g = 1, 2, \ldots, G \tag{2}$$

The *non-parametric estimation*: Let $n$ be the number of individuals under study and $n_g$ $(g = 1, 2, \ldots, G)$ be the number of individuals in stratum $g$ under study. Then $n = \sum_{g=1}^{G} n_g$ and $t_1, t_2, \ldots, t_k$ are the $k$ distinct observed survival or failure times of the $n_g$ individuals. They are also assumed to be in ascending order i.e. $t_1 < t_2 < \cdots < t_k$. Note, that for simplicity but without lose of generality, we can assume these distinct times are common for all strata. Let $d_{jg}$ denote the number of observed failures in stratum $g$ and at $t_j$, $j = 1, 2, \ldots, k$; $n_{jg}$ the number of individuals of stratum $g$ be at risk just before time $t_j$ $(j = 1, 2, \ldots, k)$ where any values that are censored at $t_j$ are also included in this count. $R_g(t_j)$ is the set of indexes of individuals of stratum $g$ that are at risk just before time $t_j$, $j = 1, 2, \ldots, k$. For a discrete distribution and stratum $g$, let $h_g(t_j)$ be the hazard function where $h_g(t_j) = P(T = t_j / T \geq t_j)$. Extending the results given by Leemis (1995), the survivor function can be written as

$$S_g(t) = \prod_{j \in R_g(t)'} [1 - h_g(t_j)] \quad t \geq 0, \qquad g = 1, 2, \ldots, G \tag{3}$$

where $R_g(t)'$ is the complement of the risk set at time $t$ for stratum $g$. Thus a reasonable estimator for $S_g(t)$ is $\prod_{j \in R_g(t)'} \left[1 - \hat{h}_g(t_j)\right]$ which reduces the

problem of estimating the survivor function to that of estimating the hazard function at each mass value. A corresponding element for the likelihood function at mass value $t_j$ is

$$h_g(t_j)^{d_{jg}}\left[1-h_g(t_j)\right]^{n_{jg}-d_{jg}} \qquad \text{for } j=1,2,\ldots,k \quad \text{and } g=1,2,\ldots,G \tag{4}$$

Note that for each $g=1,2,\ldots,G$, $d_{jg}$ is the number of failures at time $t_j$, $h_g(t_j)$ is the conditional probability of failure at time $t_j$, $n_{jg}-d_{jg}$ is the number of individuals at risk not failing at time $t_j$, and $1-h_g(t_j)$ is the probability of failing after time $t_j$ conditioned on survival to time $t_j$. Thus the likelihood function for $h_g(t_1), h_g(t_2), \ldots, h_g(t_k)$ and each $g=1,2,\ldots,G$, is

$$L[h_g(t_1), h_g(t_2), \ldots, h_g(t_k)] = \prod_{j=1}^{k} h_g(t_j)^{d_{jd}}\left[1-h_g(t_j)\right]^{n_{jg}-d_{jd}} \tag{5}$$

and the log likelihood function is

$$logL[h_g(t_1), h_g(t_2), \ldots, h_g(t_k)] = \sum_{j=1}^{k}\left[d_{jg}logh_g(t_j) + (n_{jg}-d_{jg})log\{1-h_g(t_{jg})\}\right] \tag{6}$$

The $i$th element of the score vector is

$$\frac{\partial logL[h_g(t_1), h_g(t_2), \ldots, h_g(t_k)]}{\partial h_g(t_i)} = \frac{d_{ig}}{h_g(t_{ig})} - \frac{n_{ig}-d_{ig}}{1-h_g(t_i)} \tag{7}$$

for each $g=1,2,\ldots,G$ and $i=1,2,\ldots,k$. Equating this vector to zero and solving for $h_g(t_i)$ yields the maximum likelihood estimate

$$\hat{h}_g(t_i) = \frac{d_{ig}}{n_{ig}} \tag{8}$$

This estimate for $\hat{h}_g(t_i)$ is intuitive, since for each strata $g$, $d_{ig}$ on $n_{ig}$ individuals at risk at time $t_i$ fail, so the ratio $d_{ig}$ to $n_{ig}$ is an appropriate estimate of the conditional probability of failure at time $t_i$. Using this particular estimate for the hazard function at $t_i$ the survivor function estimate becomes

$$\hat{S}_g(t) = \prod_{j\in R_g(t)'}\left[1-\hat{h}_g(t_j)\right] = \prod_{j\in R_g(t)'}\left[1-\frac{d_{jg}}{n_{jg}}\right] \qquad \text{for } g=1,2,\ldots,G \tag{9}$$

commonly known as the Kaplan-Meier or product limit estimate. Greenwood's (1226) formula can be used to give an estimator for the variance of the survivor function for each stratum $g = 1, 2, \ldots, G$, that is

$$\hat{V}[\hat{S}_g(t)] = [\hat{S}_g(t)]^2 \sum_{j \in R_g(t)'} \frac{d_{jg}}{n_{jg}(n_{jg} - d_{jd})} \tag{10}$$

Equation (10) can be used to find asymptotic confidence intervals for $S_g(t)$ $(g = 1, 2, \ldots, G)$, by using the normal critical values:

$$\hat{S}_g(t) \pm z_{a/2}\sqrt{\hat{V}[\hat{S}_g(t)]} \tag{11}$$

The maximum likelihood estimator for the parameter $c_g$, $g = 1, 2, \ldots, G$, in (1) is given by:

$$\hat{c}_g = min\{t_i : i = 1, 2, \ldots, k\} \tag{12}$$

The probability that an individual will never graduate is given by $a_g = S_g(t_{max} + \epsilon)$ where $t_{max}$ is the maximum observed duration of studies, supposed common for all departments. As stated by Gribbin & McClean (1990) the maximum likelihood estimate of $t_{max}$ is $t_k$ so, using the invariance property of likelihood estimators the maximum likelihood estimate of $a_g$ is given by

$$\hat{a}_g = \hat{S}_g(t_k) \qquad g = 1, 2, \ldots, G \tag{13}$$

The estimator of the survival function (1) then becomes

$$\hat{S}_g(t) = \begin{cases} 1 & t < \hat{c}_g \\ \hat{S}_g(t_k) + \left\{\hat{S}_g(\hat{c}_g) - \hat{S}_g(t_k)\right\} \prod_{j \in R_g(t)'} \left[1 - \frac{d_{jg}}{n_{jg}}\right] & t \geq \hat{c}_g \end{cases} \tag{14}$$

## 4 Application and Results

Model (14) was used for estimating the distribution of the duration of studies of the various departments of the previous mentioned Athenian university. There were seven strata, in our application departments ($G = 7$); department of Politics, Public Administration, Sociology, Urban & Rural Planning, Mass Media, Social Policy & Social Anthropology and department of Psychology. Historically, the first three departments are the most old ones operating in the present situation since 1983. The department

of Urban & Rural Planning was created by partition of the department of Public Administration on 1990. The last three departments are rather new ones. In particular the department of Mass Media, and Social Policy & Social Anthropology are in operation since 1990 and the department of Psychology since 1992. However we can order these departments with respect to the extend the corresponding subject of studies or curriculum is clearly defined or academically conventional. In this respect we can say that the departments of Sociology, Mass Media, Social Policy & Social Anthropology and the department of Psychology have the most clearly defined subject of studies. It follows the department of Public Administration, then the department of Politics and last comes the department of Urban & Rural Planning. We shall see that this ordering guides us providing some explanation for the differences appearing in the distribution of the duration of studies among departments.

For all strata the parameters $c_g$ were estimated using equation (12). It came out that $\hat{c}_g = 46$ months, for all $g = 1, 2, \ldots, 7$ and thus $\hat{S}_g(t) = 1$ for $t < 46$. For $t \geq 46$ the estimates of the survival functions, their standard error, and the 95% normal critical values were found using equations (14), (10) and (11) correspondingly. However, first examination of figures and plots concerning differences in survival functions among departments support the hypothesis that duration of studies does not differ among the departments of Sociology, Mass Media, Social Policy & Social Anthropology and Psychology. This hypothesis have been further formally tested using pairwise Log Rank, Breslow and Tarone-Ware statistics. It emerged that all values of test statistics are not significant at an acceptable level. Thus it is sensible for the shake of brevity in describing the patterns of the distribution of duration of studies of the departments, to pool data over these last four departments. For conciseness we shall call this pooling group of departments "the Combined Department" (CD). The new four departments appear to be significantly different concerning the distribution of duration of studies. In the sequel we describe the results came out from the analysis. In Table 1 results are reported concerning the Kaplan-Meier estimates for the survival functions and their 95% normal critical values for the CD and for the departments of, Public Administration, Politics and Urban & Rural Planning, using equations (14) and (11) respectively. Note that for this application $G = 4$ and for the four departments just mentioned we use correspondingly $g = 1, 2, 3, 4$. Figure 1 illustrates the path of the survival functions with their corresponding confidence bands.

Mindful examination of figures and plots reveal the followings: It is easy to understand that survival functions of the four departments under study exhibit almost parallel curves all the way through the follow up period, and there is clear distance among them. These reflect similarities and differences in the graduation patterns of the departments. All curves decrease steeply at the beginning of the follow up period and they then level off. This means that for all the departments and for a period of time that follows

TABLE 1. Kaplan-Meier estimates for survival functions and confidence limits

| Time $t$ in | Sociology+Mass Media+Social Policy & Anthropology Psychology | | | Public Administration | | | Politics | | | Urban & Rural Planning | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| months | $\hat{S}_t$ | Lower Limit | Upper Limit | $\hat{S}_t$ | Lower Limit | Upper Limit | $\hat{S}_t$ | Lower Limit | Upper Limit | $\hat{S}_t$ | Lower Limit | Upper Limit |
| 46 | 0.8621 | 0.8511 | 0.8731 | 0.9009 | 0.8889 | 0.9129 | 0.9363 | 0.9277 | 0.9449 | 0.9573 | 0.945 | 0.9696 |
| 50 | 0.6197 | 0.6044 | 0.6350 | 0.7152 | 0.6972 | 0.7332 | 0.7818 | 0.7671 | 0.7965 | 0.8350 | 0.8123 | 0.8577 |
| 54 | 0.4979 | 0.4820 | 0.5138 | 0.6422 | 0.6232 | 0.6612 | 0.7221 | 0.7060 | 0.7382 | 0.7748 | 0.7493 | 0.8003 |
| 58 | 0.4474 | 0.4315 | 0.4633 | 0.5655 | 0.5455 | 0.5855 | 0.6512 | 0.6340 | 0.6684 | 0.7272 | 0.6996 | 0.7548 |
| 62 | 0.3557 | 0.3402 | 0.3712 | 0.4820 | 0.4616 | 0.5024 | 0.5657 | 0.5477 | 0.5837 | 0.6630 | 0.6332 | 0.6928 |
| 66 | 0.3227 | 0.3074 | 0.3380 | 0.4463 | 0.4259 | 0.4667 | 0.5348 | 0.5166 | 0.5530 | 0.6353 | 0.6049 | 0.6657 |
| 70 | 0.3061 | 0.2910 | 0.3212 | 0.4137 | 0.3935 | 0.4339 | 0.5129 | 0.4945 | 0.5313 | 0.6165 | 0.5855 | 0.6475 |
| 74 | 0.2732 | 0.2585 | 0.2879 | 0.3762 | 0.3560 | 0.3964 | 0.4751 | 0.4565 | 0.4937 | 0.5876 | 0.5558 | 0.6194 |
| 78 | 0.2608 | 0.2461 | 0.2755 | 0.3550 | 0.3350 | 0.3750 | 0.4576 | 0.4390 | 0.4762 | 0.5750 | 0.5431 | 0.6069 |
| 82 | 0.2550 | 0.2405 | 0.2695 | 0.3376 | 0.3178 | 0.3574 | 0.4455 | 0.4269 | 0.4641 | 0.5600 | 0.5277 | 0.5923 |
| 86 | 0.2344 | 0.2201 | 0.2487 | 0.3110 | 0.2914 | 0.3306 | 0.4236 | 0.4050 | 0.4422 | 0.5490 | 0.5165 | 0.5815 |
| 90 | 0.2278 | 0.2135 | 0.2421 | 0.2936 | 0.2742 | 0.3130 | 0.4098 | 0.3912 | 0.4284 | 0.5408 | 0.5081 | 0.5735 |
| 94 | 0.2234 | 0.2091 | 0.2377 | 0.2889 | 0.2695 | 0.3083 | 0.4037 | 0.3851 | 0.4223 | 0.5348 | 0.5019 | 0.5677 |
| 98 | 0.2158 | 0.2017 | 0.2299 | 0.2771 | 0.2577 | 0.2965 | 0.3916 | 0.3728 | 0.4104 | 0.5212 | 0.4879 | 0.5545 |
| 102 | 0.2106 | 0.1965 | 0.2247 | 0.2685 | 0.2491 | 0.2879 | 0.3875 | 0.3687 | 0.4063 | 0.5167 | 0.4834 | 0.5500 |
| 106 | 0.2083 | 0.1942 | 0.2224 | 0.2617 | 0.2423 | 0.2811 | 0.3826 | 0.3638 | 0.4014 | 0.5149 | 0.4816 | 0.5482 |
| 110 | 0.2023 | 0.1882 | 0.2164 | 0.2566 | 0.2372 | 0.2760 | 0.3745 | 0.3555 | 0.3935 | 0.5114 | 0.4779 | 0.5449 |
| 114 | 0.1991 | 0.1850 | 0.2132 | 0.2532 | 0.2336 | 0.2728 | 0.3720 | 0.3530 | 0.3910 | - | - | - |
| 118 | 0.1969 | 0.1828 | 0.2110 | 0.2485 | 0.2289 | 0.2681 | 0.3693 | 0.3503 | 0.3883 | - | - | - |
| 122 | 0.1952 | 0.1811 | 0.2093 | 0.2438 | 0.2240 | 0.2636 | 0.3610 | 0.3416 | 0.3804 | 0.5023 | 0.4682 | 0.5364 |
| 126 | 0.1935 | 0.1794 | 0.2076 | 0.2426 | 0.2228 | 0.2624 | 0.3555 | 0.3359 | 0.3751 | - | - | - |
| 130 | 0.1928 | 0.1787 | 0.2069 | 0.2408 | 0.2208 | 0.2608 | 0.3539 | 0.3341 | 0.3737 | - | - | - |
| 134 | 0.1869 | 0.1726 | 0.2012 | 0.2355 | 0.2149 | 0.2561 | 0.3507 | 0.3307 | 0.3707 | - | - | - |
| 138 | 0.1861 | 0.1718 | 0.2004 | 0.2319 | 0.2111 | 0.2527 | 0.3492 | 0.3290 | 0.3694 | - | - | - |
| 142 | 0.1849 | 0.1706 | 0.1992 | - | - | - | - | - | - | - | - | - |
| 146 | 0.1825 | 0.1680 | 0.1970 | 0.2276 | 0.2056 | 0.2496 | 0.3449 | 0.3233 | 0.3665 | - | - | - |
| 150 | - | - | - | - | - | - | - | - | - | - | - | - |
| 154 | - | - | - | - | - | - | - | - | - | - | - | - |
| 158 | 0.1808 | 0.1659 | 0.1957 | - | - | - | - | - | - | - | - | - |

the threshold for graduation there is a higher but also decreasing rate for graduation than later. Towards the end of the follow up period all curves level off which means that for all the departments there is a percentage of students that effectively never graduate. However departments differ both in the rate of graduation at the beginning of the process and in the percentage of students (estimate of the parameter $a_g$) who never graduate. It is evident that survival function of students of the CD falls beneath those of the rest three departments, for the whole period of study. This means that students of this combined department graduate faster than students of the other three departments, do. We can note that graduation becomes slower and slower getting from $g = 1$, which is the CD, to $g = 4$ that corresponds to the department of Urban & Rural Planning, indicating that a "general principle" has been followed which is "the less clearly defined or academically conventional the subject of studies the longer the duration of studies". We shall shortly see that all statistics come out from the analysis are in favour of this general principle.

It is estimated that graduation just after the threshold or at the first possible opportunity emerges to the almost 14% (that corresponds to $\hat{b}_1 = 0.8621$) of the student population of the CD. The corresponding percentages for the departments of Public Administration, Politics and Urban & Rural Planning are estimated to about 10% ($\hat{b}_2 = 0.9009$), 6% ($\hat{b}_3 = 0.9363$) and 4% ($\hat{b}_4 = 0.9573$), that argue for the decreasing rate of graduation getting from $g = 1$ to $g = 4$. We get to the same conclusion taking into account the estimates of the percentiles and mean values. In particular we have the followings. The median duration of studies at the CD ($g = 1$) is estimated to be 54 months with a rather narrow 95% confidence interval under the normal approximation between 53 and 55 months. This verifies that about half of the student population of this department graduates during the first three possible opportunities. The corresponding median values for $g = 2, 3$ (departments of Public Administration and Politics) are estimated to be 62, 74 with confidence intervals between 61 and 63, and, 71 and 77. These figures show that about half of the students at the department of Public Administration graduate during the first five opportunities whereas those of Politics during the first eight. The median duration value corresponding to the department of Urban & Rural Planning ($g = 4$) can not be formally estimated since it falls out of the range of the observed values. However it can be foreseen (Table 1) it is closed to 122 months, pointing that half of this student population graduates during the first 20 opportunities. Based on the median values it is also evident that duration of studies is getting longer going from $g = 1$ to $g = 4$. We come to same point taking into account the mean values of the distributions of the duration of studies. It is estimated that the mean value for $g = 1$, is 78 months with confidence interval between 77 and 79. The corresponding estimated values for $g = 2, 3, 4$ are 86 (confidence interval 84 - 88), 98 (97 -100) and 113 (109 - 116). These mean values although underestimated, due to the fact the

TABLE 2. Pairwise comparison of survival functions

| Pair of strata $g$ | Long rank statistics | Breslow statistics | Tarone-Ware Statistics | d.f | significance level |
|---|---|---|---|---|---|
| 1,2 | 81.24 | 68.58 | 75.54 | 1 | 0.00 |
| 1,3 | 345.66 | 373.86 | 375.48 | 1 | 0.00 |
| 1,4 | 48.83 | 36.48 | 42.42 | 1 | 0.00 |
| 2,3 | 71.36 | 99.79 | 92.83 | 1 | 0.00 |
| 2,4 | 168.03 | 129.74 | 149.10 | 1 | 0.00 |
| 3,4 | 353.21 | 318.82 | 345.83 | 1 | 0.00 |

last observations are censored, they suggest that on average and for all departments graduation occurs considerable time after the threshold and the mean time for graduation is getting greater proceeding from stratum $g = 1$ to $g = 4$. In particular it is estimated that on average at the CD students graduate 32 months after the minimum time for graduation or at the ninth opportunity for graduation. At the department of Public Administration that average corresponds to 40 months or to the eleventh opportunity, at the department of Politics to 52 months or to the fourteenth opportunity and at the department of Urban & Rural Planning to 67 months or to about eighteenth opportunity for graduation.

Finally it is estimated that for the CD, $\hat{a}_1 = 0.1808 - \epsilon$, which means that somehow less than 18.08% of the corresponding student population never graduate. For the rest of the departments the corresponding percentages are estimated to higher values following the general principle mentioned earlier. Precisely at the department of Public Administration graduation do not chances to about 22.76% ($\hat{a}_2 = 0.2276 - \epsilon$) of the students, at the department of Politics to about 34.49% ($\hat{a}_3 = 0.3449 - \epsilon$) and at the department of Urban & Rural Planning to about half of the students' population, 50.23%, ($\hat{a}_4 = 0.5023 - \epsilon$).

We further test formally the null hypothesis that survival functions are the same for all strata using Log rank, Breslow and Tarone-Ware statistics. All theses statistics are asymptotically $\chi^2$ distributed with $\nu - 1$ degrees of freedom ($\nu$ is the number of strata) presupposing survival curves do not intersect and strata exhibit similar censoring patterns. Both assumptions are satisfied in our application. We first test the above mentioned hypothesis against the alternative that at least one pair of the survival functions are different. In this case all statistics turned to be significant under the significance level of 0.0000. In particular Log rank, Breslow and Tarone-Ware statistics take correspondingly the values of 575.82, 577.47, 595.17 (with 3 d.f.). We further compared all distinct pairs of survival functions. Results are reported to Table 2.

From Table 2 it is clear that survival functions for the distributions of

duration of studies differ significantly between all pairs of the four strata or departments under study.
It is clear that all results line up to the general principle mentioned above. Various explanations could be given to this finding. However one could be given with respect to the job settlement of the corresponding graduates. In particular it is sensible to suppose that the less clearly defined the subject of studies the more vague the situation concerning job settlement. So the less possibility of getting a job the less motive of getting degree.

## 5 Conclusion

We have provided a stratified non-parametric survival model for the description of the distribution of the duration of studies in various university departments. The model has been applied to a right censored data set concerning the duration of undergraduate studies in several university departments. The results give a clear picture of patterns of the distribution of the duration of studies of the various departments and reveal similarities and differences among them.

### References

Blossfeld, H-P., Hamerle A. and Mayer K. U. (1989). *Event History Analysis. Statistical Theory and Application in the Social Sciences.* Lawrence Erlbaum Associates, Inc. Hillsdale, New Jersey.

Booth, L. A. and Satchell S. E. (1995). The hazards of doing PhD: an analysis of competion and withdrawal rates of British PhD students in the 1980s. *J. R. Statist. Soc. A.* **158**, part2, 297-318.

Cox, R. D. and Oakes D. (1984). *Analysis of Survival Data.* Chapman & Hall.

Greenwood, M. (1926). *The natural duration of cancer. Reports on Public health and Medical Subjects.* Her Majesty's stationery office. London. **Vol. 33.** 1-26.

Gribbin, J. O. and McClean S. I. (1990). Modelling the duration of spells of withdrawal from a female dominated labour force. *Journal of Applied Statistics*, **Vol 17**, 369-381.

Kalamatianou, G. A. (1999). Using Life Table method on a Censored Data Set for Estimating Duration of Studies in University Departments. In *Bulletin of the International Statistical Institute*, 52nd Session. Contributed papers, Tome LVIII, Book2, pp. 101-102.

Kalamatianou, G. A. and S. McClean (1999). Survival analysis for modeling duration of undergraduate studies: The Case of Greece. In preparation

Kalbfleisch, J. D. and Prentice. L. R. (1980). *The Statistical Analysis of Failure Time Data.* John Wiley & Sons Inc. New York.

Klein, J. P. and Moeschberger M. (1997). *Survival Analysis.* New York: Springer Verlang.

Leemis, M. L. (1995). *Reliability Probabilistic Models and Statistical Methods.* Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

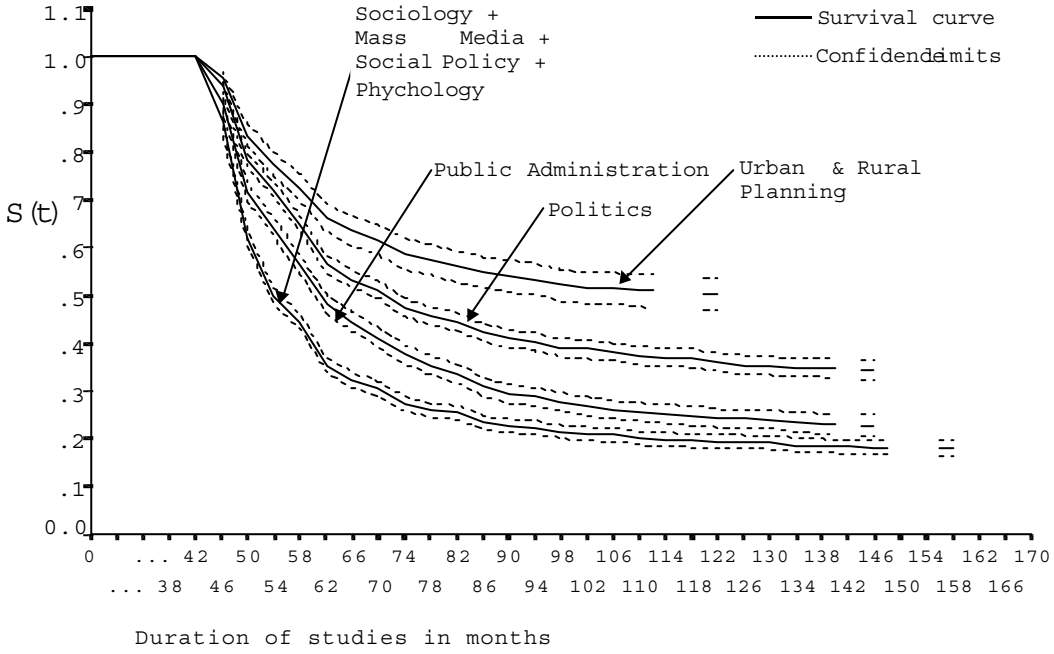



FIGURE 1. Survival functions and confidence limits

# Manuscripts in German Monasteries

Göran Kauermann[1], Ludwig Heigenhauser[1] and William Whobrey[2]

[1] Ludwig-Maximilians-Universität München, Institut für Statistik, Akademiestrasse 1, 80799 München, Germany
[2] Yale University, Department of Germanic Languages and Literatures, P.O. Box 208210, New Haven, CT 06520-8210, USA

**Abstract:** We analysis the number of manuscripts in German speaking monasteries during the middle ages. The major focus is on finding clusters of monasteries with a comparable production profile. This is done by a random effect model with monastery specific effect.

**Keywords:** German Manuscript, Random Effect Models, Nonparametric Maximum Likelihood, Cluster Analysis

## 1 Introduction

Before the invention of printing by the German Johannes Gutenberg (1400-1468) around 1450, manuscripts were the most common medium of publication. These manuscripts were almost exclusively produced in monasteries and church schools, and they vary greatly in length and quality. The oldest manuscripts produced in Germany date to the 4th century, the last era of real production is the 16th century. After this, printing replaced the time-consuming and cost-intensive handcraft of book copying. Figure 1 shows the number of extant German manuscripts classified by the century of their production. The peak in the 15th century is explained by technical developments in the cheap production of paper, which reduced the costs of handwritten books significantly.

The numbers plotted in Table 4 originate from S. Krämer (1989), who cataloged all manuscripts produced in Germany during the Middle Ages which are still in existence today. For each listed manuscript, the time and place of production is recorded. The classification of a manuscript is a matter for experts, and the assignment by century is often the closest these texts can be dated, especially for the oldest manuscripts. Finally, the year of foundation and closure as well as religious order is given for each scriptorium.

Krämer's costly and comprehensive catalog has been made numerically accessible by the third author. This now enables the application of statistical tools for the investigation of quantitative questions in the analysis of manuscripts. It is important for scholars to know how many manuscripts have been produced in a particular period in a particular scriptorium,

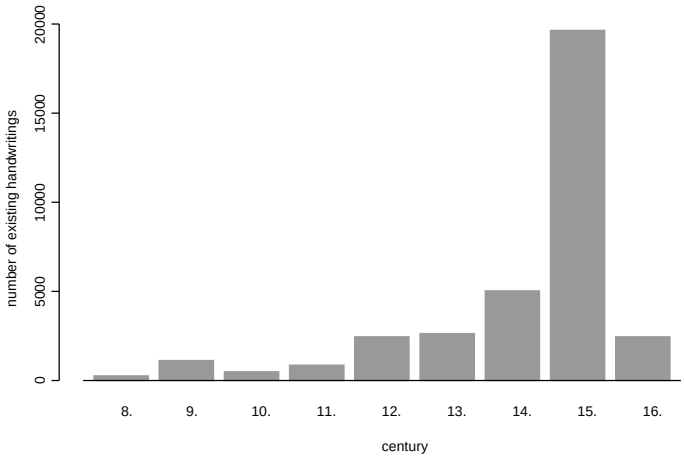



FIGURE 1. Number of known German handwritings by their year of production

since this number might provide evidence for the size, power, and orientation of that monastery. Moreover, it is well-known that a great number of manuscripts have been destroyed or lost over the centuries, but an accurate estimate of the number of lost manuscripts remains elusive.
In this work we make a first attempt in the direction of a quantitative analysis of manuscripts through the use of statistical models. We first concentrate on those production sites active from the 9th - 16 century to reduce the number of considered scriptoria to 81. Then we focus on the problem of finding clusters in the scriptoria, i.e. finding scriptoria which have the same or similar production profiles of manuscripts across eight centuries.

## 2 The Model

An initial investigation of the data uncovers strong over-dispersion when using a Poisson response model, which is not surprising since a number of relevant covariates are not recorded or completely unknown. We therefore fit a random effect model to compensate for the over-dispersion (see e.g. Aitkin, 1996). Moreover, the random effect also models the longitudinal dependence structure found in the data. This means we consider the model

$$E(y_{ij}|x_{ij}, \mathbf{z}_i) \quad = \quad \exp(x_{ij}\beta + \mathbf{z}_i) \tag{1}$$

where $y_{ij}$ gives the number of manuscripts found for the $i$-th monastery in the $j$-th century, $x_{ij}$ are additional covariates like the century and the

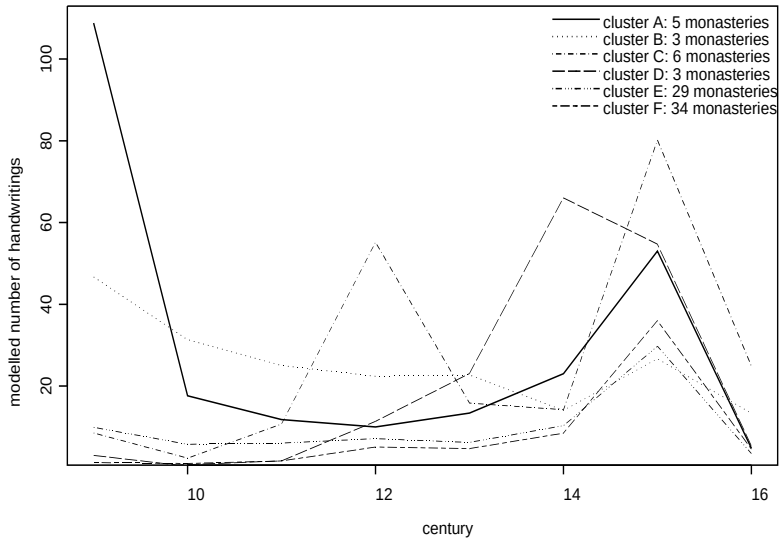



FIGURE 2. Production profile for 6 clusters in 81 Monateries

location of the monastery and $\mathbf{z}_i$ is the unobserved monastery effect which serves as random intercept in the model. Assuming normality for $\mathbf{z}_i$ seems doubtful and we therefore leave the distribution of $\mathbf{z}_i$ unspecified and make use of the Nonparametric Likelihood Estimation approach (NPML), see e.g. Aitkin (1999).

It is obvious that model (1) is unlikely to hold for all monasteries and the intention is to find clusters of monasteries where (1) holds. This is done by the following iterative procedure. The basic idea is, that we assume model (1) to hold cluster-wise, i.e. the coefficients $\beta$ differ between clusters. If we fit model (1) from all data, i.e. with coefficient $\beta$ constant for all clusters, the resulting residuals show the unexplained structure not covered by a global model. We then cluster the residuals and refit model (1) for each cluster found. This provides cluster specific coefficients $\beta$. These two steps, namely clustering fitted residuals and refitting the models separately for the found clusters, are pursued iteratively until homogeneous clusters are found and the residuals in each cluster do not show (significant) structure. Figure 2 shows the production profiles for the 6 resulting clusters. Though clusters E and F appear similar, the remaining 4 clusters distinguish by particular items. Figure 3 shows the local distribution of the clusters in Germany. While clusters A and B do show local structure, the cloisters

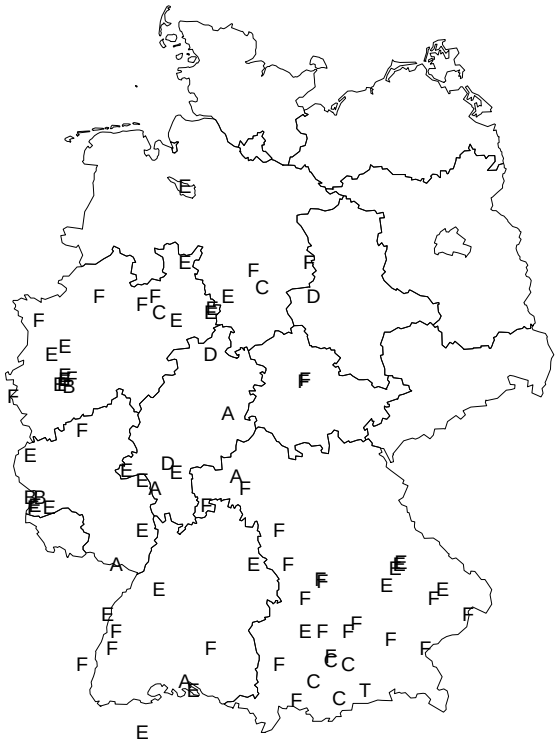


FIGURE 3. Location of Clusters of Monasteries

from cluster C clearly build a local cluster in South-Bavaria.
The modeling is extended to smaller monasteries which are not included in the first selected 81 and further investigation on a substance matter level explains the findings.

## References

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing*, **6**, 251-262.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 218-234.

Krämer, S. (1989). *Handschriftenerbe des deutschen Mittelalters, Mittelalterliche Bibliothekskataloge Deutschlands und der Schweiz* (3 volumes). München: C.H. Beck.

# Modelling Data Down 9 Kilometers into the Earth's Crust

Helmut Küchenhoff[1], Susanna Adelhardt[1] , Brian Marx[2], Helmuth Winter[3]

[1] Institut für Statistik, Ludwig–Maximilians–Universität München, Akademiestr. 1, D-80799 München, helmut@stat.uni-muenchen.de
[2] Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803-5606 USA, brian@stat.lsu.edu
[3] Institut für allgemeine und angewandte Geophysik, Ludwig–Maximilians–Universität München, winter@geophysik.uni-muenchen.de

**Keywords:** Regression; P-Splines; Geophysical data

## 1 Introduction

We present an analysis of depth-related geophysical and geochemical data from the German Continental Drilling Project (KTB) by regression modelling with varying coefficients. Especially the varying coefficients models are new procedures in analyzing and modelling borehole data. We expect a detailed understanding of the depth-changing processes which cannot be found by commonly used methods.
The KTB was designed to study the properties and processes of the continental crust by means of a deep borehole. Two boreholes were drilled into the crystalline crust of the western margin of the Bohemian Massif, Bavaria, SE Germany (Emmermann and Lauterjung, 1997; Yardley, 1997). They reached a final depth of 4 km (KTB Vorbohrung, KTB VB) and 9.1 km (KTB Hauptbohrung, KTB HB), respectively. Extensive research was done over a wide range of geoscientific questions such as the nature of geophysical structures, evolution and structure of the European Variscan basement, and more.
The investigations of drill cuttings comprise 68 variables at 5922 depth points (1 to 10 m spacing) down to the final depth of 9.1 km. Some specific characteristics arise from the borehole data. The drilled geological section is a sequence of different metamorphic rocks, mainly gneiss and metabasites, succeeding each other. This gives the data a characteristic structure like time or spatial series with autocorrelation of the data. In Section 2 we give methods and results of the regression analysis in two depth zones. In Section 3 the analysis with varying coefficients of the whole data using P–splines is presented.

## 2 Multiple Regression analysis

We investigate the characteristic properties of cataclastic fault zones on the data set of drill cuttings by multiple linear regression analysis and use a selection of 26 variables representing petrophysical properties and the chemical composition. The target variable is the logarithmic "amount of cataclastic rocks in the drill cuttings (CATR)". The multiple regression model is:

$$\ln(CATR_i + 1) = \text{LNCATR}_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i \quad (1)$$

$$E(\epsilon_i) = 0 \quad (2)$$

$$COV(\epsilon_i, \epsilon_{i+1}) = \sigma^2 * r^{(d_{i+1} - d_i)} \quad (3)$$

Here, LNCATR is the response variable, $x_{ij}$ are the $p$ regressors and $d$ is the depth variable. The model can be derived from a regression with an AR(1) error process, taking the non-equidistance of the data into account. The model is fitted by a two step procedure, which is a generalization of the Cochran Orcutt-Procedure, see e. g. Hamilton (1994, ch. 8). The variable selection process gives us different results for the two investigated sections, 1738 - 2380 m (cataclastic zone in gneiss) and 4524 - 4908 m (cataclastic zone in metabasit), see Table 1 and Table 2 below. The dependence of the results on Lithology is confirmed by geoscientific investigations, see Zulauf et al. (1999).

Table 1: *Results for the gneiss-zone from multiple regression with autocorrelated errors., Here $C$ is the carbon content, $H_2O$ the crystal water content and $\lambda$ the thermal conductivity.*

| Variable | Estimate | STD | t-value | p-value |
|---|---|---|---|---|
| INTERC | -5.816083 | 2.080966 | -2.794896 | 0.005 |
| $C$ | 3.079116 | 0.540444 | 5.697380 | 0.000 |
| $H_2O$ | 1.138937 | 0.394953 | 2.883723 | 0.004 |
| $\lambda$ | 1.249037 | 0.577929 | 2.161229 | 0.031 |

Table 2: *Results for the Metabasit-Zone., Here d is the depth, $Na_2O$ is the Natriumdioxide content, $Al_2O_3$ is the Aluminiumoxide content.*

| Variable | Estimate | STD | t-value | p-value |
|---|---|---|---|---|
| INTERC | -13.449515 | 3.946770 | -3.407727 | 0.001 |
| $d$ | 0.003588 | 0.000959 | 3.743608 | 0.000 |
| $Na_2O$ | 0.330501 | 0.081384 | 4.061003 | 0.000 |
| $H_2O$ | 1.210026 | 0.276006 | 4.384052 | 0.000 |
| $Al_2O_3$ | -0.275209 | 0.072504 | -3.795757 | 0.000 |

The estimated autocorrelation are very high in both models. The results differed substantially when the depth structure is not into account. Many variables which were not significant in the above model would have significant if the error structure was ignored.

## 3 Regression models with depth- varying coefficients

In the previous section, the estimated regression coefficients have single values. However, we might expect that since the observations are ordered with depth of the borehole, then these coefficients should vary smoothly (interact) along the depth index. Hastie and Tibshirani (1993) developed varying coefficient models. We construct a generalized additive model that implements smoothness and varying coefficients using penalized B-spline techniques or P-splines (Marx and Eilers, 1998). A nice feature of this approach is that smoothness is coerced into linear structures. Consider rewriting (1) as

$$E[\ln(\text{CARTR})]_i = f(d_i) + \sum_{j=1}^{p} x_{ij}\beta_j(d_i), \tag{4}$$

where $p$ is the number of regressors in the model. Notice that in the above expression the value of the coefficient depends on depth $d$. The smooth term $f(d)$ can be thought of as a varying intercept or trend in depth.
Suggested routines, such as backfitting, can be eliminated by projecting $\beta_j(d_i)$ onto a smooth subspace, such as a B-spline basis. Consider the above equation as

$$E[\ln(\text{CARTR})]_i = B\delta_d + \sum_{j=1}^{p} \text{diag}(x_{ij})B_j\gamma_j, \tag{5}$$

where $B$ is the B-spline basis. The B-spline coefficients $\delta$ and $\gamma$ correspond to the the smooth depth term and the $p$ varying coefficient terms respectively. A generous number of equally-spaced knots are used to construct the B-spline bases and a separate difference penalty is attached to the log-likelihood function for $\delta$ and each $\gamma$, see Eilers and Marx (2000) for details. In Figure 1, one example of smooth depth varying coefficients is provided for the regressor coefficient of the amount of crystal water. This figure gives an impact in partial rate of change in ln(CARTR), that varies with depth, crystal Water increases one unit. In that model all other variables from two the models above $(C, \lambda, Na_2O, Al_2O_3)$ were included with smooth depth varying coefficients. Since the data are highly multivariate, the results have to be interpreted only in the sense of exploratory data analysis. In the talk, we present different results and discuss strategies to find an appropriate final candidate model. We also compare our methods to those of local polynomials, see e. g. Kauermann and Tutz (2000).

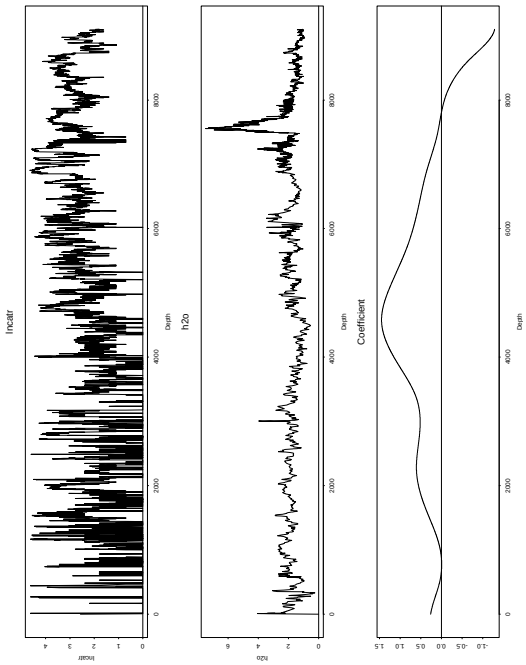



FIGURE 1. *Result of the fit of model (4) for the depth varying coefficient of the amount of crystal water.*

## References

Eilers, P.H.C. and Marx, B.D. (2000). Generalized Linear Additive Smooth Structures (submitted).

Emmermann, R. and Lauterjung, J. (1997),German Continental Drilling Program KTB: Overview and major results. *J. Geophys. Research,* **102** (B8), 18179-18201.

Kauermann, G. and Tutz, G. (2000) Local Likelihood Estimation in Varying-Coefficient Models Including Additive Bias Correction. *Journal of Nonparametric Statistics* , (to appear).

Marx, B. D. and Eilers, P.H.C (1998) Direct Generalized Additive Modelling with penalized likelihood. *Computational Statistics and Data Analysis* **28**, 193-209.

Yardley, B. (1997) Probe of a plate interior. *Nature*, **389**, 792.

Zulauf, G., Palm,S , Petschick, R. and Spies, O. (1999) Element mobility and volumetric strain in brittle and brittle-viscous shear zones of the superdeep well KTB (Germany). *Chemical Geology*, **156**, 135-149.

# HELP for Missing Data

Hung-kung Liu[1], Gene Hwang[2] and Gerard Stenbakken[1]

[1] National Institute of Standards & Technology, 100 Bureau Dr, Stop 8980, Gaithersburg, MD 20899-8980, USA
[2] Cornell University, Ithaca, NY 14853, USA

**Abstract:** Many engineering problems involve high-dimensional observations with mean vectors sitting in a lower dimensional space. Exhaustive measurement of all the elements of an observation is often time consuming and expensive. Applying a traditional multivariate linear model, one can incorporate a small number of the elements of the observation with a known design matrix to predict the rest of the elements. However, for a complicated engineering system, the design matrix is often hard to fully determine. We investigate an empirical linear model, in which we allow ourselves to use the data to determine the size of the design matrix and to estimate the unknown part of the design matrix. This estimated model is then used to construct point and interval estimation for the future observation. This technique is called HELP (High-dimensional Empirical Linear Prediction). In this paper, using the concept of Expectation Maximization (EM), we extended the HELP methodology to a more complex model.

**Keywords:** Empirical modeling; Singular value decomposition; EM algorithm.

## 1 Introduction to HELP

In the industrial manufacturing of a certain complex device, quality assurance for each product requires making a large number, say $m$, of measurements to determine conformance to specifications. For a given product, $y$ denotes its *discrepancy vector*, i.e., the $m-$dimensional vector whose components are the differences between the observed measurements and the targeted specifications. The ideal product with no measurement error therefore corresponds to a discrepancy vector which is the zero vector. Exhaustive checking of all test points for each product is often time consuming and expensive. However, the behavior of a product often depends only on a small number of independent variables. That is, we assume, by using a Taylor expansion, the model

$$y = X\gamma + \varepsilon, \tag{1}$$

where $\gamma$ is the $\ell$ dimensional unknown column vectors, $X$ is the $m \times l$ design matrix, and the measurement error $\varepsilon$ is distributed $N(0, \sigma^2 I)$. Applying a traditional multivariate linear model, one can combine a small subset of

the elements of $y$, denoted by a $t$–dimensional column vector $y_1$, with the known design matrix $X$ to predict the rest of the elements. We shall assume without loss of generality that the observed elements $y_1$ are from the first $t$ coordinates. Hence we may write $y' = (y_1', y_2')$, where $y_2$ consists of the unobserved elements of $y$. We shall partition $X$ as $(X_1', X_2')'$, where $X_1$ consists of the first $t$ rows of $X$. Assume further that there are training data

$$y^i = X\gamma^i + \varepsilon^i, \quad i = 1, \ldots, n, \tag{2}$$

from the same manufacturing process available to estimate $\sigma$. A standard prediction set based on $y_1$ is described by the inequality

$$\frac{1}{m-t}(y_2 - X_2\hat{\gamma}(X_1))'\Sigma^{-1}(X)(y_2 - X_2\hat{\gamma}(X_1))/\hat{\sigma^2}(X) \leq F \tag{3}$$

where $F$ is the $1-\alpha$ quantile of an $F$ distribution with $m-t$ and $n(m-\ell)$ degrees of freedom,

$$\Sigma(X) = I + X_2(X_1'X_1)^{-1}X_2',$$

$$\hat{\sigma^2}(X) = \frac{1}{n(m-\ell)}[\sum_{i=1}^{n} \mid y^i - X\hat{\gamma^i}(X) \mid^2] \tag{4}$$

$$\hat{\gamma^i}(X) = (X'X)^{-1}X'y^i = X'y^i,$$

and

$$\hat{\gamma}(X_1) = (X_1'X_1)^{-1}X_1'y_1.$$

Assuming $\varepsilon$ and $\varepsilon^i$ are i. i. d. $N(0, \sigma^2 I)$, it follows from the classical theory about linear models that the prediction set (3) has coverage probability exactly $1-\alpha$. Also, without normality assumption of $\varepsilon^i$'s, the coverage probability is asymptotically $1-\alpha$. However, for a complicated engineering system, the design matrix is often hard to fully determine. We, therefore, investigate empirical models. That is, we assume that

$$y = \mu + \chi\beta + X\gamma + \varepsilon. \tag{5}$$

Here $\beta$ and $\gamma$ are respectively the $k$ and $\ell$ dimensional unknown column vectors. Also the $m \times l$ matrix $\chi$ represents the part of the design that is unknown, and the $m \times l$ matrix $X$ represents the part of the design that is known. Model (5) is therefore called the *empirical* linear model since part of the design matrix $\chi$ can only be determined empirically. To make $\chi$ and $\beta$ identifiable, we first write $\chi = SDV$ by some specific singular value decomposition. Model (5), after replacing $DV\beta$ by $\eta$, becomes

$$y = \mu + S\eta + X\gamma + \varepsilon. \tag{6}$$

If $S$ were known, one can incorporate $y_1$ with the known design matrix to predict the rest of the elements as in (3). Since $S$ is unknown, it will be estimated by using a set of $y^i$'s, called the training data. We can express the training data from the same manufacturing process as

$$y^i = \mu + S\eta^i + X\gamma^i + \varepsilon^i, \quad i = 1, \ldots, n. \tag{7}$$

To illustrate how to estimate $S$, consider the special case of (6) and (7),

$$y = S\eta + \varepsilon, \text{ and} \tag{8}$$

$$y^i = S\eta^i + \varepsilon^i, \tag{9}$$

i.e., $\mu = 0$ and $l = 0$, the case when the design matrix is totally unknown. We use basically the principal components of $y^i$ to estimate the column space of $S$. Precisely, let $Y = (y^1, \ldots, y^n)$ be the $m \times n$ matrix of the training data and perform a singular value decomposition on $\frac{1}{\sqrt{n}}Y$ so that

$$\frac{1}{\sqrt{n}}Y = (\widehat{S}, \widehat{S}_0)(\widehat{D})\widehat{V} \tag{10}$$

where $(\widehat{S}, \widehat{S}_0)$ and $\widehat{V}$ are orthogonal matrices of sizes $m$ and $n$; $\widehat{S}$ is $m \times k$; and $\widehat{D}$ is an $m \times n$ diagonal matrix whose diagonal elements $\widehat{d}_i$'s, called the singular values are decreasing in $i$. The decomposition is always possible for any matrix. Assuming that $k$ is known, we then estimate $S$ by $\widehat{S}$, which has size $m \times k$ as well, and $\eta^i$ by the estimator

$$\widehat{\eta}^i(\widehat{S}) = (\widehat{S}'\widehat{S})^{-1}\widehat{S}'y^i = \widehat{S}'y^i, \tag{11}$$

where for any matrix $M$, $M'$ denotes the transpose of $M$.
Then the estimated model, together with a $t$–dimensional subvector $y_1$ of $y$, may be used to estimate the rest of the components $y_2$ of $y$.
What are the reasonable point predictors which allow us to do so? How does one assess the prediction error? We tackle the problems (especially the second one) by constructing prediction sets, confidence sets and Scheffé type (1959) simultaneous intervals. This prediction based on the empirical linear model is called the high–dimensional empirical linear prediction (HELP).

## 2 HELP for mission data

As part of the work on developing efficient new testing strategies for software-embedded systems, we have collaborated on tests of an extension of the HELP algorithm to devices following a model outside the usual non-software-embedded framework. For example, we assume that the training data follows the usual model as in (9), but at the future time $j$, observations for the $i$-th product follows

$$y_j^i = S\eta^i + R\delta_j^i + \varepsilon_j^i, \tag{12}$$

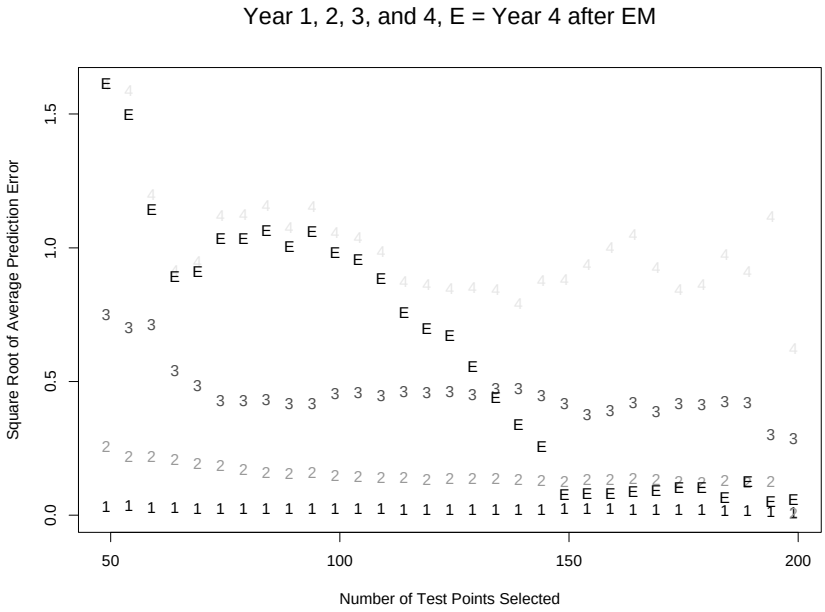



FIGURE 1. The square root of the average predicting error versus the number of test points selected ($t_p$) are plotted for years 1, 2, 3, and 4, and for "E", the EM estimates for year 4. Our analytical result on the lower bound for $t_p$ implying identifiability is 149. The graph reflects this in the fact that the the size of the square root of the average predicted error of the E's behaves like the 4's for small $t_p$, and comes down to a reasonable small size around $t_p = 149$.

where $R$ is linearly independent of $S$. This new device model, while simpler than the models ultimately applicable to software-embedded systems, provides a readily-available starting point for testing an extension of the HELP methodology using the concept of Expectation Maximization (EM) which has potential importance for software-embedded systems. The EM approach is attractive because it would provide an efficient method for extending HELP, or other testing tools, to the more complex device model. We studied patterns of the observed data to see if they can identify the model being considered. Our results not only may help to resolve the issue about whether EM works for this situation but also can help engineers to properly design their experiment in such a way that the principle components can still be identified when some of the data are missing.

# Prediction of Shot Noise

Robert Lund[1], Ronald Butler[2], Robert Paige[3]

[1] Department of Statistics,
University of Georgia,
Athens, Georgia, 30602,USA
[2] Department of Statistics, Colorado State University, Fort Collins, Colorado, 80523, USA
[3] Department of Mathematics, University of Northern Iowa, Cedar Falls, Iowa, 50614, USA

**Keywords:** Prediction, Shot Noise, Mean Squared Error, Saddlepoint Appoximation.
*Appeared in Journal of Applied Probability, Volume 36, Number 2*

We consider a shot noise process $\mathbf{A} = \{A_t, t \geq 0\}$ defined for $t \geq 0$ by

$$A_t = \sum_{i=1}^{N_t} Y_i h(t - \tau_i), \tag{1}$$

where $\mathbf{N} = \{N_t, t \geq 0\}$ is a Poisson process with arrival times $\{\tau_i, i \geq 1\}$ and rate generation parameter $\lambda > 0$, and $\{Y_i, i \geq 1\}$ is an independent and identically distributed sequence of "shot marks" that is independent of $\mathbf{N}$.

The model in (1) arises whenever Poissonian arriving shot marks are additive in a super-positioning sense after accounting for a time delay since the arrival of each shot. One simple example of such a process consists of a telephone station where calls arrive via a Poisson process. If each call lasts one time unit, then $A_t$ is the total number of calls in progress at time $t \geq 0$ when one takes $Y_i \equiv 1$ and $h(t) = \mathcal{I}_{[0,1]}(t)$. Further studies of shot noise models include atmospheric noise and radar cluttering application, membrane noise analyses, astronomical applications, and traffic flow studies. Many theoretical stochastic models also follow (1). When $h(t) = \mathcal{I}_{[0,\infty]}(t)$, $\mathbf{A}$ is the well-known compound Poisson process; when $h(t) = e^{-\gamma t}$ for $t \geq 0$ and some $\gamma > 0$, $A_t$ can be viewed as the content of water in a reservoir at time $t$ when $A_0 = 0$ and water is released from the reservoir at rate $\gamma u$ when the content of the reservoir is $u > 0$.

In this work, we consider the problem of predicting future values of $\mathbf{A}$ from a history of observations on a discrete lattice; specifically, we consider prediction of $A_{t_{n+1}}$ from $A_{t_1}, \ldots, A_{t_n}$, where $0 = t_0 < t_1 < \cdots < t_{n+1}$. For simplicity, we assume that the observations of $\mathbf{A}$ are equally spaced:

$t_i = i\Delta$ for some $\Delta > 0$. Since the time scale of $\mathbf{A}$ can always be changed, we take $\Delta = 1$ without loss of generality.
For the ensuing computations, we will require the joint moment generating function (MGF) of $(A_1, \ldots, A_{n+1})^T$ for all $n \geq 0$. The following result identifies the form of this MGF; we use the notation $\psi(s) = E[e^{sY_i}]$ for the MGF of $Y_i$:

$$M(\underline{s}) = \exp\left\{\lambda \int_0^1 \sum_{i=1}^{n+1} \left[\psi\left(\sum_{k=i}^{n+1} s_k h(u + (k-i))\right) - 1\right] du\right\}. \tag{2}$$

Throughout this work we assume that $Y_i$ has the exponential density

$$f(y) = \rho^{-1} e^{-y/\rho} \mathcal{I}_{[0,\infty)}(y) \text{ where } \rho > 0. \tag{3}$$

In addition, we shall consider the exponential shot function, $h(t) = e^{-\gamma t}, \gamma > 0$, the step shot function $h(t) = \mathcal{I}_{[0,1]}(t) + 2^{-1}\mathcal{I}_{(1,2]}(t) + 4^{-1}\mathcal{I}_{(2,3]}(t)$ and linear shot function, $h(t) = \max(1 - t/r, 0)$. These are common shot functions which also allow explicit evaluation of the integral in (2).
With explicit joint MGFs at hand, predictive inference is greatly facilitated by Skovgaard's approximation to $F(a|\underline{A}_n)$, the conditional CDF of $A_{n+1}$ given $\underline{A}_n = (A_1, \ldots, A_n)^T$, and the double saddlepoint approximation to $f(a|\underline{A}_n)$, the conditional density of $A_{n+1}$ given $\underline{A}_n$.
For the exponential and step shot functions, the structure of $h$ over $(n-1, n]$ is linearly related to that of $h$ over $(n, n+1]$ for each $n \geq 1$, i.e.,

$$h(n + \gamma) = \beta_n h(n - 1 + \gamma) \quad \text{for} \quad \gamma \in (0, 1]. \tag{4}$$

We describe such shot functions as being *interval similar*. For interval similar shot functions, $\mathbf{A}$ is autoregressive so that

$$A_{n+1} = \sum_{j=1}^{n} \alpha_j A_{n+1-j} + Z_{n+1} \tag{5}$$

where $Z_{n+1}$ is distributionally equivalent to $A_1$. As such, the conditional MGF of $A_{n+1}$ given $\underline{A}_n$ is also available and given as

$$M(s_{n+1}|\underline{A}_n) = \exp\{s_{n+1}B_n + \lambda \int_0^1 [\psi(s_{n+1}h(u)) - 1]du\} \tag{6}$$

where $B_n$ is a constant determined from the shot noise history.
This conditional MGF forms the basis for two types of saddlepoint approximation procedures. The first procedure involves direct application of the Lugannani and Rice approximation to approximate $F(a|\underline{A}_n)$ and the single saddlepoint approximation to approximate to $f(a|\underline{A}_n)$. In second procedure requires that we first remove the point mass of size $e^{-1}$ at $B_n$ and then approximate to the continuous part of $A_{n+1}$ given $\underline{A}_n$ with the Lugannani and Rice or single saddlepoint approximations.

Before proceeding, we should note that saddlepoint approximation of mixed distributions is not justified by existing theory which only applies to strictly continuous or discrete random variables. We use complex-analytic methods to justify the use of saddlepoint approximations for mixed random variables.

For our numerical work we consider the prediction of $A_6$ based upon $A_1, \ldots, A_5$ for each of our three shot functions. Here we assume the Poisson process driving our shot noise process has rate generation parameter $\lambda = 1$ and the density of $Y_i$ has parameter $\rho = 2$. The values of $A_i$, $1 \leq i \leq 5$, are chosen so that

$$\left( \left( V\left[A_i\right] \right)^{-\frac{1}{2}} \left( A_i - E\left[A_i\right] \right) : i = 1, \ldots 5 \right) = \left( \frac{1}{4}, \frac{1}{8}, 0, -\frac{1}{8}, -\frac{1}{4} \right). \qquad (7)$$

The shot noise history is chosen in this way to facilitate comparison of the various saddlepoint approximation procedures.

First, we consider the exponential shot function with $\gamma = 2$. Here $A_6|\underline{A}_5$ has a point mass of size $e^{-1}$ at $B_5 = 0.087$. From (6) the conditional mean point predictor, $E[A_6|\underline{A}_5]$, is 0.952 and associated conditional mean squared prediction error, $V[A_6|\underline{A}_5]$, is 1.963.

Next, we address the more general problem of reproducing the distribution of $A_6|\underline{A}_5$. The following table shows the accuracy of three different saddlepoint approximations of $F(a|\underline{A}_5)$ at the quantiles listed in the first column. The second column probabilities associated with each quantile were determined by inverting the empirical CDF of 1 million simulations.

| Quantile | Probability | LR Direct | LR Cts | Skovgaard |
|---|---|---|---|---|
| 0.145 | 0.4000 | 0.3856 | 0.4020 | 0.3845 |
| 0.625 | 0.6000 | 0.6040 | 0.6072 | 0.6039 |
| 1.580 | 0.8000 | 0.8021 | 0.8039 | 0.8026 |
| 2.640 | 0.9000 | 0.9001 | 0.9012 | 0.9001 |
| 3.770 | 0.9500 | 0.9496 | 0.9502 | 0.9496 |
| 6.560 | 0.9900 | 0.9897 | 0.9899 | 0.9897 |
| 7.770 | 0.9950 | 0.9948 | 0.9948 | 0.9948 |

The *LR Direct* values are the values of the Lugannani and Rice approximation at the chosen quantiles based on the conditional MGF in (6). The *LR Cts* values are approximations which apply the Lugannani and Rice approximation only to the continuous part of $F(a|\underline{A}_5)$. Skovgaard's approximation is based upon the joint MGF and is also used to find the following point estimates: $\hat{E}[A_6|\underline{A}_5] = 0.965$ and $\hat{V}[A_6|\underline{A}_5] = 1.943$. Of the three, procedure LR Cts is best while LR Direct and Skovgaard are very accurate and quite similar in performance. Surprisingly, the Skovgaard and LR Direct procedures comparable performance since Skovgaard procedure

is based high-dimensional joint MGF, which, in principle, leaves greater room for numerical error.
Next, we consider the step shot function. Here $A_6|\underline{A}_5$ has a point mass of size $e^{-1}$ at $B_5 = 1.12$, $E[A_6|\underline{A}_5] = 3.12$ and $V[A_6|\underline{A}_5] = 8.00$. Again, we present a table like the one above;

| Quantile | Probability | LR Direct | LR Cts | Skovgaard |
|---|---|---|---|---|
| 1.30 | 0.4000 | 0.3643 | 0.4003 | 0.3644 |
| 2.65 | 0.6000 | 0.5833 | 0.5996 | 0.5834 |
| 4.85 | 0.8000 | 0.7948 | 0.7983 | 0.7948 |
| 6.94 | 0.9000 | 0.8978 | 0.8985 | 0.8978 |
| 8.95 | 0.9500 | 0.9488 | 0.9489 | 0.9488 |
| 13.50 | 0.9900 | 0.9899 | 0.9899 | 0.9899 |
| 15.40 | 0.9950 | 0.9950 | 0.9950 | 0.9950 |

From Skovgaard's approximation it is found that $\hat{E}[A_6|\underline{A}_5] = 3.13$ and $\hat{V}[A_6|\underline{A}_5] = 8.06$. Approximation LR Cts is again most accurate while LR Direct and Skovgaard are nearly identical in performance.
Lastly, we considered the linear shot function with $r = 3$. The following table lists quantile values, Skovgaard's approximation and the numerically integrated double saddlepoint approximation at selected probabilities:

| Probability | Skovgaard | Double Density |
|---|---|---|
| 0.4000 | 1.31 | 1.47 |
| 0.6000 | 2.30 | 2.20 |
| 0.8000 | 4.10 | 4.00 |
| 0.9000 | 5.84 | 5.71 |
| 0.9500 | 7.54 | 7.39 |
| 0.9900 | 11.4 | 11.2 |
| 0.9950 | 13.0 | 12.9 |

Here, it is found that $\hat{E}[A_6|\underline{A}_5] = 2.64$ and $\hat{V}[A_6|\underline{A}_5] = 5.81$.
For this last example, simulation conditional on $\underline{A}_5$ not appear practically feasible. In addition, The linear shot function does not allow explicit computation of the left edge of support of the conditional distribution of $A_6$ given $\underline{A}_5$. However, this value may be characterized as the smallest value of $A_6$ for which one can compute Skovgaard's approximation. This value was estimated numerically as 0.48.

# Statistical models with mediation variables for family moral thought

Kenan M. Matawie

[1] School of Business and industry Operations Management, University of Western Sidney, PO Box 10, Kingswood NSW 2747 Australia. email : k.matawie@uws.edu.au

**Abstract:** This study investigated the mediatory role of family processes in the adolescent-parent moral thought relationship. the family processes include the cohesion and adaptability as measured by Olson's(1992) Family Adaptability Cohesion and Evaluation Scale (FACES II) and family communication as measured by Barnes and Olson's (1985) Parent Adolescent Communication Scale (PACS), Moral thought was measured by White's (1997) revised Moral Authority Scale (MAS-R), the MAS-R measures five sources of moral influence: Family, Educators, Self Interest, Society's Welfare and Equality. Results involving 218 adolescent-parent dyads revealed the significance and contribution of mediatory role of family processes, this was obvious and varies according to each parent and the specific moral views being transmitted to adolescents.

**Keywords:** Linear regression; Mediators.

## 1 Family Processes and Moral Thought

The aim of this research is to investigate the extent to which family process variables such as family cohesion, adaptability and communication mediate the parent-adolescent moral thought relationship. It is anticipated that such an analysis will clarify how the content of moral thought, specifically the five sources of moral authority, are transmitted from mothers and fathers to adolescents.

## 2 Method

151 families from regional areas of north New South Wales participated. There were 106 adolescent-parent triads and 45 adolescent-mother dyads. The 110 Female and 41 male adolescent participants were aged between 14 and 19 years, 85% were born in Australia, 10% in Asia and 5% in Europe. The mothers involved were aged between 33 and 59 years, 55% were born in Australia,15% in Asia, 19% in Europe and 11% in the Middle East. The fathers involved were aged between 35 and 76 years, 54Australia, 9% in Asia, 29% in Europe and 8% in the Middle East. Of the mothers

participated, 85% were married and 15% were either divorced or separated, whereas all of the fathers participated remains married.

## 3 Results

Five subscale scores relating to the Family, Educators, Society's Welfare, Equality and Self Interest sources of moral authority as measured by the MAS-R. The justification for analysing and reporting each source separately is revealed in the correlational analyses presented in Table 1. ,

Table 1: Pearson Product Moment Correlation matrix
of adolescent responses to the five sources of moral authority

| | Society | Equality | Family | Educators | Self Interest |
|---|---|---|---|---|---|
| Society | 1.00 | | | | |
| Equality | 0.70 | 1.00 | | | |
| Family | 0.30 | 0.30 | 1.00 | | |
| Educators | 0.39 | 0.30 | 0.69 | 1.00 | |
| Self Interest | 0.40 | 0.32 | 0.40 | 0.31 | 1.00 |

Pearson correlation analysis for the above five sources reveals that although there is some significant association between Society's Welfare and Equality sources and the Family and Educators sources, overall, there is sufficient evidence amongst the five subscales to justify a separate analysis of each source of moral authority.

## 4 Family processes as mediators of the transmission of moral thought from parents to adolescents

Baron and Kenny (1986 p.1176) Defined the term mediator as "A given variable may be said to function as a mediator to the extent that it accounts for the relation between the predictor and criterion". In accordance with this definition, eight regression analysis were conducted for each of moral authority in order to test the mediating role of family cohesion, adaptability and communication (mediating variable) in the transmission of moral thought from parents (predictor variable) to adolescents (criterion variable).,

Specifically, the findings reveal that mothers' thoughts on Society's Welfare have a stronger and significant influence on adolexcent thoufgts on Society's Welfare, when compared to fathers' thoughts. Moreover, it would appear that family cohesion and adaptability make a significant contribution with mothers' moral thought in influencing adolescents' societal concerns. The best fit regression equation, with standardised variables, for adolescents' societal concerns is,, Society = 0.20(mother's beliefs on similar issues) + 0.10 (family adaptability),, ($p < .01$)., Similarly, the following best fit

regression equations,with standardised variables, for adolescents' reveals the nature of the mediating role of family processes in the transmission of moral thought concerning Equality, Family, Educators and Self Interest., Equality=0.30(father's beliefs on similar issues)+0.10 (family cohesion)„ ($p < .01$), Family=0.30 (Mother's beliefs on similar issues)+0.40 (family cohesion)„ ($p < .01$)., Educators=0.30 (father's beliefs on similar issues)+0.15 (family cohesion)„ ($p < .01$)., Self Interest=0.15(father's beliefs on similar issues)+0.15(family cohesion)„ ($p < .01$).,

## 5 Discussion

Generally, the findings suggest that parents play an influential role in their children's moral thinking. It was found that parents play an important role though the content of moral thought to which they ascribe, but also through the way in which they communicate these thoughts, the level of emotional closeness between themselves and their children and the frequency in which they and their children change together. In other words, it was found that family processes such as cohesion, adaptability and communication play a significant mediatory role in the transmission of the content of moral thought from parents to adolescents.

## 6 references

Baron, R.M.and Kenny, D.A. (1986). The moderator-mediator variablr distinctio in social psychological research: Conceptual, strategic, and statistical considerations.*Journal of Social Psychology* **51(6)**, 1173-82.

Baron, H.L. and Olson, D.H. (1985). Parent-adolescent communication and the Circumplex Model. *Child Development* **56**, 438-447.

Olson, D. H., McCubbin, H.I., Larsen,A., Muxen,M. and Wilson, M. (1992). Family inventories: Inventories used in national survey of families (rev. ed). *Los Angeles : Sage.*

White, F.A. (1997). Measuring the content of moral thought: The Revised Moral Authority Scale (MAS-R). *Social Behaviour and Personality* **25, (4)**, 321-334.

# Transverse and Longitudinal Analysis, using Orthogonal Contrasts, for Rank one Common Structures

João Mexia[1], Manuela Oliveira[2]

[1] Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2825 Monte da Caparica, Portugal

[2] Departamento de Matemática, Universidade de Évora, Colégio Luis António Verney, Rua Romão Ramalho 59, 7000 Évora, Portugal e-mail:manuelaoliveira@mail.telepac.pt

**Abstract:** The relations between studies in a series, when STATIS (Structuration des Tableaux À Trois Indices de la Statistique) is applied, are expressed by the Hilbert-Schmidt products of the operators associated to those studies. Studies in a series are said to have a common structure whenever the matrix of Hilbert-Schmidt products is the sum of a rank one matrix $\lambda \vec{\alpha}\vec{\alpha}^t$ with an error matrix. When such a common structure exists, the components of the structure vector $\lambda \vec{\alpha}$ will suffice to locate the studies in the serie Assuming that, for all level combinations of $L$ factors, there are matched series of studies presenting common structures, standardized orthogonal matrices are used to analyze the action of those factors on the corresponding structure vectors. Both transverse analysis on homologuous components of the structure vectors, and longitudinal analysis on linear combinations of those components, are carried out. When the studies in each series are taken at different times, longitudinal analysis characterizes the effect of the factors on the evolution of the series. An application to local Portuguese elections will be presented.

**Keywords:** STATIS, Hilbert-Schmidt products, Structure Vectors, F Tests, Orthogonal Contrasts, Standardized Qrthogonal Matrices, ANOVA.

## 1 Introduction

Superscripts will indicate the number of components of vectors. In what follows studies will be matrix triplets $(X_i, D_{p_i}, D_n)$ constituted by an objects x variables data matrix, and two diagonal weight matrices one for objects and the other for variables. The STATIS method, see Lavit (1988), is widely used in the study of series of studies. The relations between the studies in a series being expressed by the Hilbert-Schmidt products between the matrices $X_i D_{p_i} X_i^t D_n$, $i = 1, \cdots, k$. When the studies have the same weight the series is balanced, and, see Oliveira & Mexia (1998), there

is a rank one common structure if the matrix $S$ of these products is the sum of a rank one matrix $\lambda\alpha^k{\alpha^k}^t$ with an error matrix whose elements are iid $N(0,\sigma^2)$. The studies in the series will correspond to the components of the structure vector $\beta^k = \lambda\alpha^k$. Then, see Oliveira & Mexia (1998), $\alpha^k$ may be estimated by the first eigenvector $\gamma^k$ of $S$, $\beta^k$ by $\tilde{\beta}^k = S\gamma^k$ and, with $S = [s_{i,j}]$, and $g = k(k-1)$, $\sigma^2$ by $\tilde{\sigma}^2 = \frac{D}{g}$, where $D = \sum_{i=1}^{k}\sum_{j=1}^{k} s_{i,j}^2 - \|\tilde{\beta}^k\|^2$.

If for the level combination of $L$ factors we have matched balanced series of studies with rank one common structures, it may be interesting to study the action of the factors on their structure vectors. When we study that action on sets of homologuous components we are carrying out a transverse analysis. We can also apply longitudinal analysis in which the action of the factors on the evolution of the series is considered.

## 2 Orthogonal Contrasts

Let the $L$ factors have $J_1, \cdots, J_L$ levels. The level combinations $(j_1, \cdots, J_L)$ can be ordered by the indexes $i = j_1 + \sum_{l=2}^{L}(j_l - 1)\prod_{h=1}^{l-1} J_h$. For the different series we will have the pairs $(D_i, \tilde{\beta}_i^k)$, $i = 1 \cdots, \overline{J} = \prod_{l=1}^{L} J_l$. To use orthogonal contrasts we will assume that these pairs are independent, that the $D_i$ are the products by $\sigma^2$ of central chi-squares with $g$ degrees of freedom and that the $\tilde{\beta}_i^k$ are normal with mean vectors $\beta_i^k$ and variance- covariance matrices $\sigma^2 I_k$. Clearly these assumptions are only approximatively met but it is well known, for instance, see Ito (1980), that, in the balanced case, ANOVA is robust both towards heterocedasticity and especially against non-normality.

An orthogonal $J \times J$ matrix is standardized orthogonal if the elements in it's first line are equal to $J^{-1/2}$. Let $P_l$ be a standardized orthogonal $J_l \times J_l$ matrix, $l = 1, \cdots, L$. Writing the Kronecker matrix product as $\otimes$, $\bar{P} = P_L \otimes \cdots \otimes P_1$ will also be, see Mexia (1988), standardized orthogonal. Thus the line vectors of $\overline{P}$ will be mutually orthogonal and all but the first one will have null component sums. These matrices can be used, see again Mexia (1988), to apply orthogonal contrasts in ANOVA for complete lay-outs. Each line of $\bar{P}$ is obtained through Kronecker products of lines of the $P_1, \cdots, P_L$, and so it can be associated with the set $C$ of the indexes of those matrices that did not contribute with their first line. Let $A(C)$ be the sub-matrix of the lines associated with $C$. Then the hypothesis

$H_0(C)$ of absence of effects [interactions] for the factor [factors] with index [indexes] in $C$ can be rewritten as $H_0(C) : A(C)\mu^{\overline{J}} = 0^{g(C)}$, with $g(C) = \prod_{l \epsilon C}(J_l - 1)$ and $\mu^{\overline{J}}$ the vector of treatment mean values. If there are $\overline{g}$ degrees of freedom for the error and, with $R$ replicates, $Y^{\overline{J}}$ is the vector of treatment totals, the statistic of the $F$ test for $H_0(C)$ will be $\Im(C) = \frac{\overline{g}}{g(C)} \frac{\| A(C)Y^{\overline{J}} \|^2}{R * SSE}$, where $SSE$ is the sum of squares for the error.

To apply this technique to transverse analysis we have only to replace $\overline{g}$ by $g\overline{J}$, $Y^{\overline{J}}$ by the vectors of homologuous components of the estimated structure vectors and $R * SSE$ by $\sum_{i=1}^{\overline{J}} D_i$. Likewise in longitudinal analysis we study the action of the factors on the values of ${c^k}^t \beta_i^k$, $i = 1, \cdots, \overline{J}$, where $c^k$ is a norm one vector. In the application we try to show how the results of the transverse analysis may lead to the choice of $c^k$.

## 3 An Application

We considered the results of six local elections in three Portuguese districts along the coast and three other interior districts. The objects were the counties and the variables where the percentages of votes for the main parties: PSD, PS, CDS, and PCP, the other parties (OP), the null votes (NV) and abstentions (ABS). We thus had a first factor with three levels: North (N), Center (C) and South (S) and a second factor with two levels: Coast (Ct) and Interior (I). In table 1 we present the estimated structure vectors, and in table 2 the $F$ tests carried out in the transverse analysis for the first and second factors $F(1)$ and $F(2)$ and their interaction $F(1,2)$.

| **N** | | **C** | | **S** | |
|---|---|---|---|---|---|
| Ct | I | Ct | I | Ct | I |
| 22.43 | 23.89 | 30.86 | 26.78 | 24.98 | 23.19 |
| 21.74 | 24.99 | 28.80 | 26.79 | 24.22 | 21.97 |
| 22.13 | 23.38 | 27.25 | 24.52 | 23.69 | 22.40 |
| 22.16 | 28.28 | 25.20 | 25.35 | 25.55 | 22.48 |
| 21.91 | 25.22 | 24.01 | 24.65 | 25.43 | 21.92 |
| 20.92 | 22.52 | 23.74 | 23.75 | 25.80 | 23.28 |

Table 1: *Structure Vectors*

| F(1) | F(2) | F(1,2) |
|---|---|---|
| 7.05 ∗ ∗ | 0.02 | 1.80 |
| 5.67 ∗ ∗ | 0.39 | 0.89 |
| 2.55 | 0.01 | 0.78 |
| 1.58 | 1.97 | 1.29 |
| 1.67 | 1.55 | 0.22 |
| 1.12 | 1.78 | 0.80 |

Table 2: *F tests*

where ** stands for significant at the 1% level.

There where only significant results for the first two elections which point towards the political homogeneization of the country. This conclusion agrees with the common knowledge of the increasing dominance of $PSD$ and $PS$ in Portuguese politics. To show the differences in the evolution of the six series we carried out a longitudinal analysis on the values of $\frac{1}{\sqrt{2}}\beta_{i,6} - \frac{1}{\sqrt{2}}\beta_{i,1}$. The values for this contrast being: $-1.07, -0.97, -5.03, -2.14, 0.58$ and $1.48$. The $F$ tests obtained were $F(1) = 5.21\ **$, $F(2) = 0.61$ and $F(1,2) = 0.50$ so only the $F$ test for the the first factor was significative. These results are coherent with those of the transverse analysis in which the only significant results were for the first factor.

## 4 Final Comments

We used the well established ANOVA robustness, see Ito (1980), to carry out the simultaneous treatment of matched series of studies. In a previous paper, Oliveira & Mexia (1999), the series corresponded to the levels of one factor while now more than one factor are considered.

## References

Ito, P.,K. (1980). Robustness of Anova and Macanova Test Procedures. P. R. Krishnaiah ed, Handbook of Statistics, Vol. I. North Holland.

Lavit C. (1988). Analyse Conjointe de Tableaux Quantitatifs. Collection Méthods+Programmes, Masson.

Mexia, J. T. (1988). Standardized Orthogonal Matrices and the Decomposition of the sum of Squares for Treatments. *Trabalhos de Investigazão*, **2** , FCT/UNL, 1-54.

Oliveira, M., M., and Mexia, J., T. (1998). Tests for the rank of Hilbert-Schmidt product matrices. *Advances in Data Science and Classifications (Rizzi, Vichi and Bock, Ed.)*, 619-625. Springer.

Oliveira, M., M., and Mexia, J., T. (1999). F tests for Hypothesis on the Structure Vectors of Series. *Discussiones Mathematicae. Algebra and Stochastic Methods*, **19**, 345-353.

# Prediction in ARCH-M models: Bootstrap versus parametric methods

Jesús Miguel [1]and Pilar Olave[1]

[1] Dpto. Métodos Estadísticos , Facultad de Ciencias Económicas , Universidad de Zaragoza
[2] Address of second author

**Abstract:** In this paper, we obtain all the theoretical conditional moments of the multi-step prediction error distribution from an ARMA process, when its innovations are represented by a GARCH(1,1) model and whose conditional variance is a regressor variable. Such moments are used to determine the forecast intervals. We propose an alternative bootstrap method for constructing prediction intervals for this model. Finally, we compare both methods in several simulations for different ARMA-GARCH(1,1)-M models. These methods are then applied to analyze the time-varying risk premium in the Spanish stock market.

**Keywords:** ARCH-M model, Bootstrap method, Prediction

## 1 Introduction

ARCH models and their subsequent generalizations have proven to be an ideal technique by which the conditional variance of financial time series can be modelled.
The first paper to apply the ARCH-in-mean specification to asset pricing models was that of Engle, Lilien and Robins (1987). These models are used in most financial series to test several hypotheses about the rationality of the market expectations.
Baillie and Bollerslev (1992) consider forecasting the conditional mean and variance from an ARMA process with GARCH innovations, but they leave out the ARCH-M model. In this paper, we present the prediction of an ARMA model with GARCH(1,1) innovations and whose conditional variance is a regressor variable, that is to say, an ARMA-GARCH(1,1)-M model. Expressions for the MSE predictor are given and all the theoretical moments of the multi-step prediction error distribution are calculated.
This study allows us to determine the conditional distribution of the forecast error ; this information can then be used to construct the forecast intervals, since the usual forecast intervals do not work well. These expansions are based on parametric hypotheses on the moments of the conditional one-step-ahead distribution.

We propose a alternative bootstrap method to construct forecast intervals for an ARMA-GARCH(1,1)-M model. This method estimates the conditional distribution in $t+s$, given the information up to time $t$. We only resample ahead conditional on estimates of the model parameters, that is to say, resample forward values conditional on all previous values of the series. It is a generalisation of that proposed Miguel and Olave (1999) for ARCH models.
Finally, both methods are compared in several simulations. They are then used to analyze the time-varying risk premium in the Spanish stock market, using daily observations from January 1997 to December 1999.

## 2 Prediction intervals

The univariate ARMA(p,q)-GARCH(1,1)-M model is given by

$$\begin{aligned} y_t &= \mu + \delta\sigma_t^2 + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\ \epsilon_t &= z_t\sigma_t \end{aligned}$$

where $\{z_t\}$ is a sequence of zero-mean independent random variables with common distribution $F$ with variance equal to one. The unknown (p+q+5)-vector of parameters is given by $\vartheta = (\mu, \delta, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \omega, \alpha, \beta)$.
The ARMA(p,q) model for the conditional mean is assumed to be covariance-stationary and invertible. Furthermore, we assume $\phi_p \neq 0$ and $\theta_q \neq 0$. The most basic assumptions on the process defined above are those required in order for the process to be well defined and for $\epsilon_t$, and hence $y_t$, to have finite and constant second moments and stationary conditional variance.
In the presence of ARCH, the unconditional distribution of $\epsilon_t$ has fatter tails than the conditional one-step-ahead prediction error distribution. Similarly, the conditional distribution of $\epsilon_{t+s}$ for $s > 1$, given information through time t, differs from the conditional distribution for s=1, because the prediction error distribution generally depends nontrivially on the information set at time t.

### 2.1 Parametric method

The parametric method consists in approximating the quantile of the prediction error distribution by means of the Cornish-Fisher expansions using the conditional moments. We shall make use of the higher-order conditional moments for this process. For simplicity, we assume that this conditional distribution is symmetric with all the existing even-ordered moments proportional to the corresponding powers of the conditional variances.

From the model, the optimal s-step-ahead predictor is readily seen to be

$$\hat{y}_{t+s} = \mu \sum_{i=1}^{s} \xi_{s-i} + \delta \sum_{i=1}^{s} \xi_{s-i} E_t[\sigma^2_{t+i}] + \sum_{i=0}^{p-1} e_1^{'} \Phi^s e_{i+1} y_{t-i} + \sum_{i=0}^{q-1} e_1^{'} \Phi^s e_{p+i+1} \epsilon_{t-i}$$

where $\xi_{s-i} = e_1^{'} \Phi_{s-i} e_1$ with $\Phi$ is the matrix of parameters under the ARMA model and $e_j$ is a (p+q)-vector of zeros, except the jth element which is one. The GARCH(1,1) model can be expressed as an ARMA(1,1) model in the square errors and, hence, the optimal s-step-ahead predictor for the conditional variance is similar to the predictor for the conditional mean. Therefore, the forecast error for the s-step-ahead predictor for the conditional mean is given by

$$e_{t,s} = y_{t+s} - \hat{y}_{t+s} = \sum_{i=1}^{s-1} \gamma_{s-i} \nu_{t+i} + \sum_{i=1}^{s} \psi_{s-i} \epsilon_{t+i}$$

where $\gamma_{s-i} = \delta\alpha \sum_{j=i}^{s-1} \xi_{j-i} (\alpha + \beta)^{s-j-1}$ and $\nu_t = \epsilon_t^2 - \sigma_t^2$. This expression clearly reveals that the s-step-ahead forecast error depends on the future errors for the conditional mean and the future errors for the conditional variance. Such representation will be used in order to determine the theoretical conditional moments up to the fourth order. These moments are used to approximate the quantile of the s-step-ahead prediction distribution [see Miguel and Olave (2000), for more details].

### 2.2 Bootstrap method

However, the previous method does not work well when the one-step-ahead conditional distribution is asymmetric, or does not verify the necessary hypotheses. We propose a bootstrap method, which generally leads to obtaining accurate asymptotic coverage probabilities without parametric hypotheses on the error distribution.
The bootstrap that we describe is designed to mimic the distribution of the forward values, conditional to all the series. Thus, the method is more realistic, given that the ARCH model constructs the volatility based on the lagged residuals.The proposed method does not require a knowledge of the backward residuals, which are difficult to estimate in dynamic models with time-dependent conditional variances. The algorithm of the proposed bootstrap consists of resampling the standardized residuals to construct the forward values using the estimated model. Its steps are similar to those proposed in Miguel and Olave (1999) in the ARCH model.

## 3 Empirical study

We have compared both methods in several simulations with different one-step-ahead distributions and various sets of parameters in the conditional

mean and variance. The results obtained are satisfactory, over all when the distribution is leptocurtic or asymmetric.
Finally, we have studied the IBEX35 Spanish stock market index with daily observations. The results obtained show that the proposed bootstrap works well and presents an advantage when the one-step-ahed conditional distribution does not verify the neccessary hypothesis. This is a frequent ocurrence when we analyze financial time series.

## References

Baillie, R.T. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, **52**, 91-113.

Engle, R. F.; Lillien, D.M. and Robins, R.P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, **55**, 391-408.

Miguel, J.A. and Olave, P. (1999). Bootstrapping forecast intervals in ARCH models. *TEST*, **8**, 345-364.

Miguel, J.A. and Olave, P. (2000). Prediction in ARCH-M models: A theoretical view. *Pre-print*.

# Bayesian Inference in GARCH Models

Laura Muñoz[1] and Manuel Salvador[1]

[1] Department of Statistical Methods, University of Zaragoza, 50005 Zaragoza, Spain

**Abstract:** This paper describes Bayesian Inference for a linear conditionally heterokedastical time series with stable innovations. Our procedure uses Monte-Carlo Methods to generate samples from the joint posteriori density and takes into account the uncertainity concerning the true model We illustrate our approach using simulated data from some particular ARMA-GARCH models and applying the methodology to the Spanish stock market.

**Keywords:** Bayesian inference, GARCH models, MCMC methods

## 1 Introduction

In spite of the well known properties of GARCH models, early empirical work with ARCH models for daily exchange rates indicated that the implied unconditional distributions of estimated ARCH models were not sufficiently leptokurtic to represent the distribution of returns. A parametric alternative which has received little attention in the literature is the stable distributions family. Stable distributions have long been regarded as important generalizations of the normal distributions whose location-scale families are closed under convolution. In a practical setting, stable distributions have attracted considerable interest, because they can allow for skew and for arbitrarily larger tis the case with the normal distribution.
Mandelbrot (1963) and Fama (1965) were among the first to attempt to use stable distributions in modelling stock market price changes; Unfortunately, general statistical analysis based on stable distributions has been hindered by the nonexistence of a simple form of probability density representation. However, Buckle (1995) has recently obtained a particular mathematical representation involving the denstity of a stable distributions. Our proposal extends the ideas of Barnett et al.(1997) to a heteroskedastic framework. For ARMA-GARCH models with stable innovations Bayesian inference using Monte Carlo methods allows the simultaneous handling of parameter estimation, order selection and the law of the error process specification

## 2 Class of Models, Assumptions and Prior Distributions

The class of models we propose is the ARMA-GARCH family with stable innovations. A model is given by:

$$\begin{cases} \Phi(B)\,(w_t) = \Theta(B)\epsilon_t \\ \dfrac{\epsilon_t}{\sigma_t}|\mathcal{F}_{t-1} \sim S_\alpha\,(\beta, 0, 1) \\ \sigma_t^2 = \omega + \displaystyle\sum_{j=1}^{p} \alpha_j \sigma_{t-j}^2 + \sum_{k=1}^{q} \beta_k \epsilon_{t-k}^2 \end{cases}$$

ASSUMPTION 1: The autoregressive and moving-average orders $nar_{\max}$, $nma_{\max}$ and the ARCH and GARCH orders $p_{\max}$ and $q_{\max}$ are specified by the user and represent the maximal degree of the ARMA-GARCH polynomials. The invertibillity and stationarity for the ARMA component is assumed and also the stationarity of the GARCH component.
ASSUMPTION 2: We assume that either $\phi_{nar_{\max}} \neq 0$ or $\theta_{nma_{\max}} \neq 0$ and either $\alpha_{p_{\max}} \neq 0$ or $\beta_{\mathrm{q}_{\max}} \neq 0$
>From the expression provided by Buckle (1995) for the joint density of $T$ i.i.d. observations from a stable distribution the density of the standarized error term is:

$$f\left(\frac{\epsilon_t}{\sigma_t}, z_t|\alpha, \beta\right) = \frac{\alpha}{|\alpha - 1|}\exp\left\{-\left|\frac{\epsilon_t}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}\right\}\left|\frac{\epsilon_t}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}\left|\frac{\sigma_t}{\epsilon_t}\right|$$

where $\{z_t\}_{t=1}^T$ are auxiliary variables used in the representation of the density function of a stable variable and:

$$t_{\alpha,\beta}(y) = \left(\frac{\mathrm{sen}[\pi\alpha + \eta_{\alpha,\beta}]}{\cos\pi y}\right)\left(\frac{\cos\pi y}{\cos\left[\pi\alpha + \eta_{\alpha,\beta}\right]}\right)^{\frac{\alpha-1}{\alpha}}$$

with $\alpha \in (0,1) \cup (1,2]$, $\beta \in [-1,1]$ $z \in (-\infty, \infty)$, $y \in (-1/2, 1/2)$ and with $\eta_{\alpha,\beta} = \beta \min\left(\alpha, 2-\alpha\right)\dfrac{\pi}{2}$
We begin by defining the corespondents indicators variables $I_j; j = 1, ..., nar_{\max}$ $J_i; j = 1, ..., nma_{\max}$, $L_i; i = 1, ..., p_{\max}$, $M_j; j = 1, ..., q_{\max}$ in order to carry out the selection of the order model and use vectorial notation $\overline{I}$, $\overline{J}$,$\overline{L}$ and $\overline{M}$ for these indicator variables. In order to enforce both the stationarity and invertibility of the process, let $(\phi_1^*, ..., \phi_{\mathrm{nar}}^*)$ be the first $nar$ autocorrelations of a stationary autoregressive process with autoregressive polynomial $\Phi(B)$ and let $(\theta_1^*, ..., \theta_{\mathrm{nma}}^*)$ be the first $nma$ partial autocorrelations of an stationary autoregressive process with autoregressive polynomial $\Theta(B)$. We also assume prior indiference about the model order. The invertibility and stationarity assumptions over the ARMA process can be

guaranteed by assuming a uniform prior distribution over $[-1,1]$. We proceed by specifying a prior beta distribution over the parameter $2\delta^*-1$, with $\delta^* = \sum_{i=1}^{p}\alpha_i + \sum_{j=1}^{q}\beta_j$ being the persistence in the GARCH(p,q) model and a Dirichlet distribution over the parameters: $(\alpha_i^* = \frac{\alpha_i}{\delta^*}; i = 1,...,p_{\max}, \beta_j^* = \frac{\beta_j}{\delta^*}; j = 1,...,q_{\max})$. Finally for the parameter $\omega^* = \frac{\omega}{\delta^*}$ as ussually we adopt an inverse gamma prior distribution.

## 3 Posterior Distribution

From Bayes theorem, the posteriori distribution is given by:

$$\pi(\Phi^*, \overline{I}, \Theta^*, \overline{J}, \omega^*, \Gamma^*, \overline{K}, \eta^*, \overline{L}, \phi^*, \overline{M}, \alpha, \beta | z, \overline{\epsilon}_0, \overline{w}_0, \overline{\sigma}_0^2) \propto$$

$$\propto L(\Phi^*, \overline{I}, \Theta^*, \overline{J}, \omega^*, \delta^*, \Gamma^*, \eta^*, \alpha, \beta, z, w | \overline{w}_0, \overline{\epsilon}_0, \overline{\sigma}_0^2)\pi(\Phi^*, \overline{I})\pi(\Theta_j^*, \overline{J})$$

$$\times\pi(\omega^*, \delta^*, \Gamma^*, \eta^*) = \alpha^T over \alpha\, |\alpha - 1|^T \prod_{t=1}^{T}\left|\frac{m_t(w_t, \Phi^* o\overline{I}, \Theta^* o\overline{J}, \overline{w}_0, \overline{\epsilon}_0)}{\sigma_t^\alpha}\right|^{\frac{1}{\alpha-1}} \times$$

$$\times \int_{-1/2}^{1/2}\left[\exp\left\{-\left|\frac{m_t(w_t, \Phi^* o\overline{I}, \Theta^* o\overline{J}, \overline{w}_0, \overline{\epsilon}_0)}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}\right\}\left|\frac{1}{t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}} dz_t\right] \times$$

$$\times\pi(\Phi^*, \overline{I})\pi(\theta^*, \overline{J})\pi(\omega^*, \delta^*, \Gamma^*, \eta^*, \overline{L}, \overline{M})$$

## 4 Sampling Scheme and Conditionals Distributions

We propose to use the Gibbs Sampling algorith, which is carried out by the following sampling scheme, where $S(.)$ denotes:

$$S(.) = \exp\left\{-\sum_{t=1}^{T}\left|\frac{.)}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}\right\}\prod_{t=1}^{T}\left|\frac{m_t(.)}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}$$

and $m_t(.) = m_t(w_t, \Phi^* o, \Theta^* o\overline{J}, \overline{w}_0, \overline{\epsilon}_0)$

**(i)** For each $w_t$, sapling $z_t, t = 1,..,T$ from the distribution:

$$\pi\left(z_t | \Phi^*, \overline{I}, \Theta^*, \overline{J}, \omega^*, \Gamma^*, \overline{L}, \eta^*, \overline{M}, w_t, \alpha, \beta, \overline{w}_0, \overline{\epsilon}_0\right)$$

$$= C\exp\left\{1 - \left|\frac{m_t(.)}{\sigma_t}\right|^{\frac{\alpha}{\alpha-1}}\right\}\left|\frac{m_t(.)}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}$$

**(ii)** Sampling $\alpha$ from the density:

$$\pi\left(\alpha|\beta,\Phi^*,\overline{I},\Theta^*,\overline{J},\omega^*,\Gamma^*,\overline{L},\eta^*,\overline{M},w_t,\overline{w}_0,\overline{\epsilon}_0\right) \propto \left(\frac{\alpha}{|\alpha-1|}\right)^T S(.)$$

**(iii)** Sampling $\beta$ from the density:

$$\pi(\beta|\alpha,\Phi^*,\overline{I},\Theta^*,\overline{J},\omega^*,\Gamma^*,\overline{L},\eta^*,\overline{M},v,\overline{w}_0,\overline{\epsilon}_0,\overline{\sigma}_0^2) \propto \prod_{t=1}^{T}\left|\frac{dt_{\alpha,\beta}(z_t)}{dz_t}\right|^{-1}_{t_{\alpha,\beta}(z_t)=v_t}$$

where $v_t$ are obtained after the reparametrization trhough the transformation $v_t = t_{\alpha,\beta}(z_t)/m_t(w_t,\Phi^* o\overline{I},\Theta^* o\overline{J},\overline{w}_0,\overline{\epsilon}_0)$

**(iv)** Generate jointly $(I_j,\phi_j^*)$ for $j=1,...,nar_{\max}$ a sample from:

$$\pi(I_j,\phi_j^*|\Phi^*_{i\neq j},I_{i\neq j},\Theta^*,\overline{J},\omega^*,\Gamma^*,\overline{L},\eta^*,\overline{M},w,z,\alpha,\beta,\overline{w}_0,\overline{\epsilon}_0,\overline{\sigma}_0^2) \propto$$

$$\propto \frac{1}{2}\pi(I_j=1)S(.)$$

**(v)** Generate jointly a sample $(J_i,\theta_i^*)$ from the distribution:

$$\pi(\theta_i^*,J_i|\theta_{j\neq i},J_{j\neq i},\Phi^*,\overline{J},\mathcal{R}_2,\alpha,\beta,\overline{\epsilon}_0,\overline{w}_0,\overline{\sigma}_0^2) \propto \frac{1}{2}\pi(I_j=1)S(.)$$

**(vi)** Generate a sample $(\overline{L},\overline{M})$ from:

$$\pi(\overline{L},\overline{M}|\Phi^*,\overline{I},\theta^*,\overline{J},\omega^*,\delta^*,\Gamma^*,\eta^*,w,z,\alpha,\beta,\overline{w}_0,\overline{\epsilon}_0,\overline{\sigma}_0^2) \propto$$

$$\propto \exp\left\{-\sum_{t=1}^{T}\left|\frac{m_t(.)}{\sigma_t t_{\alpha,\beta}(z_t)}\right|^{\frac{\alpha}{\alpha-1}}\right\}\left|\frac{1}{\sigma_t}\right|^{\frac{\alpha}{\alpha-1}}\pi(\overline{L},\overline{M})$$

**(vii)** Generate a sample $(\delta^*,\omega^*,\Gamma^*,\eta^*)$ from the density

$$\pi(\delta^*,\omega^*,\Gamma^*,\eta^*|\overline{L},\overline{M},w,z,\alpha,\beta,\overline{w}_0,\overline{\epsilon}_0,\overline{\sigma}_0^2) \propto \pi(\delta^*,\omega^*,\Gamma^*,\eta^*|\overline{L},\overline{M})S(.)$$

## References


BARNETT, G., ROBERT, K. and SHEATER, S. (1993) : Robust Bayesian Estimation of Autoregressive-Moving Average Models. *Journal of Time Series Analysisis*,**Vol. 18**, N.1, pp. 11-28

BUCKLE, D.J. (1995): Bayesian Inference for stable distributions. *J. Am. Stat. Assoc.*, **Vol 90**, pp.605-13.

SANGJOON, K., SHEPHARD, N. and CHIB, S. (1998): Stochastic Volatitility: Lilelihood Inference and Comparison with ARCH Models. *The Review of Economic Studies Limited*, **Vol. 65**, pp. 361-393.

# Perturbations in Sub-Normal Models

Célia Nunes [1], João Mexia [2]

[1] Departamento de Matemática/Informática, Universidade da Beira Interior, 6201-001 Covilhã, Portugal

[2] Departamento de Matemática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Quinta da Torre, 2825-114 Caparica, Portugal e-mail:manuelaoliveira@mail.telepac.pt

**Abstract:** In sub-normal models the observations vector is the sum of two independent components. The first of these stands for what is measured, the second one is a standard normal error vector. To carry out the inference it is assumed that the first component lies inside a given subspace $\Omega$. Then $F$ tests can be derived with optimal properties. We now study the effects of adding a perturbation vector to the model. Two special situations are singled out. The first of these is based on a mixed model studied by Michalski & Zmyslony (1996), while the second one is the case of designs with replicates.

**Keywords:** Sub-Normal Models, Mixed Models, $F$ Tests,Variance Components, Quadratic Spaces.

## 1 Introdution

Superscripts will indicate the number of components of vectors. In the usual sub-normal models the observations vector $Y^n$ is the sum of two independent vectors $Z^n$ and $e^n$. The first of these stands for what is measured, the second one is a normal error vector with null mean vector and variance-covariance matrix $\sigma^2 I_n$, we put $e^n \sim N(0, \sigma^2 I_n)$. To carry out the inference it is assumed that $Z^n \epsilon \Omega$, with $\Omega$ a subspace with dimension $m$.
Let $\omega$ be a subspace of $\Omega$ with dimension $p$. With $\nabla^\perp$ the orthogonal complement of $\nabla$ and $\nu^n_\nabla$ the orthogonal projection of $\nu^n$ on $\nabla$, when $H_0 : Z^n \epsilon \omega$ holds, the test statistic $\Im = \dfrac{n-m}{m-p} \dfrac{\|Y^n_{\overline{\omega}}\|^2}{\|Y^n_{\Omega^\perp}\|}$ ,where $\overline{\omega} = \omega^\perp \cap \Omega$, has an $F$ central distribution with $m-p$ and $n-m$ degrees of freedom. Actually, see Mexia & Dias (1999), through the introduction of a norm in the quotient vector space $\frac{\Omega}{\omega}$, more general hypothesis may be considered. We thus have been able to widen the family of the models for which exact $F$ tests can be derived since we only require that $Z^n$ is independent from $e^n$ and belongs to $\Omega$. When there are perturbations we add a third random vector $W^n$ to the model in order to describe their effects. We will assume that these perturbations act independently from what is measured and from the usual errors, so that $W^n$ is taken to be independent of both $Z^n$ and $e^n$. Moreover to clearly separate $Z^n$ from $W^n$ we will require that $W^n \epsilon \Omega^\perp$.

## 2 Test Power

Since $\overline{\omega}$ and $\Omega^{\perp}$ are mutually orthogonal $e^n_{\overline{\omega}}$ and $e^n_{\Omega^{\perp}}$ will be independent, thus $Y^n_{\overline{\omega}} = Z^n_{\overline{\omega}} + e^n_{\overline{\omega}}$ will be independent from $Y^n_{\Omega^{\perp}} = W^n_{\Omega^{\perp}} + e^n_{\Omega^{\perp}}$. Besides this, when $Z^n = z^n$ and $W^n = w^n$, $\|Y^n_{\overline{\omega}}\|^2$ and $\|Y^n_{\Omega^{\perp}}\|^2$ will be the products by $\sigma^2$ of independent chi-squares with $m-p$ and $n-m$ degrees of freedom and non-centrality parameters $u = \frac{1}{\sigma^2}\|z^n_{\overline{\omega}}\|^2$ and $v = \frac{1}{\sigma^2}\|w^n_{\Omega^{\perp}}\|^2$. Thus the conditional distribution of $\Im$ will be the doubly non central $F$ distribution with $m-p$ and $n-m$ degrees of freedom and non centrality parameters $u$ and $v$, $F(x|m-p, n-m, u, v)$.
Let $U = \frac{1}{\sigma^2}\|Z^n_{\overline{\omega}}\|^2$ and $V = \frac{1}{\sigma^2}\|W^n_{\Omega^{\perp}}\|^2$ have distributions $G_1(u)$ and $G_2(v)$. Since they are independent their joint distribution will be $G(u,v) = G_1(u)G_2(v)$ and the unconditional distribution of $\Im$ will be,
$F(x|m-p, n-m, G) = \int_0^{+\infty}\int_0^{+\infty} F(x|m-p, n-m, u, v)dG(u,v)$.
Since, $F(x|m-p, n-m, u, v)$ increases with $v$

$$\begin{aligned} F(x|m-p, n-m, G) &> F(x|m-p, n-m, G_1) \\ &= \int_0^{+\infty} F(x|m-p, n-m, u, 0)dG_1(u) \end{aligned}$$

Thus when using the critical value $k$, the power of the test in the absence of perturbations, which is $1-F(k|m-p, n-m, G_1)$, is larger than when we have them since now it is $1-F(k|m-p, n-m, G)$. This conclusion holds also for the usual fixed effects models since then we have only to assume that $Z^n$ is fixed having a degenerate distribution. Moreover the expression of $F(x|m-p, n-m, G_1)$ enables us to show, see Mexia & Dias (1999), that, when there are no perturbations, the $F$ tests for sub-normal models are $UMP$ in the family of the tests whose conditional power, given $U = u$, is a function of $u$.

## 3 An Application

Michalski & Zmyslony (1996) present tests for hypótesis $H_{0,i} : \theta_i = 0,\ i = 1, \cdots, m-1$, for mixed models with observations vector $Y^n \sim N(X\beta^k, \sum_{i=1}^{m-1} \theta_i V_i + \theta_m I_m)$ where $\theta_i \geq 0$ $i = 1, \cdots, m-1$, and $\theta_m > 0$. With $\triangle$ the range space of matrix $X$, and $M$ the matrix of orthogonal projection on $\triangle^{\perp}$ it was assumed that the matrices $W_i = MV_iM^t,\ i = 1, \cdots, m-1$, and $W_m = M$ commute and are linearly independent. Thus the quadratic space generated by the $W_1, \cdots, W_m$ will be commutative and have, see Seely (1971) a basis constituted by the orthogonal projection matrices $E_j$ on mutually

orthogonal spaces $\nabla_j$, $j = 1, \cdots, k$. The direct sum of these subspaces will be $\triangle^{\perp}$ and so $W_m = M = E_1 + \cdots + E_k$. Thus $W_i = \sum_{j=1}^{k} a_{i,j} E_j$, $i = 1, \cdots, m$, and $\sum_{i=1}^{m} \theta_i W_i = \sum_{j=1}^{k} \eta_j E_j$ ,with $\eta^k = A^t \theta^m$, where $A = [a_{i,j}]$. Since the matrices $W_1, \cdots, W_m$ are linearly independent, $rank(A) = m$ and, with $B = [b_{i,j}]$ the Moore-Penrose inverse of $A^t$, we have $BA^t = I_m$ and $\theta^m = B\eta^k$. Due to $W_m = \sum_{j=1}^{k} E_j$ we will have $a_{m,1} = \cdots = a_{m,k} = 1$ as well as $\sum_{j=1}^{k} b_{i,j} = 0$, $i = 1, \cdots, m-1$, and $\sum_{j=1}^{k} b_{m,j} = 1.$,
If in the i-th line of $B$, $i = 1, \cdots, m-1$ there are only two non null elements they will be symmetrical and so we will have $\theta_i = b_i(\eta_{j^+} - \eta_{j^-})$ with $j^+$ and $j^-$ the column indexes of the positive and the negative elements in that line. With $u_{j^+} = dim(\nabla_{j^+})$ and $u_{j^-} = dim(\nabla_{j^-})$ we have the $\Im$ test statistic $\Im_i = \frac{u_{j^-}}{u_{j^+}} \frac{\|Y^n_{\nabla_{j^+}}\|^2}{\|Y^n_{\nabla_{j^-}}\|^2}$ , with distribution $F(z|u_{j^+}, u_{j^-}, 0, 0)$ when $H_{i,0}$ holds. We now point out that the occurrence in matrix $B$ of a line with only two non null elements follows from the set of matrices used, and so does not depend on $Y^n$ being normal. Thus if we go over to sub-normal models $Y^n = Z^n + e^n$, with $\Omega$ the direct sum of $\triangle$ and the $\nabla_1, \cdots, \nabla_{k-1}$, and $e^n \sim N(0^n, \theta_m I_n)$, we can have $F$ tests for the $H_{i,0}$, when in the i-th line there are only two non null elements $b_{i,j^+}$ and $b_{i,k}$. Since, now $\nabla_k = \Omega^{\perp}$, we will have $\Im_i = \frac{n-m}{u_{j^+}} \frac{\|Y^n_{\nabla_{j^+}}\|^2}{\|Y^n_{\Omega^{perp}}\|^2}$ and our previous results apply in straightforward way.

## 4 Replications

An interesting instance is when there are $r$ replicates. The observations are thus grouped, each group having $r$ observations. Then $\Omega$ is the subspace of vectors with groups of $r$ identical components since it is natural to assume that the components of $Z = Y^n - e^n$ corresponding to repeated measurements are equal. Perturbations will now correspond to components of $Y^n - e^n$ that differ inside one or more of the groups. These differences being quite distinct from those that originate the error vector $e^n$ so that, to account for them, a perturbation vector $W^n \epsilon \Omega^{\perp}$ may be added to the model.

## References

Mexia, J. T., and Dias, G. C. (1999). F Tests for Generalized Linear Hypothesis in Sub-Normal Models. Workshop Grónow'99, 20-24.

Michalski, A., and Zmyslony, R. (1996). Testing Hypothesis for Variance Components in Mixed Linear Models. *Statistics*, **27**, 297-310.

Seely, J. (1971). Quadratic Subspaces and Completeness. *Annals of Mathematical Statistics*, **42**, 710-721.

# Working Covariance Structures for Binary Spatial Data

Samuel D. Oman[1], Yohay Carmel[2], Ronen Kadmon[3]

[1] Department of Statistics, Hebrew University, Mount Scopus, Jerusalem, 91905 Israel
[2] Department of Forest Sciences, Colorado State University, Fort Collins, CO 80523, USA
[3] Department of Evolution, Systematics and Ecology, Institute of Life Sciences, Hebrew University, Givat Ram, Jerusalem, 91904 Israel

**Abstract:** An important stage in modeling a binary response, observed at points of a lattice, is fitting a working spatial covariance structure. The usual approach uses covariance models appropriate for normally distributed data, and may not adequately capture the correlations present in the data. We describe an alternative structure, which reflects the binary nature of the response and may be more realistic. We illustrate using data on forest vegetation dynamics.

**Keywords:** Exploratory data analysis; Generalized estimating equations; Hierarchical generalized linear models.

## 1 Introduction

In many applications one models binary variables $Y_i$, observed at sites $s_i$ in a spatial lattice, in terms of corresponding vectors $x_i$ of explanatory variables. In analyzing such data it is important to fit an approximate covariance structure. As a purely descriptive tool, the scale of the covariance between $Y_i$ and $Y_j$, in relation to the distance $d_{ij}$ between $s_i$ and $s_j$, may give insight into its cause. On a more analytic level, if the $Y_i$ are modeled using generalized estimating equations (Albert & McShane, 1995; Chung, 1997) then a working covariance matrix is required. In the context of longitudinal data analysis, the importance that the matrix realistically reflect the data is well known (Crowder, 1995; Sutradhar & Das, 1999). The covariances can also help to specify the dependence structure of the underlying random field, if the data are analyzed using a hierarchical generalized linear model (Clayton & Kaldor, 1987; Breslow & Clayton, 1993; Waller et al 1997; Diggle et al 1998; Heagerty & Lele, 1998).
Since covariance models for normal data are well understood, it is natural to apply them to the Pearson residuals $e_i = (Y_i - p_i)/\{p_i(1-p_i)\}^{1/2}$, where $p_i = E(Y_i)$. For example, Albert & McShane (1995) and Chung (1997) use

models such as

$$\gamma_1 e^{-\gamma_2 d_{ij}} \tag{1}$$

for $\mathrm{cor}(e_i, e_j)$. However, this approach does not reflect the fact that $Y$ is binary, and thus may not adequately model the correlation present in the data. To see this, observe from the marginal bivariate distribution of $Y_i$ and $Y_j$ that their correlation $r_{ij}$ must satisfy

$$r_{ij} \leq \left\{ \frac{p_i/(1-p_i)}{p_j/(1-p_j)} \right\}^{1/2} \tag{2}$$

whenever $p_i \leq p_j$. If a model such as (1) is used, all correlations at a given distance must be equal; and (2) shows that this common value will be very small if there is even one pair $(s_i, s_j)$ of sites at that distance apart with $p_i \ll p_j$. Thus (1) will not be able to model larger correlations which might be present between other pairs of sites the same distance apart, but with more similar $p_i$ and $p_j$.

## 2 A Class of Working Covariances

We propose taking the working correlations to be simple approximations to the true correlations which would obtain if $Y$ were modelled by a hierarchical generalized linear model with a probit link and latent Gaussian field $Z = (Z_i)$. If $c_{ij} = \mathrm{cov}(Z_i, Z_j)$, then an approximation due to Pearson (1901) suggests

$$\bar{c}_{ij} = \phi_i \phi_j c_{ij} \tag{3}$$

as a working covariance between $Y_i$ and $Y_j$, where $\phi_i = \phi(\Phi^{-1}(p_i))$ and $\phi$ and $\Phi$ denote the standard normal density and distribution functions. Since the resulting correlations depend on $p_i$ and $p_j$, they avoid the problem discussed following (2). Moreover, if the $Y_i$ do, in fact, arise as threshold variables (as in Albert & McShane,1995 and Heagerty & Lele 1998), then (3) should give more realistic working correlations than (1).
To implement this approach, we first obtain estimates $\hat{p}_i$ of the $p_i$, possibly under the assumption of independence, and define $\hat{\phi}_i = \phi(\ Phi^{-1}(\hat{p}_i))$. Next observe that

$$E(Y_i - Y_j)^2 = f_{ij} + g_{ij} c_{ij}$$

where $f_{ij} = p_i + p_j - 2p_i p_j$ and $g_{ij} = -2\phi_i \phi_j$. Assuming $c_{ij}$ to be a function $c(d_{ij})$ of $d_{ij}$ alone, we compute estimates $\hat{c}_d$ of $c(d)$, for various values of $d$, by weighted regressions of the appropriate $W_{ij} = (Y_i - Y_j)^2$ on estimated $f_{ij}$ and $g_{ij}$. Finally, we graph the $\hat{c}_d$ vs $d$ and, if appropriate, fit the points by a parametric function such as ( refccc).

## 3 Application to Vegetation Dynamics

We illustrate using data on vegetation dynamics for a $600 \times 750m^2$ region (divided into $40 \times 50$ grid cells of $15m \times 15m$) of natural forest in the Galilee mountains in northern Israel. The data are part of a larger set, described in more detail and analyzed from a different point of view by Carmel et al. (1999). Data on tree cover, obtained using image processing of historical aerial photographs from 1964 and 1992 (Carmel & Kadmon, 1998), showed the percentage of cells in which trees were the dominant growth form (compared to shrubs and herbaceous vegetation) to increase from 1% in 1964 to 70% in 1992. It was desired to model tree dominance in 1992 in terms of initial vegetation conditions as well as cell-specific environmental factors (e. g., slope, aspect, distance to nearest stream bed) and anthropogenic factors (such as type and intensity of controlled livestock grazing) which were present during the 28-year period.

Our intent here is not to give a complete analysis of the data, but rather to compare the usual and the proposed approaches in the exploratory step of fitting a working covariance structure. We first estimated $p_i$ from a probit regression of the $Y_i$ (= 1 if trees were the predominant growth form in cell $s_i$ in 1992, and 0 if shrubs or herbaceous growth dominated) on the explanatory variables described above. Plots of $Y_i$ and $\hat{p}_i$ over the region indicated that the covariates captured the major components of spatial trend present in the data, and a directional correlogram of the residuals did not indicate anisotropy.

Following the usual approach, we first computed the empirical correlogram of the Pearson residuals. Letting $r_P(d)$ denote the resulting correlation between $Y_i$ and $Y_j$ at sites separated by distance $d$, we fit the curve $0.70\ times(0.36)^d$ to the graph of $r_P(d)$ versus $d$ (where $d$ is in units of $15m$). This fit predicted the higher correlations slightly better than the "quadratic" fit $0.32 \times (0.75)^{d^2}$. We then used our proposed method to compute the estimates $\hat{c}_d$. The quadratic fit $0.70 \times (0.77)^{d^2}$ predicted the larger $\hat{c}_d$ somewhat better than the linear fit $1.52\ times(0.38)^d$.

As fitting the Pearson and latent correlations lead to different covariance models, it is of interest to know which set of correlations more accurately reflects the correlations actually present in the data. To answer this, we divided the interval $[0, 1]$ into 5 subintervals by the points 0.1, 0.3, 0.7 and 0.9. For each combination $(I, J)$ of two of these subintervals, and distance $d < 4$, we then computed the sample correlation over all pairs $(Y_i, Y_j)$ such that $d_{ij} = d, \hat{p}_i \in I$ and $\hat{p}_j \in J$. Each empirical correlation significantly different from 0 was then compared with the average of the correlations $r_L(i, j) = \hat{\phi}_i \hat{\phi}_j \hat{c}_d / \{\hat{p}_i(1 - \hat{p}_i)\hat{p}_j(1 - \hat{p}_j)\}^{1/2}$ predicted using the latent approach and with the "average" of the corresponding correlations $r_P(d_{ij})$ from the correlogram (which are all the same since $d_{ij} \equiv d$). The $r_L$ fitted the sample correlations (especially the larger ones) better than did the $r_P$, with a root mean squared error of 0.057 as opposed to 0.069.

## References

Albert, P. S. and McShane, L. M. (1995). A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. *Biometrics*, **51**, 627-38.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9- 24.

Carmel, Y. and Kadmon, R. (1998). Computerized classification of Mediterranean vegetation using panchromatic aerial photographs. *J. of Vegetation Science*, **9**, 445-454.

Carmel, Y., Kadmon, R. and Nirel, R. (1999). Spatio-temporal predictive models of Mediterranean vegetation dynamics. *Ecological Applications*, submitted.

Chung, Y.-F. (1997). A central limit theorem for spatial regression based on generalized estimating equations. Ph D thesis, Univ. of Maryland at College Park.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age- standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671-81.

Crowder, C. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, **82**, 407-10.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299-350.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.*, **93**, 1099-111.

Pearson, K. (1901). Mathematical contributions to the theory of evolution - VII. On the correlation of characters not quantitatively measurable. *Philosoph. Transact. of the Royal Soc. of London, Ser. A*, bf 195, 1-47.

Sutradhar, B. C. and Das, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika*, **86**, 459-65.

Waller, L. A., Carlin, B. P., Xio, H. and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease reates. *J. Amer. Statist. Assoc.*, **92**, 607-17.

# Analysis of the Dynamics of an Air Pollutant by Gaussian Hidden Markov Models

Roberta Paroli[1] and Luigi Spezia[2]

[1] Istituto di Statistica, Università Cattolica S.C., Milano, Italia
e-mail: rparoli@mi.unicatt.it
[2] Dipartimento di Ingegneria, Università degli Studi di Bergamo, Dalmine, Italia
e-mail: spezia@unibg.it

**Abstract:** Data sets about air pollution may be examined by hidden Markov models: the dynamics of mean concentrations of a pollutant is supposed to depend on the dynamics of the unobserved climatic situations, described by a Markov chain. The probability density function of every observed variable, given a state of the Markov chain, is supposed gaussian. We show the maximum-likelihood estimators of the parameters of the model obtained by the EM algorithm.

**Keywords:** Discrete Stochastic Processes; EM Algorithm; Sulphur Dioxide.

## 1 Introduction

Air quality control includes the study of data sets recorded by air pollution testing stations. We consider one of the five stations placed in Bergamo (a town in Northern Italy, with 116.000 inhabitants), that working in via Goisis. It records seven kinds of pollutants and each hour we have the mean concentration of every pollutants. We are interested in the analysis of the dynamics of daily mean concentrations of sulphur dioxide ($SO_2$). We suppose that the dynamics of $SO_2$ is influenced by the dynamics of an unobserved process which describes the evolution of the possible climatic situations in Bergamo. The unobserved process is modelled by a Markov chain and the probability density function (*pdf*) of every observation at any time is determined only by the current state of the chain.
Hidden Markov models (HMMs) are discrete-time stochastic processes $\{Y_t; X_t\}$ such that $\{Y_t\}$ is an observed sequence of random variables and $\{X_t\}$ is an unobserved Markov chain. $\{Y_t\}$, given $\{X_t\}$, is a sequence of conditionally independent random variables (*conditional independence condition*) with the conditional distribution of $Y_t$ depending on $\{X_t\}$ only through the contemporary variable $X_t$ (*contemporary dependence condition*).
In this paper the special HMM in which the *pdf* of every observed variable, given a state of the Markov chain, is gaussian is examined; so we have the special model $\{Y_t; X_t\}$ called *gaussian hidden Markov model* (GHMM).

The aim of this paper is to show how the maximum likelihood estimators of the parameters of GHMMs may be suitably obtained using the EM algorithm. The basic model used to study univariate gaussian stationary time series will be introduced in Section 2, then the maximum likelihood estimators will be obtained in Section 3, finally, in Section 4, a highlight of the results obtained in the application on air pollution data will be shown.

## 2 Definition of the model

Let $\{X_t\}_{t\in(1,\dots,T)\subset\mathcal{N}}$ be a discrete, homogeneous, aperiodic, irreducible Markov chain on a finite state-space $S_X = \{1, 2, \dots, m\}$.
The transition probability from state $i$, at time $t-1$, to state $j$, at time $t$, is denoted by $\gamma_{i,j}$, for any state $i, j$ and for any time $t$: $\gamma_{i,j} = \mathrm{P}(X_t = j \mid X_{t-1} = i) = \mathrm{P}(X_2 = j \mid X_1 = i)$. The transition probabilities $(m \times m)$ matrix is $\Gamma = [\gamma_{i,j}]$, with $\sum_{j\in S_X} \gamma_{i,j} = 1$, for any $i \in S_X$. The initial distribution is the vector $\delta = (\delta_1, \delta_2, \dots, \delta_m)'$, where $\delta_i = \mathrm{P}(X_1 = i)$, for any $i = 1, 2, \dots, m$, with $\sum_{i\in S_X} \delta_i = 1$. Since $\{X_t\}$ is a homogeneous, irreducible Markov chain, defined on a finite state-space, it has an initial distribution $\delta$ which is stationary, that is, for any time $t$, $\delta_i = \mathrm{P}(X_t = i)$, for any state $i = 1, 2, \dots, m$. Since $\delta$ is a stationary distribution, the equality $\delta' = \delta'\Gamma$ holds: $\delta$ is the left eigenvector of the matrix $\Gamma$, associated with the eigenvalue one, which always exists. Finally, the hypotesis characterizing HHMs is that the Markov chain $\{X_t\}$ is unobservable.
Let $\{Y_t\}_{t\in(1,\dots,T)\subset\mathcal{N}}$ be some discrete stochastic process, on a continuous state-space $S_Y \equiv \mathcal{R}$, such that it satisfy the conditional independence condition and the contemporary dependence condition. By these two conditions, given a sequence of lenght $T$ of observations, $y_1, y_2, \dots, y_T$ and a sequence of lenght $T$ of unobserved states $i_1, i_2, \dots, i_T$, it results

$$f(y_1, y_2, \dots, y_T \mid i_1, i_2, \dots, i_T) = \prod_{t=1}^{T} f(y_t \mid i_t),$$

where the generic $f(y \mid i)$ is the *pdf* of the random variable $Y_t$, when $X_t = i$, for any $1 \le t \le T$, henceforth denoted $Y_{t(i)}$, for any $i = 1, \dots, m$. The expression of the *pdf* $f(y \mid i)$ may coincide with the expression of a special discrete or continuous random variable, so we can define several types of HMMs. If the *pdf* is gaussian, i.e. $Y_{t(i)} \sim N\left(\mu_i, \sigma_i^2\right)$ and $f(y \mid i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_i}{\sigma_i}\right)^2\right]$, we have the GHMM.

## 3 Parameters estimation

In the case of GHMMs the parameters to be estimated are the entries of the matrix $\Gamma$, the entries of the vector $\delta$ and the parameters of the $m$ gaussian

distributions of $Y_{t(i)}$, for any $i = 1, \ldots, m$. In this paper we obtain, by the EM algorithm, the maximum likelihood estimators of the parameters $\gamma_{i,j}$'s, $\mu_i$'s and $\sigma_i^2$'s, for any $i = 1, \ldots, m$, while the estimator of the initial distribution $\delta$ will be obtained by the equality $\delta' = \delta'\Gamma$, after the estimation of $\Gamma$.

Let $\phi$ be the vector of the $m^2 + m$ unknown parameters, whose entries are the $m^2 - m$ transition probabilities $\gamma_{i,j}$'s, e.g. the off-diagonal entries of $\Gamma$ (the diagonal entries are obtained by difference, given that $\Gamma$ is a stochastic matrix and therefore each row sum equals one: $\gamma_{i,i} = 1 - \sum_{j \neq i \in S_X} \gamma_{i,j}$, for any $i \in S_X$), the $m$ parameters $\mu_i$'s and the $m$ parameters $\sigma_i^2$'s:

$$\phi = \left(\gamma_{1,2}, \gamma_{1,3}, \ldots, \gamma_{m,m-1}, \mu_1, \ldots, \mu_m, \sigma_1^2, \ldots, \sigma_m^2\right)'.$$

Let $y = (y_1, \ldots, y_T)'$ be the vector of the observed data, that is the sequence of the realizations of the stochastic process $\{Y_t\}$; the vector $y$ is incomplete because the sequence of the states of the chain $\{X_t\}$ is missing. Moreover let $L_T(\phi)$ be the likelihood function of observed data:

$$L_T(\phi) = \sum_{i_1 \in S_X} \sum_{i_2 \in S_X} \cdots \sum_{i_T \in S_X} \delta_{i_1} f(y_1 \mid i_1) \prod_{t=2}^{T} \gamma_{i_{t-1}, i_t} f(y_t \mid i_t).$$

To obtain the maximum likelihood estimators of $\phi$, given that we are in a situation with incomplete data, performing the EM algorithm is easier than solving directly the likelihood system. The algorithm is based on an iterative procedure with two steps at each iteration: the first step, E-step, provides the computation of an *Expectation*; the second step, M-step, provides a *Maximization* (for details on the EM algorithm, see McLachlan and Krishnan, 1997).

Let $\phi^{(k)} = \left(\gamma_{1,2}^{(k)}, \gamma_{1,3}^{(k)}, \ldots, \gamma_{m,m-1}^{(k)}, \mu_1^{(k)}, \ldots, \mu_m^{(k)}, \sigma_1^{2^{(k)}}, \ldots, \sigma_m^{2^{(k)}}\right)'$ be the vector of estimates obtained at the $k^{th}$ iteration of the EM algorithm. At the $(k+1)^{th}$ iteration in the E-step, we obtain the function that must be maximized with respect to $\gamma_{i,j}$, $\mu_i$, $\sigma_i^2$, for any $i, j \in S_X$; in the M-step we obtain the vector of estimates $\phi^{(k+1)}$, whose entries are:

$$\gamma_{i,j}^{(k+1)} = \frac{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \gamma_{i,j}^{(k)} f^{(k)}(y_{t+1} | j) \beta_{t+1}^{(k)}(j)}{\sum_{t=1}^{T-1} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)};$$

$$\mu_i^{(k+1)} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i) y_t}{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)}; \quad \sigma_i^{2^{(k+1)}} = \frac{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i) \left(y_t - \mu_i^{(k+1)}\right)^2}{\sum_{t=1}^{T} \alpha_t^{(k)}(i) \beta_t^{(k)}(i)}$$

(for proofs, see Paroli and Spezia, 1999), where

$$\alpha_1^{(k)}(i) = \delta_i^{(k)} f^{(k)}(y_1 \mid i), i = 1, \ldots, m$$
$$\alpha_t^{(k)}(j) = \left(\sum_{i \in S_X} \alpha_{t-1}^{(k)}(i) \gamma_{i,j}^{(k)}\right) f^{(k)}(y_t \mid j), t = 2, \ldots, T, j = 1, \ldots, m$$

and

$$\beta_T^{(k)}(i) = 1, i = 1, \ldots, m$$
$$\beta_t^{(k)}(j) = \sum_{j \in S_X} f^{(k)}(y_{t+1} \mid j)\beta_{t+1}^{(k)}(j)\gamma_{i,j}^{(k)}, t = T-1, \ldots, 1, j = 1, \ldots, m$$

are the *forward* and the *backward pdf*s at the $k^{th}$ iteration, $f^{(k)}(y_t \mid i)$ is the *pdf* of $Y_{t(i)} \sim \mathcal{N}\left(\mu_i^{(k)}; \sigma_i^{2^{(k)}}\right)$, for any $i = 1, \ldots, m$, and $\delta^{(k)}$ is the left eigenvector of the matrix $\Gamma^{(k)} = \left[\gamma_{i,j}^{(k)}\right]$, such that $\delta'^{(k)} = \delta'^{(k)}\Gamma^{(k)}$.
If the algorithm converges at the $(k+1)^{th}$ iteration, that is if the difference $\ln L_T(\phi^{(k+1)}) - \ln L_T(\phi^{(k)})$ is less than or equal to an arbitrary sufficiently small value, then the vector $\phi^{(k+1)}$ is the maximum likelihood estimator of the unknowns parameters of the GHMM.
Explicit formulas of the maximum likelihood estimators of the parameters allow us not to use at each M-step a numerical method of maximization, such as Newton-Raphson's. Hence convergence occurs more quickly and the algorithm is more stable.

## 4 Application to air pollution data

We implemented the foregoing iterative procedure in a GAUSS code, to study a time series of the daily mean concentration of $SO_2$, in the second semester of 1998, recorded by the air pollution testing station placed in via Goisis in Bergamo. To estimate the dimension $m$ of the state-space of the Markov chain, we use the *Bayesian Information Criterion* and we choose a three-state Markov chain. The estimations of the parameters of the three gaussian *pdf*s are $\mu = (3.37, 9.05, 20.71)'$; $\sigma^2 = (2.68, 7.48, 55.44)'$ and those of the Markov chain are

$$\Gamma = \begin{bmatrix} 0.8898 & 0.1102 & 0.0000 \\ 0.1698 & 0.7479 & 0.0822 \\ 0.0312 & 0.1277 & 0.8411 \end{bmatrix}; \qquad \delta = (0.5266; 0.3120; 0.1614)'.$$

### References

McLachlan G. J. and Krishnan T. (1997).

*The EM algorithm and extensions.*

John Wiley & Sons, New York.

Paroli R. and Spezia L. (1999).

*Gaussian hidden Markov models: parameters estimation and applications to air pollution data.*

Serie E.P. n. 94, Istituto di Statistica, Università Cattolica S.C., Milano.

# Finite Sample Properties of a QML Estimator of SV Models with Long Memory

Ana Pérez[1], Esther Ruiz[2]

[1] Department of Statistics and Econometrics, Universidad de Valladolid, Avda Valle Esgueva, 6, 47011 Valladolid, Spain. email: perezesp@eaee.uva.es
[2] Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Madrid, 126, 28903 Getafe, Spain. email: ortega@est-econ.uc3m.es

**Abstract:** We analyse the finite sample properties of a spectral QML estimator of Long Memory Stochastic Volatility models. We show up its poor performance for the parameter values usually encountered in practice. We also discuss a problem of identification when the volatility has a unit root. An empirical analysis of the IBEX35 index of the Madrid Stock Exchange illustrates our findings.

**Keywords:** Fractional integration; Stochastic Volatility; QML estimator.

## 1 Introduction

High frequency financial time series of returns are often characterised by having excess kurtosis and autocorrelated squared observations. Moreover, these autocorrelations tend to decay very slowly indicating that volatility could be characterised as a long memory process; see Ding et al. (1993). To model this behaviour, Harvey (1998) and Breidt et al. (1998) propose Long Memory Stochastic Volatility (LMSV) models and suggest estimating their parameters by Quasi-Maximum Likelihood (QML) in the frequency domain. Breidt et al. (1998) prove the strong consistency of this estimator and analyse its finite sample properties for some of the parameters of the model in the stationary case. We extend their study to all the parameters of the model and to a bigger range of parameter values that include more realistic and non-stationary cases.

## 2 Estimation of LMSV models

We consider the AutoRegressive Long Memory Stochastic Volatility (ARLMSV) model defined by the following equations:

$$y_t = \sigma_* \sigma_t \epsilon_t \quad ; \quad (1-\phi L)(1-L)^d log(\sigma_t^2) = \eta_t \tag{1}$$

where $\sigma_*$ is a scale parameter, $\sigma_t$ is the volatility of series $y_t$, $\epsilon_t$ is a Gaussian white noise process with unit variance, $\eta_t$ is $NID(0, \sigma_\eta^2)$ and independent

of $\epsilon_t$, and $d$ can be non-integer. This model is stationary and invertible when $|\phi| < 1$ and $-1/2 < d < 1/2$, and has long memory if $0 < d < 1/2$ . When $\phi = 0$ , model (1) becomes the basic LMSV model in Harvey (1998); if $d = 0$ and $|\phi| < 1$, we obtain the ARSV (AutoRegressive SV) model in Harvey et al. (1994); when $\{d = 0, \phi = 1\}$ or $\{\phi = 0, d = 1\}$, model (1) becomes the RWSV (Random Walk SV) model.
Model (1) can be linearised by taking the logarithm of $y_t^2$ as follows:

$$log(y_t^2) = \mu + log(\sigma_t^2) + \xi_t \tag{2}$$

where $\mu = log(\sigma_*^2) + E(log(\epsilon_t^2))$ and $\xi_t = log(\epsilon_t^2) - E(log(\epsilon_t^2))$. QML estimation is then carried out by maximising the discrete Whittle approximation to the likelihood function of $log(y_t^2)$ in the frequency domain, treating $\xi_t$ as if it were Gaussian. In the stationary case, Breidt et al. (1998) prove that the QML estimator is consistent. If $0.5 \leq d \leq 1$, the model is no longer stationary, but taking first differences in (2) yields a stationary model for the series $\Delta log(y_t^2)$ and QML estimation can be carried out for that series. Once the parameters have been estimated, the underlying volatility is estimated as the product $\hat{\sigma}_* exp\{0.5\tilde{h}_{t/T}\}$, where $\tilde{h}_{t/T}$ is the smoothed estimated log-volatility series, obtained with the smoothing algorithm in Harvey (1998), and $\hat{\sigma}_*$ is an estimator of the scale factor.

## 3 Finite sample properties of QML estimator

Table 1 displays simulation bias and standard deviations, across 3000 replications, for some selected SV models and for three sample sizes, $T = 1024$, $T = 4096$ and $T = 8192$. We first observe that in the basic LMSV models ($\phi = 0$), the variance of the QML estimators always decrease as $d$ and $T$ increase. In the stationary cases ($0 < d < 1/2$ and $|\phi| < 1$), we also observe a negative bias in the estimator of $d$ and a positive bias in the estimator of $\sigma_\eta^2$ , both decreasing the bigger is $T$, except for the model with $\phi = 0.99$. In this case, the bias of $\hat{d}$ suddenly becomes positive and quite large and the bias of $\hat{\sigma}_\eta^2$ becomes negative, and none of them decreases with $T$. On the other hand, if the parameter $d$ in the ARLMSV model is close to the boundary of nonstationarity ($d = 1/2$), estimators behave in the opposite way: the estimator of $\phi$ is largely overestimated and has much dispersion, while d is always underestimated. This illustrates the problems faced in the estimation of ARLMSV models due to a lack of identification of the parameters when the volatility is close to have a unit root. Regarding the non-stationary LMSV model with $d = 0.75$, it seems that the QML estimator behaves better than in the stationary cases, with small standard deviations and smaller, but positive, bias for $\hat{d}$, and smaller bias for $\hat{\sigma}_\eta^2$.

TABLE 1. Simulation bias and standard deviation (in parenthesis) for the spectral QML estimator under several ARLMSV models

| Parameters | $\phi$ | | | d | | | $\sigma_\eta^2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $(\phi, d, \sigma_\eta^2)$ | T=1024 | T=4096 | T=8192 | T=1024 | T=4096 | T=8192 | T=1024 | T=4096 | T=8192 |
| (0,0.2,0.1) | - | - | - | -0.026 (0.171) | -0.027 (0.157) | 0.000 (0.148) | 0.121 (0.265) | 0.056 (0.149) | 0.036 (0.115) |
| (0,0.4,0.1) | - | - | - | -0.100 (0.178) | -0.026 (0.111) | -0.011 (0.081) | 0.137 (0.265) | 0.061 (0.137) | 0.035 (0.095) |
| (0,0.45,0.1) | - | - | - | -0.101 (0.161) | -0.032 (0.086) | -0.017 (0.063) | 0.141 (0.258) | 0.060 (0.124) | 0.034 (0.083) |
| (0,0.75,0.1) | - | - | - | 0.044 (0.152) | 0.020 (0.081) | 0.015 (0.056) | 0.053 (0.156) | 0.012 (0.064) | 0.004 (0.040) |
| (0.2,0.4,0.1) | 0.167 (0.520) | 0.073 (0.478) | 0.027 (0.425) | -0.136 (0.189) | -0.070 (0.131) | -0.034 (0.092) | 0.116 (0.271) | 0.057 (0.156) | 0.041 (0.121) |
| (0.2,0.45,0.1) | 0.174 (0.530) | 0.072 (0.471) | 0.026 (0.424) | -0.161 (0.181) | -0.065 (0.117) | -0.032 (0.081) | 0.120 (0.271) | 0.059 (0.155) | 0.044 (0.123) |
| (0.99,0.2,0.1) | -0.022 (0.019) | -0.018 (0.010) | -0.018 (0.007) | 0.224 (0.139) | 0.275 (0.070) | 0.286 (0.049) | -0.065 (0.052) | -0.087 (0.018) | -0.090 (0.013) |

## 4 Empirical analysis of IBEX35

Several SV models have been fitted to a series of daily returns of the IBEX35 index of the Madrid Stock Exchange, from 7/1/87 to 30/12/98 (2991 observations), after having removed some small correlation structure in the mean. The kurtosis of the series is 8.321 and the Box-Ljung statistic for the first ten autocorrelations in the squares is $Q_2(10) = 1129.1$, which is highly significant. The series of squared and log-squared returns have significant autocorrelations, even for high lags, with a very slow decay, especially remarkable in the log-squared returns.

We first estimate the ARSV model by the QML method of Harvey et al. (1994), with the following results (standard deviation in parenthesis): $\hat{\phi} = 0.9898(0.0057)$, $\hat{\sigma}_\eta^2 = 0.0168(0.0042)$, $\hat{\sigma}_* = 0.9297(0.0374)$. As expected, the autoregressive parameter is estimated very close to one, indicating persistency in the volatility. We have also fitted a RWSV model with very similar results.

Regarding long memory models, we have first fitted a stationary basic LMSV model with the result that $d$ is estimated on the boundary of nonstationarity ($\hat{d} = 1/2$). Therefore, following Harvey (1998), a model with $d > 1/2$ has been estimated, obtaining $\hat{d} = 0.7538$, $\hat{\sigma}_\eta^2 = 0.0906$, $\hat{\sigma}_* = 1.5112$. Observe that no standard deviation of the estimators is displayed because the asymptotic distribution of QML estimators for LMSV models is still unknown. We have also fitted an ARLMSV model, and the estimated parameters have been $\hat{\phi} = 0.6632$, $\hat{d} = 0.7035$, $\hat{\sigma}_\eta^2 = 0.0155$, $\hat{\sigma}_* = 1.5484$.

Several sample moments of the standardized observations from the models previously estimated have been calculated and it has been checked that the fit from long memory models is better than from the other models,

with correlations in the squares being no longer significant at the 1% level. However, looking at the diagnostics on the residuals, it is difficult to distinguish between the LMSV model and the ARLMSV model. In order to choose between both long memory models, we have carried out a likelihood ratio test to gauge the significance of the AR component in the ARLMSV model. The value of the test statistic is LR=1.0885, which does clearly not reject the null $H_0 : \phi = 0$ compared to the $\chi^2_1$ distribution. Therefore, the LMSV model seems to be preferable.
The implied smoothed volatility of the IBEX35 returns for the four estimated models has also been obtained. As expected, there is no much difference between the volatilities estimated by the ARSV and the RWSV models. Moreover, the volatility implied by the long memory models also show the same pattern, although their fit is better, especially in the high volatility periods.

## 5 Conclusions

We have analysed the finite sample properties of a QML estimator of Long Memory Stochastic Volatility models in the frequency domain. This estimator is shown to behave poorly when the volatility evolves smoothly over time and/or is close to be nonstationary. Even worse, in the unit root case, the parameters of the ARLMSV model are not identified so inference on those parameters is not reliable at all. The poor small sample properties of the QML estimators make it difficult to establish concluding results about the presence of long memory in the volatility of the IBEX35 index.

**Acknowledgements:** We acknowledge financial support from projects SEC97-1379 and PB95-0299 by the Spanish Government.

### References

Breidt, F.J., N. Crato and P.J.F. de Lima (1998). The detection and estimation of long-memory in stochastic volatility. *Journal of Econometrics*, **83**, 325-348.

Ding, Z., C.W.J. Granger and R.F. Engle (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83-106.

Harvey, A.C. (1998). Long Memory in Stochastic Volatility. In J. Knight and S. Satchell (eds.), *Forecasting Volatility in Financial Markets* London: Butterworth-Haineman.

Harvey, A.C., E. Ruiz and N.G. Shephard (1994). Multivariate Stochastic Variance Models. *Review of Economic Studies*, **61**, 247-264.

# Estimation of Total and Direct Effects of Residence in Special Dementia Care Units on Function using Clustered Longitudinal Data

Eva Petkova[1], Jeanne Teresi[2] and Jian Kong[2]

[1] Columbia University, Department of Biostatistics, 1051 Riverside Drive, Unit 48, New York, NY 10032, USA, phone: (212) 543-5875; fax: (212) 543-5599; e-mail: ep120@columbia.edu

[2] Columbia University, Stroud Center

**Abstract:** Data from a large number of nursing homes collected by the National Institute of Aging is used to study the effect of residence in Special Care Units (SCUs) for demented subjects on functioning of residents. Demented subjects from non-SCUs are used for comparison. Subjects are assessed three times six months apart. The total effect of residence, which is a sum of direct effect and indirect effect that occurs through other factors, is estimated using Mixed Effects Models (MEMs) and conditioning on the baseline cognitive level and other covariates. Since a previous study shows that time affects differently the course of cognitive decline of SCU and non-SCU residents, the direct effect of residence in SCU is estimated using MEMs with time and current cognitive level as covariates. The covariance structure of the data, modeled to accommodate the complex sampling scheme and the repeated measurements on a subject, is evaluated for additional information about the process of functional decline. The evidence for beneficial effect of residence in SCUs on functioning is critically evaluated in terms of generalizability of the results and limitations of the causal interpretation.

**Keywords:** mixed effects models; direct and total effects; clustered data; observational study; causal effect.

## 1 Introduction

In this paper MEMs (Laird *et al.* (1982), Beacon *et al.* (1996),Diggle *et al.* (1994)) are used to study the processes of functional decline of institutionalized, cognitively impaired individuals in order to identify factors that influence decline, and to estimate the effect of special dementia care units on functional outcome. Building on path analysis methodology (Freedman (1988), Draper (1995) and deLeeuw *et al.* (1995)) inferences are made here about the direct and total effects of residence in special care units on the functional decline of demented residents, and limits are established regarding the causal relationships that can be inferred.

Residence in various care settings can have both a direct and an indirect effect on the functional outcome of demented residents. The direct effect will result from activities aimed at maintaining the functional independence of the residents. The indirect effect occurs as a result of the residence affecting other aspects of life related to functionality, such as for example, cognition and health. The sum of these two effects is the total effect of residence on the functioning of demented residents. Both direct and total effects are of interest when assessing the utility of special care dementia units and each has separate implications for designing such units, although the observational nature of the available data limits the scope of the inferential conclusions that can be made.

## 2 Results

The total effect was estimated through the following model, which has only one timevarying covariate, i.e., time.

$$Y_{ikt} \quad = \quad age_{ik1} + totpro^2_{ik1} + scu_{ik} * time_{ikt}. \tag{1}$$

According to this model, the slope with respect to time is higher for patients from SCUs indicating that, conditional on baseline cognition ($totpro_{ik1}$), the limitation of subjects residing in SCUs increases faster than does that of subjects from non-SCUs. At the same time, for every given combination of baseline cognition level and age, the intercept of the line for SCU residents is smaller than the intercept for non-SCU residents. This suggests that for a given baseline age and cognitive level the SCU residents have fewer functional limitations at baseline, but that their functional limitation increases faster than does that of non-SCUs residents.
The direct effect was assessed using two "best" fitting models. The first one included as a time varying covariate only cognition:

$$Y_{ikt} = age_{ik} + scu_{ik} * totpro^2_{ikt} \tag{2}$$

This model postulates that all effects of time on the process of functional limitation is through cognitive decline, i.e. the effect of residence on functioning depends only on the cognitive level and not on time. Here all effect of time on function is through its effect cognitive decline.
The second model for estimating the direct effect is the following:

$$Y_{ikt} = age_{ik} + totpro^2_{ikt} + scu_{ik} * time_{ikt} * totpro_{ikt} \tag{3}$$

It has two time varying covariates and is more difficult to interpret than the models above. Based on it, the following observations can be made. First, the effect of time on functional decline depends on the current cognitive level and this dependence is different for residents of SCUs than it

is among their counterparts in non-SCUs. Second, for every given cognitive level the relationship between time and functional limitations is linear, and for fixed cognitive level the functioning of non-SCUs residents remains constant, whereas the functional limitations of SCUs residents increase or decrease over time depending on the cognitive level. Third, except for very cognitively impaired residents ($totpro < 10$) SCUs residents manifest less functional limitation than do non-SCU residents at all times and the difference is larger for residents with higher cognitive capacity.

Based on models (1), (2), (3) and from results of previous work, Liu *et al.* (2000), which examines the effect of time on cognition among SCU and non-SCU residents, it can be concluded that residence in an SCU is associated with better functioning at most cognitive levels. At the same time, residence in SCUs affects cognition negatively in the sense that the cognitive decline is faster in SCUs as compared with non-SCUs, Liu *et al.* (2000). This indirect effect of residence on functioning through cognition results in faster increase of functional limitations for SCU residents.

## 3 Conclusions

Turning to defining the scope of our conclusions, the data analyzed here come from a complex multistage cluster random sampling of patients residing in large nursing homes in the state of New York. The response rate at the facility-level was very high, 70%, and at resident-level was greater than 90%, supporting the conclusion that the sample is not a *sample of convenience*, but satisfies the *exchangeability* principle, which makes the results generalizable to nursing homes in New York State. Careful planning and execution of the sampling design as well as the low rate of missing data supports a conclusion for lack of bias in the results. Therefore, of most concern in terms of making any causal inference is the lack of randomization of patients to an SCU or a non-SCU.

Subjects from SCUs and non-SCUs were sampled after they have been admitted to the respective unit. Moreover, subjects from facilities that did not have SCU could not have been assigned to an SCU, while patients from facilities with an SCU could have been assigned to either an SCU or a non-SCU, depending on the judgment of the staff. This creates the possibility that patients from SCUs are different from those in non-SCUs in a way that can be important for making conclusions about the effect of residence. If the patients from SCUs and non-SCUs differed only with respect to these two variables, or more generally, with respect to only observed variables, the problem would not be unsurpassable, because the *ignorability* principle, required for valid causal conclusions, would have been satisfied. However, it is very likely that doctors and staff at nursing homes employ a variety of judgments when assigning patients to different units, and that these judgments take into account factors other than subjects' measurable characteristics. In light of this, attribution of the effects estimated with this

analysis solely to residence in an SCU or a non-SCU, has to be made with extreme caution.
It is emphasized that the most current knowledge has been used in the design and execution of this study and that all factors known or speculated to influence the functional abilities of nursing homes residents with dementia have been included in our analyses. Nevertheless, because a randomized controlled study is not feasible in this setting, the causal interpretation of the results should be viewed with healthy skepticism. The best strategy for cross-validation of the results would be replication, using perhaps data from some of the other NIA studies conducted in different states.

**Acknowledgements:** The authors would like to thank the members of the Columbia University, Biostatistics in Psychiatry Workshop for their helpful comments and suggestions, provided during several earlier presentations of these data. Additionally, the authors are grateful to Douglas Holmes for valuable comments and review of the manuscript.

## References

Beacon, H. and Thompson, S. (1996) Multi-level models for repeated measurement data: Application to quality of life data in clinical trials. *Statistics in Medicine*, **15**.

Diggle, P.J., Liang, K-Y and Zeger S.L. (1994). *Analysis of longitudinal data.* Oxford: Calderon Press.

Draper, D. (1995). Inference and hierarchical modeling in the social sciences. *J. Educational and Behavioral Statistics*,**20**.

Freedman D. (1988) As others see us: A case study in path analysis. *J. Educational Statistics*,**12**.

Laird N.,Ware, J. (1982). Random effects models for longitudinal data. *Biometrics*, **38**.

de Leeuw, J. and Kreft, I. (1995). Questionaing Multilevel Models. *J. Educational and Behavioral Statistics*,**20**.

Liu X, Teresi J, Waternaux C. (2000). Modelling the decline pattern in functional measures from a prevalent cohort study. *Statistics in Medicine* (In press).

# On the Number of Letters Being Sent. An Applications of Semiparametric Mixed Models.

Christian Pfeifer[1] and G.U.H. Seeber[1]

[1] Institut für Statistik, Universität Innsbruck, Universitätsstraße 15, A–6020 Innsbruck (Austria)

**Abstract:** Random coefficients are introduced into the splined variable of a semiparametric regression model (Green and Silverman (1994, p. 64 ff.) to describe heterogeneity of a temporal effect. Estimates for random coeffcients are used to classify units by a hierarchical cluster algorithm.

**Keywords:** Semiparametric regression; linear mixed model; cluster analysis.

## 1 Introduction

For the purposes of cost accounting and controlling the federal Austrian postal system seeks, for each of about 2300 post offices, quarterly estimates for the number of letters that, through various channels, get into and out of the postal system.
Data are collected by a variation of stratified random sampling, where post offices are assigned to strata defined by regional and organizational criteria. Within each stratum six days are randomly selected for each post office, on which, for each input and output channel, the number of letters have to be counted. No other measurements, such as weighing or counting jars, are taken.
Estimates for the totals are based on a statistical model fitted to the data. Based on exploratory analyses, comparison with other kinds of models, as well as other considerations we use a model of the form

$$\log(y_{ijt}) = \mu + \alpha_i + \beta_{ij} + \gamma_{jt} + f(t) + \varepsilon_{ijt}, \tag{1}$$

where $y_{ijt}$ denotes the counted number of letters in post office $i$ on day of the week $j$ in week $t$. $\gamma_{jt}$ corresponds to a dummy variable indicating whether or not day $j$ in week $t$ follows a holiday with post offices closed. Errors $\varepsilon_{ijt}$ are assumed to be independent and distributed as Gaussian with mean 0 and constant variance $\sigma^2$. The *temporal component* $f(t)$ describes the general pattern of outcomes as it developes over time. $f(t)$ is unique to and characterizes a stratum. We used both purely non-parametric estimates for $f(t)$ and smooth estimates, obtained by loess or smoothing splines.

The defintion of strata is crucial for obtaining good estimates for the total sum of letters, which requires strata to be as homogenous as possible with respect to seasonal fluctuations described by $f(t)$. Bootstrap procedures have been used to compute confidence bands for $f(t)$ and to gain insight as to the variability of estimated totals.

## 2 A Semiparametric Mixed Effects Model

Assume $f(t)$ to be estimated by cubic regression B-splines with $K$ interior knots, i.e.

$$f(t) \approx \sum_{l=1}^{K+3} \zeta_l B_l(t). \tag{2}$$

Then model (1) reads as

$$\log(y_{ijt}) = \mu + \alpha_i + \beta_{ij} + \gamma_{jt} + \sum_{l=1}^{K+3} \zeta_l \cdot B_l(t) + \varepsilon_{ijt}. \tag{3}$$

For each post office $i$ we now introduce random coefficients $\eta_{il}$, $l = 1, \ldots, K+3$ into the splined variable

$$\log(y_{ijt}) = \mu + \alpha_i + \beta_{ij} + \gamma_{jt} + \sum_{l=1}^{K+3} (\zeta_l + \eta_{il}) \cdot B_l(t) + \varepsilon_{ijt}. \tag{4}$$

Random coefficients $\eta_{il}$ are assumed to be independent of $\varepsilon_{ijt}$ and distributed as Gaussian with mean zero and variance $\tau_l^2$. The sum in (4) may be written as

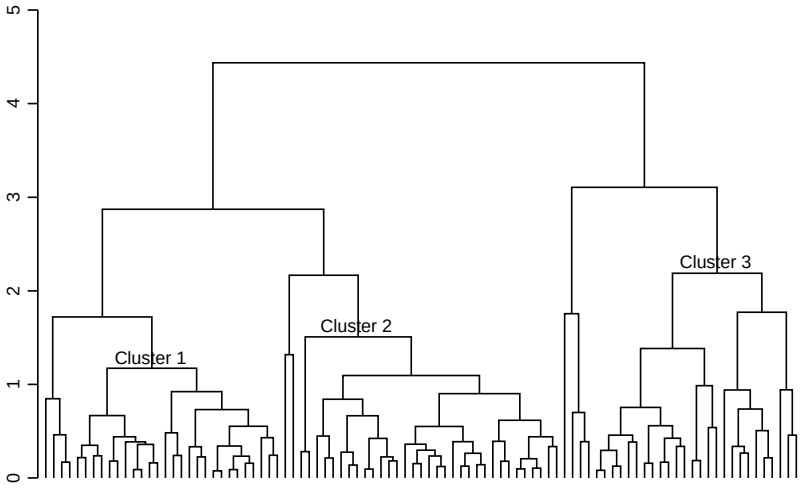
$$\boldsymbol{B} \cdot (\boldsymbol{\zeta} + \boldsymbol{\eta}_i), \tag{5}$$

where the $[n \times (K+3)]$ matrix $\boldsymbol{B}$ denotes the B-representation for the given set of $K$ interior knots (see, e.g. de Boor 1978, chapter IX). Formally (4) defines a linear mixed model.

## 3 Classifying Post Offices

Note that estimates $\widehat{\boldsymbol{\zeta}} + \widehat{\boldsymbol{\eta}}_i$ describe the temporal component for post office $i$. This is a $(K+3)$ vector of real numbers. The Euclidean distance between two such vectors defines a sensible measure of dissimilarity of temporal patterns associated with post offices, which can be used to classify offices by an appropriately chosen clustering algorithm.
For illustrative purposes we fitted model (4) to data collected for 95 large post offices in Tyrol from April 1, 1998 until September 30, 1999. $K = 5$

Figure 1: Dendrogram for large Tyrolean post offices




equidistant interior knots had been selected. BLUP estimates were obtained for the random coefficients. See Venables and Ripley (1997, section 10.3) for a description of `lme`, the S-Plus function we have used on a Linux platform. Figure 1 displays the dendrogram describing the agglomeration scheme for a hierarchical, complete-linkage clustering algorithm. We selected three clusters with a total of 85 post offices as indicated in the graph. The table below gives estimates for the standard deviations of random components for the entire set of post offices and the three clusters, showing strikingly smaller values for the latter.

**Estimated Standard Deviations $\widehat{\tau}_l$ for Variance Components**

| | all offices | cluster 1 | cluster 2 | cluster 3 |
|---|---|---|---|---|
| 1 | 0.3153 | 0.2451 | 0.1767 | 0.1283 |
| 2 | 0.2746 | 0.0047 | 0.0008 | 0.4460 |
| 3 | 0.3399 | 0.0320 | 0.0115 | 0.0220 |
| 4 | 0.3531 | 0.1396 | 0.0009 | 0.1491 |
| 5 | 0.1701 | 0.1403 | 0.0004 | 0.3358 |
| 6 | 0.1532 | 0.0005 | 0.0964 | 0.0025 |
| 7 | 0.3260 | 0.0005 | 0.0028 | 0.2511 |
| 8 | 0.1046 | 0.0178 | 0.0043 | 0.1289 |

Figure 2: Number of letters posted in large Tyrolean post offices

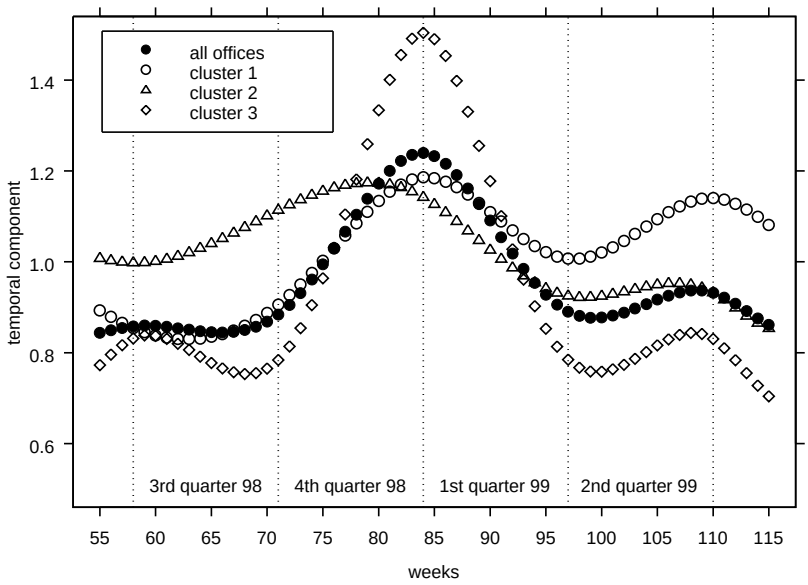



Figure 2 shows estimated temporal components based on fixed effects only. Vertical lines denote the position of interior knots, which are placed at the end of a quarter year. Post offices in cluster 1 seem to have a slightly increasing trend, as opposed to those found in cluster 2. Cluster 3 shows large seasonal variation.

In conclusion we found the combination of mixed models, regression splines and cluster analysis useful in classifying units by their performance over time, thus providing better estimates for the required totals. There is, however, heavy demand on computational resources.

## References

de Boor, C. (1978). *A Practical Guide to Splines.* New York: Springer.

Green, P.J., and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach.* London: Chapman & Hall.

Venables, W.N., and Ripley, B.D. (1997). *Modern Applied Statistics with S-PLUS. Second Edition.* New York: Springer.

# Modeling with the Laplace distribution: Goodness of Fit tests

Pedro Puig[1] and Michael A. Stephens[2]

[1] Unitat d'Estadística , Departament de Matemátiques, Universitat Autónoma de Barcelona, 08193 Bellaterra, Spain
[2] Department of Mathematics and Statistics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6

**Abstract:** Tests are given for the Laplace or double exponential distribution. The test statistics are based on the empirical distribution function (EDF) , and include the Cramér-von Mises, Anderson-Darling and Watson statistics, and the Kolmogorov-Smirnov family. An application of the Laplace distribution is in LAD (or $L^1$) regression, used when the errors appear to have long tails. The tests are also useful in this context.

**Keywords:** EDF tests; Cramér-von Mises statistics; LAD regression.

## 1 The Laplace Distribution

In this paper we give tests of fit for the Laplace, or double-exponential, distribution with density function:

$$f(x;\alpha,\beta) = \frac{1}{2\beta} e^{-|x-\alpha|/\beta} \qquad -\infty < x < \infty \tag{1}$$

where $\alpha$ and $\beta$ are location and scale parameters which are unknown.
The Laplace distribution is used to model symmetric data with long tails. This distribution also arises directly when a random variable occurs as the difference of two variables with exponential distributions with the same scale.
The Laplace distribution, used to model the errors, provides a motivation for the use of LAD (or $L_1$) regression, where parameter estimates are based on minimizing the sum of absolute values of the residuals, rather than on least squares.
On the theoretical side, a justification of the Laplace distribution has been given by Gnedenko (1982). The classical Central Limit Theorem establishes that the distribution of a large sum $S$ of independent and identically distributed variates with finite variance can be approximated by a normal (Gaussian) distribution and this result is often used to justify use of the normal distribution for a random variable. However, the number $n$ of variates may not be fixed but is itself a random variable, and then the Gaussian

approximation generally fails. Gnedenko (1982) establishes that if $n$ has a geometric distribution with a large mean then the distribution of $S$ can be approximated by the Laplace distribution. Gnedenko gives an application to reliability data.
Another property relates the Gaussian and Laplace distributions. A Laplace variate can be generated by a Gaussian with variance following an exponential distribution. Hsu (1979) gives this result and uses the Laplace distribution to model the position errors observed in large navigation systems. Finally, the properties of the Laplace distribution, and many references to other applications, are given by Johnson, Kotz, and Balakrishnan (1990).

## 2 Goodness-of-fit tests

The null hypothesis is $H_0$: the random sample $x_1, x_2, ..., x_n$ comes from the Laplace distribution (1).
The test procedure is as follows. Suppose the order statistics (ascending) of the sample are $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

a) Find the maximum likelihood estimates of the parameters as follows. The estimate of $\alpha$ is the median of the sample; thus $\hat{\alpha} = x_{(\frac{n+1}{2})}$ if $n$ is odd, and $\hat{\alpha} = \frac{1}{2}\{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\}$ if $n$ is even. The estimate of $\beta$ is $\hat{\beta} = \sum_{i=1}^{n} |x_i - \hat{\alpha}|/n$.

b) Make the transformation $z_{(i)} = F(x_{(i)}; \hat{\alpha}, \hat{\beta})$, for $i = 1, ..., n$, where $F$ is the distribution function of (1). The $z_{(i)}$ will be in ascending order.

c) The Cramér-von Mises statistic is computed from

$$W^2 = \sum_{i=1}^{n} \left\{z_{(i)} - (2i-1)/(2n)\right\}^2 + 1/(12n),$$

the Watson statistic is

$$U^2 = W^2 - n\left(\bar{z} - \frac{1}{2}\right)^2,$$

where $\bar{z}$ is the mean of the $z_i$, and the Anderson-Darling statistic is

$$A^2 = -n - (1/n)\sum_{i=1}^{n}(2i-1)\left[\log(z_{(i)}) + \log(1 - z_{(n+1-i)})\right].$$

The Kolmogorov statistics are computed from

$$D^+ = \max_i \left\{i/n - z_{(i)}\right\}; D^- = \max_i \left\{z_{(i)} - (i-1)/n\right\};$$

then $D = \max(D^+, D^-)$ and $V = D^+ + D^-$.

d) Select the table for the chosen statistic, say $Z$, from the paper of Puig and Stephens (2000). Interpolate to find the percentage point $Z^*$, at a given significance level, say $\alpha^*$, for the sample size $n$. If $Z$ is greater than $Z^*$, the null hypothesis is rejected at level $\alpha^*$.

On the whole, the Kolmogorov-Smirnov statistics are less powerful than the Cramér-von Mises statistic, and the latter are the recommended statistics for practical use. Also, the distributions are easier to work with, since the asymptotics can be found and the finite-$n$ distributions converge quickly to the asymptotic. Of this family, $U^2$ is the best statistic against the symmetric alternatives that we have checked.
The consequences of reject or accept $H_0$ are important from the practical point of view. For instance, Bain et al. (1973) presented a data set traditionally considered Laplace distributed. This data set has been used by Kappenman (1977) and Shyu and Owen (1986a, 1986b) to illustrate their procedures for computing tolerance intervals for the Laplace distribution. However the EDF tests reject that data follows a Laplace distribution. Then a nonparametric method for computing the tolerance interval would be better.

## 3 The Laplace distribution and LAD regression

Suppose the regression model is

$$y_i = \beta_0 + x_i'\beta + \sigma e_i,\ i = 1, ..., n \tag{2}$$

where the $x_i$ are vectors of covariates, $\beta_0$, $\beta$ and $\sigma$ are unknown parameters, and $e_i$ are independent random errors. As stated earlier, the estimation of the parameters in LAD regression is done by minimizing the sum of the absolute values of the residuals. For an exposition of LAD regression see Bloomfield and Steiger (1983) and Birkes and Dodge (1993).
The technique is important when the errors might be supposed to have long-tailed distributions. In particular, when the distribution is Laplace, the LAD estimator of $(\beta_0, \beta)$, is the maximum likelihood (ML) estimator $(\hat{\beta}_0, \hat{\beta})$, and the ML estimator of $\sigma$ is $\hat{\sigma} = SAR/n$, where $SAR = \sum |y_i - x_i'\hat{\beta} - \hat{\beta}_0|$. Thus if the errors $e_i$ can be modelled as coming from a standard Laplace distribution, maximum likelihood gives justification to the use of LAD regression.
The fitted regression will then give estimated residuals which are correlated, but the important results of Pierce and Kopecky (1979) can be used, applied to Laplace distributed errors; EDF statistics computed from the regression standardized residuals will have the same limiting distributions as they would if the residuals are treated as a random sample (that is, uncorrelated), with unknown parameters.

Some examples of application of the EDF goodness of fit tests to LAD regression can be found in Puig and Stephens (2000), one of them related with the paper of Parker (1988), where a kind of Box-Cox transformation is presented in this context.

## References

Bain, Lee J. and Englehardt, Max. (1973) Interval Estimation for the Two-Parameter Double Exponential Distribution. *Technometrics* **15**, 875-887.

Birkes, D. and Dodge, Y. (1993) *Alternative Methods of Regression.* New York: John Wiley & Sons.

Bloomfield, P. and Steiger, W. (1983) *Least Absolute Deviations: Theory, Applications, and Algorithms.* Boston: Birkhäuser.

Gnedenko, B. V. and Gnedenko, D. B. (1982) Laplace distributions and the logistic distribution as limit distributions in probability theory. *Serdica* **2**, 229–234.

Hsu, D.A. (1979) Long-tailed Distributions for Position Errors in Navigation. *Appl. Statist.* **28**, 62-72.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1990) *Distributions in Statistics, Volume 1: Continuous Univariate Distributions.* Boston: Houghton Mifflin.

Kappenman, R. F. (1977) Tolerance Intervals for the Double Exponential Distribution. *JASA* **72**, 908-909.

Parker, I. (1988) Transformations and Influential Observations in Minimum Sum of Absolute Errors Regression. *Technometrics* **30**, 215-220.

Pierce, D.A. and Kopecky, K.J. (1979) Testing goodness of fit for the distribution of errors in regression models. *Biometrika* **66**, 1-5.

Puig, P. and Stephens, M.A. (2000) Tests of fit for the Laplace Distribution with applications. *Technometrics.* To appear in November 2000.

Shyu, J.C. and Owen, D.B. (1986a) One-Sided tolerance intervals for the two-parameter Double Exponential Distribution. *Commun. Statist.B -Simula.* **15(1)**, 101-119.

Shyu, J.C. and Owen, D.B. (1986b) Two-Sided tolerance intervals for the two-parameter Double Exponential Distribution. *Commun. Statist.B -Simula.* **15(2)**, 479-495.

# Evaluating Agreement in a Study on Diagnosis of Silicosis

Arminda Lucia Siqueira[1], Otaviano Francisco Neves[1], Ana Paula Scalia Carneiro[2]

[1] Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, C.P. 702, 30161-970, Belo Horizonte, MG, Brazil. E-mail: arminda@est.ufmg.br
[2] Departamento de Medicina Preventiva e Social, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil.

**Abstract:** In this paper we compare the results obtained by two image methods for diagnosis of silicosis using weighted kappa and loglinear models. The interobserver agreement was assessed through kappa. The confidence interval for kappa was constructed using exact method and the Monte Carlo procedure was needed in some cases. We fitted nested loglinear models in order to choose the one better describing the structure of agreement in the data.

**Keywords:** Loglinear model; Silicosis; Weighted kappa.

## 1 Introduction

Silicosis is an occupational lung disease caused by the inhalation of crystalline silica, common among underground mine workers. Although potentially an avoidable disease it still represents a serious health problem in developing countries, where the incidence and prevalence remain at high levels. The diagnosis in the initial stages is not trivial and requires a specific training. At present, Chest Radiography (CR) is the principal instrument used in the diagnosis, following the International Labor Organization (ILO) rules. Recently, the possibility of evaluating the incipient cases by means of High Resolution Computed Tomography (HRCT) has been considered. One disadvantage is the lack of film models to help the diagnosis on a comparative basis, like the ILO CR collection.

In Brazil, the State of Minas Gerais has the highest register of silicosis cases due to the large number of gold miners. One estimates that, in the last 15 years, the Brazilian Social Security Institute (INSS) has given around 4500 benefits for silicosis in a certain area of Minas Gerais.

From those 4500 individuals, we consider a group of 1500 former gold miners receiving benefits due to previous diagnosis of silicosis and who are putting in a claim on the employers. They expect to receive indemnity due to the supposed diagnosis of silicosis obtained by previous CR. However, in most cases those diagnoses did not follow ILO rules. All of them are males

with age ranging from 40 to 70 years old, the majority retired and with incomplete first degree of education.
There are twelve possible classifications for CR (0/-,0/0, 0/1, 1/0, 1/1, 1/2, 2/1, 2/2, 2/3, 3/2, 3/3 and 3/+ coded as 1 to 12, respectively) and the four categories for HRCT (0, 1, 2 and 3) correspond to the numerator of CR. The higher the score the greater the severity of the disease.
In our study, there were three CR readers, who will be referred to as readers 1, 2 and 3 (the third one is a B reader) and two HRCT readers, called readers 4 and 5. Only for the cases of disagreement between HRCT readers there was a third reader (reader 6).
In Brazil, for benefit purposes, the technical definition of silicosis is the exposed worker with compatible clinical and occupational history, whose CR was classified starting from the category 1/0 by two experts who were qualified by the INSS (Algranti, 1998). Among the 135 sampled individuals, all with good quality CR performed in the ILO standards, 67 positive diagnoses were identified (diseased). The remaining 68, undefined or negative results, had HRCT investigation.
HRCT is an expensive method, its price being approximately ten times higher than CR. Furthermore, there is no consensus about the superiority of HRCT in the studies comparing both methods in the evaluation of incipient forms of silicosis. In this paper we compare the results obtained by HRCT and CR in the diagnosis of silicosis in its initial phases (dubious or borderline cases) as well as the agreement among CR and HRCT readers.

## 2 Agreement among readers

The agreement among readers of the two methods was evaluated using weighted kappa coefficient ($\kappa$) as described in Agresti (1996). All the calculations were obtained using StatXact4 and in some cases the Monte Carlo procedure was necessary to get exact results (Metha & Patel, 1999).
For CR, the interobserver agreement evaluation was based on all the 135 individuals and also on the subgroup of those 68 who had been diagnosed by HRCT. We consider first twelve categories (scored as 1 to 12) and then four categories (0, 1, 2, 3), comparable to HRCT.
Table 1 shows the exact estimates and the confidence intervals for kappa for the 135 individuals tested for CR and those 68 tested for CR and HRCT. For the 135 individuals, there is high agreement among the CR readers using twelve or four categories. For the 68 individuals tested for both methods, the agreement diminished considerably, there being no pattern for the three CR readers. This is expected because those are incipient cases for whom the diagnosis is more difficult.
There is poor agreement between the HRCT readers: the estimate and the 95% confidence interval for kappa considering readers 4, 5 and 6 are: $\hat{\kappa}_{4,5} = 0.14$ and $(-0.02, 0.31)$; $\hat{\kappa}_{4,6} = 0.16$ and $(-0.20, 0.51)$; $\hat{\kappa}_{5,6} = 0.33$ and $(0.13, 0.54)$. The last two sets of calculations are based on 40 observations.

TABLE 1. Exact estimates and 95% confidence intervals for kappa for the total tested for CR ($n = 135$) and those tested for CR and HRCT ($n = 68$)

| $n$ | Method (No. of classes) | Readers 1 and 2 | 1 and 3 | 2 and 3 |
|---|---|---|---|---|
| 135 | CR (12) | 0.88 | 0.87 | 0.79 |
| | | $(0.83, 0.92)$ | $(0.82, 0.91)$ | $(0.72, 0.86)$ |
| 135 | CR (4) | 0.83 | 0.82 | 0.72 |
| | | $(0.77, 0.89)$ | $(0.77, 0.88)$ | $(0.63, 0.81)$ |
| 68 | CR (12) | 0.45 | 0.26 | 0.26 |
| | | $(0.32, 0.58)$ | $(0.09, 0.44)$ | $(0.13, 0.39)$ |
| 68 | CR (4) | 0.45 | 0.30 | 0.15 |
| | | $(0.26, 0.65)$ | $(0.06, 0.53)$ | $(0.003, 0.30)$ |

## 3 Agreement between the two methods of diagnosis

First we compare CR (four categories) with HRCT through weighted kappa. Considering the median given by the three CR and HRCT readers, there are only two categories (0 and 1) for both methods with the following distribution: 50 cases in the cell (CR=0, HRCT=0), 8 in the cell (CR=1, HRCT=0) and 5 in the cells (CR=0, HRCT=1) and (CR=1, HRCT=1). Although 81% (55/68) of diagnosis agree, the value of kappa is only equal to 0.32 with a pretty large range of the 95% confidence interval: $(0.04, 0.61)$. Based on kappa we would say there is poor agreement between CR and HRCT, but there seems to be a problem with kappa in assessing agreement for those sparse data, as noticed in the literature.

The identification of the patterns of agreement becomes richer through models. Several models can be used for evaluating agreement between rating on an ordinal scale (e.g. Tanner & Young, 1985; Agresti, 1988; Agresti, 1992; Becker & Agresti, 1992). A recommended procedure compares several nested loglinear models, that is, the fit of a sequence of models. We used this approach for evaluating agreement between methods.

The general form of the model for $r$ readers is given by $\log m_{i,j,\ldots,q} = \mu + \lambda_i^{R_1} + \lambda_j^{R_2} + \ldots + \lambda_q^{R_r} + \delta_{i,j,\ldots,q}$, for $i, j, \ldots, q = 1, \ldots, c$ (the number of categories), where $m_{i,j,\ldots,q}$ is the expected cell count, $\mu$ represents the overal effect, $\lambda_i^{R_1}$the effect of the reader 1 on category $i$ and so on. The parameter $\delta_{i,j,\ldots,q}$ assumes different values depending on the structure of agreement. For instance, $\delta_{i,j,\ldots,q} = 0$ for the independence model; $\delta_{i,j,\ldots,q} = \delta$ if $i = j = \ldots = q$ for $i = 1, \ldots, c$ and 0 otherwise for global agreement among readers, that is, agreement among the $r$ readers. We also consider the following models: agreement among $r - 1$ readers (labelled quasi agreement among readers); global agreement between methods, that is, there is agreement between two methods if two CR readers agree with

two HRCT readers, independent of the category; category heterogeneity, that is, the previous model, but distinguishing among categories; subgroup heterogeneity; that is, the same as the global agreement between methods, but distinguishing among pairs of CR readers.
Table 2 displays the results obtained by fitting those models. As expected, the last model gives the best fit indicating the need of incorporating the difference of agreement among the CR readers. However, the category heterogeneity model, also with a good fit, is more interesting in practice. The interpretation for both models is that there is no substantial difference between methods (CR and HRCT).

TABLE 2. Results for log-linear models: deviance($D$) and degree of freedom ($df$)

| Model | $D$ | $df$ | $D/df$ |
|---|---|---|---|
| Independence | 24.9293 | 60 | 0.4155 |
| Global agreement among readers | 21.9227 | 59 | 0.3716 |
| Quasi agreement among readers | 13.4523 | 58 | 0.2319 |
| Global agreement between methods | 21.6711 | 58 | 0.3736 |
| Category heterogeneity | 10.0578 | 57 | 0.1765 |
| Subgroup heterogeneity | 4.7028 | 56 | 0.0840 |

## References

Agresti, A. (1988). A model for agreement between ratings on an ordinal scale. *Biometrics*, **44**, 539-548.

Agresti, A. (1992). A modeling patterns of agreement and disagreement. *Statistics Methods in Medical Research*, **1**, 201-218.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Algranti, E. (1998). Occupational lung diseases in Brazil. In: Banks, D. E., Parker, J. E. *Occupational Lung Diseases: an International Perspective*. London: Chapman & Hall.

Becker, M. P., and Agresti, A. (1992). Loglinear modeling of pairwise interobserver agreement on a categorical scale. *Statistics in Medicine*, **10**, 101-114.

Mehta, C. R., and Patel, N. R. (1999). StatXact4 for Windows (Version 4.0.1) User Manual.

Tanner, M. A., and Young, M. A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, **80**, 175-180.

# Exact Logistic Regression for Modelling Chagas' Disease Data

Arminda Lucia Siqueira[1], Maria Cláudia F. M. C. Souza[1], Flávia Komatsuzaki[1], Eliane C. D. M. Gontijo[2] and Leonor Bezerra Guerra[3]

[1] Departamento de Estatística, Universidade Federal de Minas Gerais (UFMG), C.P. 702, 30161-970, Belo Horizonte, MG, Brazil. E-mail: arminda@est.ufmg.br
[2] Departamento de Medicina Preventiva, UFMG, Belo Horizonte, MG, Brazil
[3] Departamento de Morfologia, UFMG, Belo Horizonte, MG, Brazil

**Abstract:** When the sample size is small, the asymptotic theory results might be inappropriate or for sparse data, the asymptotic solution of the logistic regression may not exist. In those cases, exact methods are recommended. In this paper we illustrate exact logistic regression, based on the conditional likelihood, with two data sets, an experimental design and an epidemiological study, about Chagas' disease, very common in the State of Minas Gerais, Brazil.

**Keywords:** Chagas' disease; Exact methods; Logistic regression.

## 1 Introduction

The goal of many studies is to determine the effect of one or several covariates on the response. In particular for binary response variable ($Y = 0, 1$) the logistic regression is a very common model. For $k$ covariates $(x_1, x_2, \cdots, x_k)$, the model is given by $\log[p/(1-p)] = \gamma + \sum_{i=1}^{k} \beta_i x_i$, where $p = \Pr[Y = 1]$. The effect of the covariates on the outcome can be evaluated by the coefficients of model ($\beta_i$). The usual statistical inference for them is based on asymptotic theory results.

In many data sets, the asymptotic results do not exist. This can happen for small sample size or highly unbalanced design, that is, sparse data (Mehta & Patel, 1995).

Methods based on asymptotic theory usually require large samples, but this might be unfeasible. Agresti (1996) gives a method for calculating the sample size needed for logistic regression. He shows that moderate to large sample size is required in order to have a reasonable power.

In some situations, the sample size is small for several reasons, such as, difficulty in collecting data, inclusion and exclusion criteria for medical studies, when the event of interest is rare and the study is expensive. In those cases, exact methods are recommended.

## 2 Exact Logistic Regression

Maximum likelihood is the usual method for estimating the parameters in a logistic model. When these estimates do not exist (for instance in any situation cited in Section 1), exact estimates can be obtained by an efficient algorithm presented by Hirji et al. (1987), among others.
Mehta & Patel (1995) provide details for the exact inference (estimation and test of hypotheses) on the coefficients of a logistic regression of unstratified and stratified binary data. Hirji (1988) discusses exact inference for matched case-control studies. Procedures of exact inference are implemented in the software LogXact (Mehta & Patel, 1996) for all these situations.
The main idea for exact inference is to generate the joint and conditional distributions of the sufficient statistics for the coefficients of the model. Standard tests (e.g. score, Wald and likelihood ratio tests) can be used, but instead of approximate distributions for the test statistic, the exact distribution needs to be generated. The conditional exact inference can be accomplished by deriving the permutation distributions, therefore requiring fast algorithms to be computationally feasible.
The exact method is more conservative than the asymptotic one. In several data sets there is difference between both methods for p-value and confidence interval. The exact method, considered the gold standard, is recommended when the asymptotic results are in question.

## 3 Exact Inference Applied to Chagas' Disease Data

Chagas' disease is a systemic pathology caused by *Trypanosoma cruzi*, a protozoan discovered by Carlos Chagas in 1909 and it is transmitted to humans and other animals by an insect. After the bite of this bug, the *T. cruzi* penetrates into the blood stream reaching mainly muscular tissues in the heart and digestive tract where it proliferates, producing inflammation, lesion of nerve terminals and fibrosis.
It is a tropical disease, very common in the State of Minas Gerais, Brazil, especially in the rural area where very poor housing propitiates the entrance of the vector insect.
We illustrate the exact inference for logistic model with two data sets on Chagas' disease, both conducted at the Federal University of Minas Gerais (UFMG). We used the software LogXact (Mehta & Patel, 1996).
The sample sizes are very small: 31 rats for the first study and 76 individuals, corresponding to 38 pairs for the second one. The justification for such sample sizes is the difficulty of obtaining homogeneous rats and the extremely laborious process of collecting data for the experimental design. For the epidemiological study, all cases of the disease were observed in individuals already naturally infected in the investigated group, so it is not possible to increase the sample, since it is an observational study.

### 3.1 Experimental Design

The experiment conducted in the Neurobiology Laboratory - UFMG was designed to characterize the infection by *T. cruzi* and to investigate the involvement of inflammatory cells and one of their products nitric oxide in the phenomenon of heart sympathetic denervation, that is, the lesions of nerve terminals which are important in the regulation of cardiac function. Thirty-one male adult Holtzman rats were inoculated with Y strain of *T. cruzi*. In order to investigate the role played by inflammatory cells, presence of T. cruzi in the tissue and nitric oxide in the denervation process, *T.cruzi*-infected and non-infected rats were treated with the immunosuppressive drug cyclophosphamide (CY), or with the anti-inflammatory agent dexamethasone (DX) or with the esther methyl NG-nitro-L-arginine (L-NAME), a competitive inhibitor of nitric oxide synthesis.

The outcome variable for the logistic model is denervation (yes or no) and the following covariates were considered: percentage of inflammation ($X_1$), percentage of parasite nests ($X_2$). $X_1$ and $X_2$ were categorized using the quartiles as the cutpoints. The treatment groups were identified by the following indicator variables: inoculated, control group ($I_1 = 0, I_2 = 0, I_3 = 0$), inoculated + CY ($I_1 = 1, I_2 = 0, I_3 = 0$), inoculated + DX ($I_1 = 0, I_2 = 1, I_3 = 0$) and inoculated + L-NAME ($I_1 = 0, I_2 = 0, I_3 = 1$). There is no asymptotic solution for the model with all those variables ($X_1, X_2, I_1, I_2, I_3$) and also for other models with some of those variables (for instance including $X_{1,} I_1, I_2, I_3$). Moreover, in some cases the asymptotic p-value is substantially different from the exact one.

Table 1 shows the conservative feature of the exact method. Based on asymptotic or exact inferences there is evidence that the inflammation causes the damage of cardiac noradrenergic nerve terminals. The conclusion for the effect of L-NAME depends on the estimation methods: using exact method we can conclude that it does not inhibit the sympathetic denervation, but based on asymptotic p-value we can identify some effect of this drug.

TABLE 1. Results of a logistic model

| Covariate | Method | $\hat{\beta}$ | C.I. for $\beta$ (95%) | p-value |
|---|---|---|---|---|
| Constant | asymptotic | $-3.9133$ | $(-6.8740, -0.9526)$ | 0.0096 |
| | exact | $-3.4606$ | $(-7.9189, -0.5237)$ | 0.0125 |
| Inflammation | asymptotic | 2.3344 | $(0.7708, 3.8979)$ | 0.0034 |
| | exact | 2.1147 | $(0.7477, 4.3283)$ | 0.0002 |
| $I_3$ (L-NAME) | asymptotic | $-2.9079$ | $(-6.0844, 0.2687)$ | 0.0728 |
| | exact | $-2.6983$ | $(-7.4121, 0.6339)$ | 0.1613 |

### 3.2 Epidemiological Study

This study was developed in the Chagas' Disease Ambulatory of Clinical Hospital - UFMG in partnership with the SLU, the Urban Cleaning Service of the city of Belo Horizonte, State of Minas Gerais, Brazil.
In order to assess the impact of living and working conditions, a diseased individual was matched with a non-diseased one. The sample consists of 76 SLU manual labours, that is, 38 pairs (10 of females and 28 of males).
Besides epidemiological data (age, sex, race, previous and present living conditions, etc.), there is also information on health, occupational history, habits (e.g. alcohol and smoking), previous and present anamnesis, physical examination, infection by *T.cruzi* (serologically tested) and standard electrocardiogram. There are more than 130 variables in this data set. The exact logistic regression was used to assess which variables are related to the presence of disease.
LogXact treats this matched case-control as stratified data (pairs are the strata). The difficulty appears due to the large number of covariates to be considered. Since there is no automatic selection methods, we start with the univariate analysis for the most important variables, then we add the second variable and so on. The final model included the covariates time of staying in the endemic area, coded as 1 if greater than 5 years and 0 otherwise ($p = 0.0190$), and dysphagy ($p = 0.0007$). Curiously, there is no difference between diseased and non-diseased individuals concerning absentee of work ($p = 0.1475$; estimate and 95% confidence interval for odds ratio are 0.4575 and $[0.1567, 0.2871]$) and even social and health history (results not shown). Those results demonstrate that there is no reason for stigma of having Chaga' disease as can be observed in real life.

### References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* New York: Wiley.

Hirji, K. E., Mehta, C. R., and Patel, N. R. (1987). Computing distributions for exact logistic regression. *Journal of the American Statistical Association*, **82**, 1110-1117.

Hirji, K. E., Mehta, C. R., and Patel, N. R. (1988). Exact inference for matched case-control studies. *Biometrics*, **44**, 803-814.

Mehta, C. R., and Patel, N. R. (1995). Exact Logistic regression: theory and examples. *Statistics in Medicine*, **14**, 2143-2160.

Mehta, C. R., and Patel, N. R. (1996). LogXact for Windows (Version 2.0) User Manual.

# Management of Agricultural Data Using Qualitative and Statistical Modelling

S. Tzortzios[1], N. Gitsakis[1] and G. Adam[1]

[1] University of Thessalia, Faculty of Agriculture - Crop and Animal production, Lab of Biometry, Pedion Areos 383 34 Volos, Greece, Tel/Fax: +30 421 74259, Email: stzortz@uth.gr

**Abstract:** The increasing complexity of agricultural data and their management, in finding solutions for certain farmer's problems, requires adequate tools. Fortunately, developments in computer technology continually expand the possibilities in agricultural data analysis and processing. In this study, we tackle some of the important issues in agricultural data management and processing, using an integration of statistical tools and qualitative modelling techniques, in order to describe the complex structure of agricultural processes running in a specific cattle's breed scheme. Specific methodologies were used to make more efficient use of data collected by providing a means of effectively analyzing data. The implementation of the proposed approach is based mainly upon the use of object-oriented and visual development tools applied into the creation of an interactive and friendly computer environment for agricultural data management, where various data manipulation and analysis techniques, as well as experimentation and further research in a cooperation with the farmer, can easily be contacted. The application system seems to be a very useful tool in organizing the information regarding agriculture, by providing important utilities offered by the new technology in the field of agriculture.

**Keywords:** Database management; integrated environment; qualitative modeling; data processing.

## 1 Introduction

The approaches and methodologies used for agricultural data analysis and processing in general continuously evolve (e.g. specific statistical analysis methods and tools, such as SPSS, are used quite intensively to assist the researcher's work). The basic idea in our research work is to provide an integrated environment, where various data analysis and modelling tools would be at the disposal of the researcher to be used in processing farming production problems and extracting adequate solutions. For this purpose, certain database management and qualitative modelling techniques have been used in conjunction, as an integrated computing environment, called AgroModel, and tested upon specific cattle breeding cases. Artificial intelligence and qualitative modelling techniques have been applied for a long

period of time, with quite successful results in most of the cases [5]. However in the field of agriculture there is still a need for further research work to be carried out. We decided to use and apply qualitative techniques describing the structure and performance of plants and animals within agricultural environments, in order to assist the agriculturist to manage easily complicated processes, associated in particular with cattle breeding, and provide the ability to extract and evaluate the most valuable information from a set of complicated with various factors quantitative data. The retrieval of all the relevant information on the control treatments in various agricultural cases could be considered as a quite important research material for interesting studies of the plant species or livestock breeds in various experimental environments.

## 2 Material and Methods

The overall application work was carried out using the above mentioned integrated environment AgroModel and a qualitative modelling tool, called QMTOOL [1]. In our integrated approach, we used that qualitative modelling tool in conjunction with AgroModel ][4] in order to produce working models of the agricultural process under control, and test them in order to acquire the desired functionality, prior to their farm implementation. The development of the integrated application environment of this study was based mainly upon the use of object-oriented and visual development tools, and the use of open architecture technology drivers and methods - ODBC interface, SQL- to interact initially with Microsoft Access and Excel databases and later on with certain statistical packages, such as SPSS, on a Windows 98 operating system platform, [2], [3]. In particular software modules were created as VBasic modules scripts and SQL queries in order to facilitate the communication of model components and execution of internal functions and procedures. The idea was to use an interactive and friendly environment for agricultural data management, where data analysis (e.g. statistical) and experimentation, as well as further research on contemporary data analysis techniques, could easily be contacted.

## 3 Results and Discussion

In our case study, we used only a subset of real data (of Charolais Breed - MLC Beef Cattle recording scheme) extracted from the database to work with, as an example of farmer's level specific interest. It was important to find the tools to describe and analyze such agricultural data structures (cattle breeds and their characteristics), in order to specify and select the most adequate qualitative scheme, without an in-depth requirement for

programming skills. It was also necessary to be flexible enough to allow easy modifications of the given data structure according to any new requirements. Using the facilities of the modelling tool mentioned above, system models were created by simply connecting input, state and output objects and assigning to their connections qualitative values of their magnitudes and relationships. During the execution phase, the system converts qualitative attributes into numerical data in order for the appropriate simulations calculations to take place. This conversion is based on qualitative to numerical values conversion tables, describing basic numerical and alphanumerical factors such as herd's size, dam's category, etc in the selected Charolais breed data scheme. The mapping of mathematical equations (interrelation of data variables), shown as objects connections between the input, state and output variables, into qualitative descriptions is carried out using the following functional: M+(Invar,Stvar), M-(Stvar,Stvar), f(Stvar,Outvar) arithmetic: add, minus, etc. and derivative: incr, steady, decr, etc., constraints. The qualitative marks M+, M- and f simply indicate that there is a relationship (influence), positive or negative (qualitative terms representing the magnitude of the functional relationship) between these variables. The actual value calculation of a given variable is based on its current value (state) plus a sum of influenced values of the preceding variables. A single influence is calculated taking into consideration the value of its predecessor and the magnitude of the connection expressed in qualitative terms. For instance, given that calf's quality factor (Cq) is in functional relationship with the Sire's and dam's class categories, this is expressed in the following way: CalfQuality: $fM + Sc, M + Dc$. Internally, qualitative modelling involves the interpretation and execution of such system's equations, based on qualitative methods for modelling physical systems.

As a general conclusion of the results presented could be realized that the degree to which a user could exploit the AgroModel depends on his goal and his level of knowledge and experience. In cases of general database use, an elementarily trained farmer could accomplish his main management requirements; for a more educated researcher there are advanced tools, which could be exploited according to specific scientific purposes. We have addressed the problem of modelling and managing some of the important cattle breed characteristics, using a tool that utilizes both conventional numerical methods and more advanced qualitative techniques, in order to deal efficiently with proper animals selections of the most productive ones. In particular, we were able to:

- Produce a reliable description of the specific cattle breed scheme based on qualitative models. ·
- Provide flexibility in the manipulation of various cattle parameters in qualitative forms during the model's design. ·

- Produce accurate results of cattle's behaviour comparing to physical representations.
- Reduce the cost of cattle's management by reducing the risk of taking wrong selections decisions.

Qualitative models of the selected cattle's scheme were introduced at a high-level abstraction form, using relatively small amount of information, similar to human reasoning on studying complex physical system's behaviour. This approach of the application system AgroModel among its general importance in the agricultural industry as a whole (education, research and production), seems to be a very useful tool in organizing the information regarding agricultural data, while at the same time provides the utilities for the best exploitation of the knowledge gained up today in the field of conventional agriculture. This application environment tends to be improved and incorporate further agricultural data (plant and animal), as a broader biometrical agricultural network database, in order to provide a fully integrated environment where various conventional and sustainable agricultural data could be gathered for studies of scientific and practical purposes.

# Bibliography


[1] Adam G. K. and Grant E., "QMTOOL - A qualitative modelling and simulation CAD tool for designing automated workcells," *In: Proceedings of the 1994 IEEE International Conference on Robotics and Automation*, Vol. 2, pp. 1141-1146, 8-13 May 1994, San Diego USA.

[2] Law M. Averill, Kelton W. David, "Simulation Modelling and Analysis", McGraw-Hill International Editions, Industrial Engineering Series, 1991.

[3] Riley J., "Statistical usage in biological sciences: main problems and solutions" *In: 5th Meeting of the International Biometric Society*, Network for Central America and Caribbean, August 1997, Xalapa, Mexico.

[4] Tzortzios S. and Adam G., "Proper computer procedures (AgroModel) in plant and animal selection for research and educational purposes", *In: INTAGRED-99 International Workshop Information Technology in Agricultural Education*, 27-29 September 1999, Moscow, Russia.

[5] Valavanis K. and Saridis G., "A review of intelligent control based methodologies for modelling and analysis of hierarchically intelligent systems", *In: Intelligent Control: Proceedings of the 5th International Symposium*, Northeastern University, MA, 1990, IEEE Computer Society Press.

# Internet Statistics from NetSizer

Sam Weerahandi[1]

[1] Telcordia Technologies, 445 South St., Morristown, NJ 07960

**Abstract:** NetSizer, the Telcordia Internet Sizer. is an Internet tool that keeps track of the growth of the Internet. This work deals with the problem of estimating the number of Internet hosts. A demonstration of the tool will be given and an overview of the sampling methodology and estimation methodology will be provided

**Keywords:** Internet Host Counts ; Network Monitoring; IP address.

## 1 Introduction

In the poster session we will demonstrate and discuss about statistics available from NetSizer, the Telcordia Internet Sizer. It has become a popular web site among Internet analysts to obtain various statistics related to the growth of the Internet. Its public page is located at http://www.netsizer.com. For many years Telcordia has been heavily involved in measuring, monitoring and forecasting Internet growth and has an active program to continually expand capabilities in this area. estimates a variety of important statistics related to the size and growth of the Internet, and Telcordia makes many of NetSizer's statistics publicly available free of charge. NetSizer's more advanced capabilities are not publicly available and are used by Telcordia in its consulting engagements.

## 2 NetSizer Current Statistics

NetSizer currently produces the following statistics, in particular:

- number of Internet hosts by country
- number of Internet users and subscribers by country
- number of hosts by by top level domain
- number of hosts belonging to ISPs and to other second level domains

By the latter we mean that the host counts are produced by domain by Internet Service Provider (e.g. aol.com) and other major Internet players (e.g. army.mil). Among hosts, NetSizer also estimates the number of web,

mail, and Domain Name Service (DNS) servers and can distinguish between web servers with and without secure capabilities – this is of interest since, for the most part, electronic commerce transactions take place on servers with secure capabilities. NetSizer produces the number of Internet users and subscribers by geographic region.

NetSizer produces high level counts – for example, the number of hosts and the number of web servers worldwide – in real-time. In fact by observing displays of these "Internet growth clocks", you can actually watch the Internet grow in real-time! NetSizer produces most other measures (e.g., hosts by country) on a monthly basis. After analyzing NetSizer's methodology for estimating host counts by country, Telegeography determined that Telcordia's numbers are the most accurate available, and the recently published Telegeography 2000 reference book now uses NetSizer produced data for host counts by country and domain instead of their previous source. Other firms which produce Internet host counts typically do so once or twice per year whereas NetSizer produces some outputs in real time, and others daily and monthly. Results currently produced monthly could readily be produced more frequently if needed. NetSizer stores many of the results it produces making historical data available. NetSizer also forecasts future values for many of its statistics.

Somewhat more technically stated, NetSizer's host counts are estimates of the current number of permanent IP addresses in use in the public Internet. Hence hosts counts include network elements such as routers, servers, computer hosts in universities and companies, and ports in modem banks of ISPs. Host counts do not include home computers that are assigned temporary IP addresses at the time of accessing the Internet through an ISP nor hosts belonging to private IP networks and located behind firewalls.

## 3 How does NetSizer estimate host counts?

The estimated number of permanent IP addresses in use is based on a sample of IP addresses (out of the sample space of two to the 32nd power of possible IP addresses). To evaluate how many IP addresses are in use, a random sample of approximately 150,000 of these addresses is selected and searched for every day to determine how many are registered in the DNS. Because NetSizer collects and analyzes data all day every day, many estimates are provided on a real-time basis. NetSizer produces raw estimates with confidence intervals and accuracy is improved via stratification techniques. Forecasts are based on state-of-the-art diffusion models. NetSizer also includes methodology to account for hosts in a country even when the country domain (e.g., .jp for Japan, .ca for Canada) is not used by the host's owner.

## 4 Why is NetSizer useful?

Internet is growing so fast that survey-based estimates of the growth of the Internet provided yearly, or even quarterly, quickly become outdated. NetSizer overcomes this problem by providing monthly, daily and real-time estimates based on statistical techniques. Telcordia has used data provided by NetSizer in consulting proposals and engagements for ISPs, enterprises, and governments in the areas of strategic planning, competitive analysis, market planning, business case analysis, and IP network planning.

## 5 NetSizer's future directions

Telcordia's current research and development program for NetSizer includes developing capabilities to measure bandwidth and traffic on any network link (e.g., a T1 line) accessible in the public Internet. NetSizer's current bandwidth estimates are about 20 times faster than any other tools we are aware of, and we are not aware of any other tools capable of remotely measuring traffic between any two directly connected IP addresses on the public Internet.

Once these capabilities are fully researched, developed, and tested, they will have potential applicability for IP network planning and management for extranets, IP VPNs, as well as subnets of the public Internet. The new capabilities will also provide unique information to support the development of Telcordia (tm) Indicators for the Digital Economy.

## 6 Telcordia (tm) Indicators for the Digital Economy

The Internet is likely the fastest growing medium of commerce in the US economy and is quickly becoming so in many parts of the world. Traditional methods for measuring, tracking, and forecasting the information economy are unable to keep pace with the speed at which the Internet economy grows. Telcordia has a multidisciplinary team of statisticians, economists, Internet engineers, and computer scientists all working together to enhance NetSizer's capabilities and to use the unique data produced by these new capabilities, together with other data, to develop Telcordia (tm) Indicators for the Digital Economy. Emphasis will be placed on revenue generating electronic commerce which takes place on the public Internet. This will initially include domestic and international traffic generated by consumer-to-business transactions as well as business-to-business transactions on the public Internet. The focus is not on what individual consumers are doing, but on measuring electronic commerce and its impact at the individual firm, industry, and country level.

Telcordia plans to produce measures to monitor and track Internet electronic commerce; indices against which to measure performance improvements of a firm, industry or country; and indicators to predict performance

changes. The results planned to be produced should prove to be very valuable to the private sector – particularly the investment community – as well as government agencies such as statistical bureaus and commerce departments. Potential uses of these measures, indices, and indicators include:

- Tracking and measuring growth of Internet electronic commerce
- Measuring the impact of Internet electronic commerce on a country's overall economy
- Predicting economic changes based on Internet economic activity
- Growth projections for a firm, portfolio, or industry
- Strategic planning
- Competitive analysis

# Multivariate Kurtosis for Measuring Image Sharpness

Nien Fan Zhang[1], Michael T. Postek[1], Robert D. Larrabee[1] and Andras E. Vladar[1]

[1] National Institute of Standards & Technology, 100 Bureau Dr., Stop 8980, MD 20899-8980, USA

**Abstract:** In industry applications, such as automated on-line semiconductor production, users of scanning electron microscope (SEM) metrology instruments would like to have these instruments function without human intervention for long periods of time, and to have some simple criterion (or indication) of when they need servicing or other attention. At the present time, no self testing is incorporated into these instruments to verify that the instrument is performing at a satisfactory performance level. Therefore, there is a growing realization of the need for the development of a procedure for periodic performance testing. Postek and Vladar (1996) published a procedure which was based on the objective characterization of the spatial Fourier transform of the SEM image of a test object for this purpose and the development of appropriate analytical algorithms for characterizing sharpness. In this paper, an alternative approach based on the multivariate kurtosis is proposed to measure the sharpness of SEM image.

Scanning electron microscopes are being utilized extensively in the production environment. Since these instruments are approaching full automation, objective diagnostic procedures must be implemented to ensure data and measurement fidelity. One approach to this issue is the sharpness technique. It is known that the low-frequency changes in the video signal contain information about the large features and that the high-frequency changes carry information of finer details. When an SEM image has fine details at a given magnification, namely, when there are more high-frequency changes in it, we say it is sharper.

Since an SEM image is composed of a two-dimensional array of data, it can be expressed as $I(u,v)$, ($u = 1, ..., n$ and $v = 1, 2, ..., n$), where $u$ and $v$ are spatial indices. The corresponding two-dimensional finite Fourier transform is

$$f(x,y) = \sum_{u=1}^{n} \sum_{v=1}^{n} I(u,v) e^{i(xu+yv)} \tag{1}$$

where the spatial frequencies $x$ and $y$ have indices from $-N$ to $N$. Based on $f(x,y)$, we observe that when an SEM image is visually sharper than a second image, the high spatial frequency components of the first image are larger than those of the second. The following is an illustrative example. Part a of the attached figure shows the performance of a SEM on a heavy gold-coated oxide test sample at low accelerating voltage. This micrograph was taken following a tip change. This image appears to be far less sharp and lacking in resolution when it compared

with a similar micrograph (Part c of the figure) taken when the same instrument was operating more optimally. Parts b and d show the magnitude distribution of the two-dimensional spatial Fourier transform for the images in Parts a and c, respectively. From these figures it is clear that for the sharper image in Part c of the figure, its cone in Part d, which represents the magnitude distribution of the Fourier transform of the image, is wider than that in Part b of the image in Part a of the figure.

For a given univariate random variable $Z$ with mean $\mu_z$ and finite moments up to the fourth, the kurtosis is defined as

$$\beta_2 = \frac{m_4}{m_2^2}$$

where $m_4$ and $m_2$ are the fourth and second central moments respectively, that is

$$m_4 = E[(Z - \mu_z)^4] \text{ and } m_2 = \sigma_z^2 = E[(Z - \mu_z)^2].$$

For any univariate normal distribution, $\beta_2 = 3$. Therefore the value of $\beta_2$ can be compared with 3 to determine whether the distribution is "peaked" or "flated-topped" relative to a Gaussian. Note that kurtosis is a dimensionless ratio.

Four separate distribution density functions with zero mean and unit variance were compared by Kaplansky (1945) to illustrate the properties of kurtosis. His results show that the smaller the kurtosis, the flatter the top of the distribution. Finucan (1964) also discussed the interpretations of kurtosis. The conclusion is that the distribution with smaller kurtosis is more flat-topped or has a large shoulder than that with larger kurtosis.

Note that, since kurtosis is a dimensionless ratio of the moments, a value of kurtosis can be calculated for any positive function when the area underneath the function curve is finite and when the curve has finite moments up to the fourth. Let $w = f(z)$ be such a discrete univariate function. Then

$$\beta_2 = \frac{\sum_{j=1}^{n} (z_j - \mu_z)^4 f(z_j)}{[\sum_{j=1}^{n} (z_j - \mu_z)^2 f(z_j)]^2}$$

where $\mu_z = \sum_{j=1}^{n} z_j f(z_j)$.

Multivariate kurtosis has been proposed by Mardia (1970). Let $W$ be a $p$-dimensional random vector with finite moments up to the fourth. Let $\mu$ be the mean vector and $\Gamma$ be the covariance matrix of $W$. The kurtosis of $W$ is defined by

$$\beta_{2,p} = E[(W - \mu)^T \Gamma^{-1} (W - \mu)]^2,$$

where $T$ denotes the transpose of a vector. When $p = 1$, $\beta_{2,1}$ becomes the univariate kurtosis $\beta_2$. When $p = 2$, for a two-dimensional random vector $W = (X, Y)^T$, the two-dimensional kurtosis is given by

$$\beta_{2,2} = \frac{\gamma_{4,0} + \gamma_{0,4} + 2\gamma_{2,2} + 4\rho(\rho\gamma_{2,2} - \gamma_{1,3} - \gamma_{3,1})}{(1 - \rho^2)^2} \tag{2}$$

where $\gamma_{k,l}$, $(k, l = 0, 1, 2, 3, 4)$ is

$$\gamma_{k,l} = \frac{E[(X - \mu_x)^k (Y - \mu_y)^l]}{\sigma_x^k \sigma_y^l} \tag{3}$$

and $\mu_x$ and $\mu_y$ are the marginal means and $\sigma_x$ and $\sigma_y$ are the marginal standard deviations and $\rho$ is the correlation between $X$ and $Y$. In particular, when a two-dimensional random vector $W$ has a discrete probability distribution $f(x_j, y_k)$, $j = 0, 1, ..., n$ and $k = 0, 1, .., m$, the two dimensional kurtosis can be calculated similarly with

$$\gamma_{k,l} = \frac{\sum_{j=0}^{n} \sum_{k=0}^{m} f(x_j, y_k)(x_j - \mu_x)^k (y_k - \mu_y)^l}{\sigma_x^k \sigma_y^l}. \tag{4}$$

From (3) or (4), the marginal kurtoses of the marginal distribution of $X$ and $Y$ are

$$\beta_{2,x} = \gamma_{4,0} \text{ and } \beta_{2,y} = \gamma_{0,4} \tag{5}$$

From Priestley (1981), the normalized spectral density with zeros frequency at the center can be treated as a probability density function. The kurtosis of the spectral density estimate corresponding to (1) of an SEM image can be calculated by (2) and (3) or (4). A sharper SEM image corresponds to a spectrum which has a large shoulder or has a flatter shape. Thus, it can be concluded that the corresponding kurtosis of the sharper image is smaller. Therefore, an increase in kurtosis over some preestablished reference portends that the sharpness of an SEM image has been degraded relative to the existing at the time the reference value was established.

The marginal kurtoses defined in (5) are used to measure the shapes of the marginal distributions. The difference between the marginal kurtoses is a measure of the shape difference between the marginal distributions. It can be used to detect possible instrument vibration. To show how the kurtosis can be applied to the SEM images, a series of five micrographs were taken on the same instrument with only one parameter changed for each image. Two-dimensional kurtoses were calculated for the five micrographs. The result shows the kurtosis measure gives the correct ranking of the sharpness of the images.

**Keywords:** Image analysis, kurtosis, discrete Fourier transform, spatial frequency.

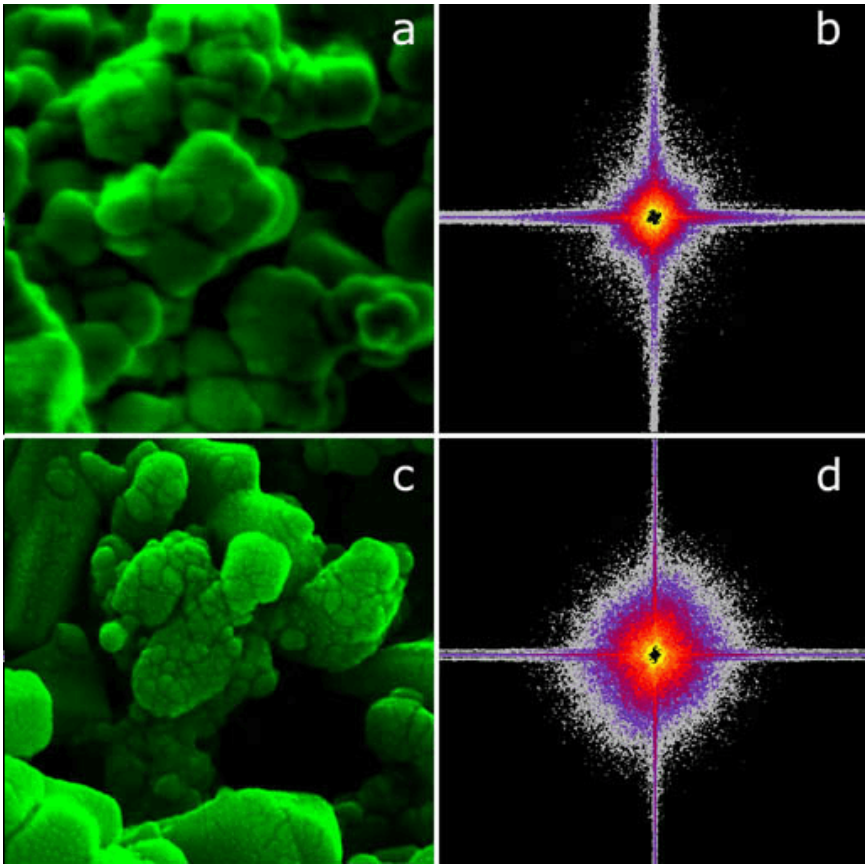



FIGURE 1. Scanning electron micrographs and their two-dimensional Fourier frequency magnitude distributions. (a) Scanning electron micrograph of heavily coated oxide taken following a tip change. (b) Two-dimensional Fourier frequency magnitude distributions of image (a). (c) Scanning electron micrograph of the same sample as above, but taken when the instrument is functioning at a high level of performance. (d) Two-dimensional Fourier frequency magnitude distributions of image (c). Note that there are more high frequency elements present in the Fourier frequency magnitude distribution from (c)