SMS

Statistical Modelling Society

Description of Short Courses

- The Craft of Smoothing (Eilers & Marx)

- An Introduction to Bayesian Statistics with Applications in Biostatistics and Epidemiology (Lesaffre)

- Longitudinal and Incomplete Data (Verbeke & Molenberghs)

(Version: January 2010)

The Craft of Smoothing

Paul Eilers
`p.eilers@erasmusmc.nl`

Brian Marx
`bmarx@lsu.edu`

In the "The Craft of Smoothing," we describe in detail the basics and use of P-splines, a combination of regression on a B-spline basis and difference penalties (on the B-spline coefficients).

Our approach is practical. We see smoothing as an everyday tool for data analysis and statistics. We emphasize the use of modern software and we provide functions for R/S-Plus and Matlab.

There will be eight sessions:

- Session 1 presents the idea of bases for regression. It will show why global bases, like power functions or orthogonal polynomials are ineffective and why local bases (gaussian bell- shaped curves or B-splines) are attractive.

- In Session 2 penalties are introduced, as a tool to give complete and easy control over smoothness. The combination of B-splines and difference penalties will be studied for smoothing, interpolation and extrapolation.

- In the first two sessions the data are assumed to be normally distributed around a smooth curve. In session 3 we extend P-splines to non-normal data, like counts or a binomial response. The penalized regression framework makes it straightforward to transplant most ideas from generalized linear models to P-spline smoothing. Important applications are density estimation and variance smoothing.

- Any smoothing method has to balance fidelity to the data and smoothness of the fitted curve. An optimal balance can be found by cross-validation or AIC. This subject is studied in Session 4, as well as the computation of error bands of an estimated curve. We also show how optimal smoothing performs on simulated data, to give you confidence in that it makes the right choices.

- Session 5 places P-splines in a wider perspective. It presents Bayesian and mixed model interpretations of P-splines. Special attention is being paid to streamlined computation

- In the first five sessions we only consider one-dimensional smoothing. When there are multiple explanatory variables, we can use generalized additive models, varying-coefficient models, or combinations of them. Tensor products of B-splines and multi-dimensional difference penalties make an excellent tool for smoothing in two (or more) dimensions. This is the subject of session 6.

- The difference penalty is sufficient to control smoothness in most standard situations. On other occasions we may want to push a curve fit into special directions, because the data are (quasi-)periodic or the shape of the fitted curve has to be monotone or convex. Session 7 shows how specialized penalties can be designed and used.

- The final session looks at the use of P-splines in regression problems with very many variables, which are ordered, like in optical spectra. In the chemometric literature this is known as multivariate calibration.

In addition there are two computer lab sessions, in which R software will be used to solve a number of smoothing problems. One session will concentrate on simple functions with limited goals. This will improve your understanding of what is going on "under the hood". The other session will use the `mgcv` package, written by Simon Wood, a large but powerful tool that can handle a variety of situations.

We very much hope that you will find this course useful, interesting, and enjoyable.

An Introduction to Bayesian Statistics
with Applications in Biostatistics
and Epidemiology

Presenter:

Emmanuel Lesaffre

Katholieke Universiteit Leuven and
Erasmus University Medical Center

email: `emmanuel.lesaffre@med.kuleuven.be`

Abstract:

Bayesian statistics has received a great deal of attention as a method to tackle complex statistical analyses. The purpose of this course is to gradually and smoothly introduce the participants into the Bayesian philosophy and terminology. The early examples are simple but inspired by clinical trial examples and simple epidemiological studies. To this end the menu-driven software FirstBayes will be employed. While in general the technical level of the course is low, the different numerical techniques to calculate (or to sample from) the posterior distribution will be treated. The recent Markov Chain Monte Carlo (MCMC) techniques will be explored quite extensively thereby trying to keep the technical level low. WINBUGS will be used to illustrate the power of the MCMC methods with a variety of biostatistical and epidemiological applications. For instance we will consider the use of Bayesian methods in disease mapping, meta-analyses, survival analysis, errors-in-variables problems, missing data problems, clinical trials such as the use of Bayesian methods for the planning of trials, in compliance problems, data and safety monitoring boards, etc.

Outline:

Day 1 morning
      Frequentist and likelihood approach in contrast to Bayesian philosophy
      Bayes theorem and Bayesian summary statistics: binomial likelihood
Day 1 afternoon
      Bayes theorem and Bayesian summary statistics: normal likelihood

Day 2 morning
      Tutorial with FirstBayes
Day 2 afternoon
      Sampling from the posterior distribution with applications in R (or S+)
      Choosing the prior distribution

Day 3 morning
      Towards real life problems: more than 1 parameter involved
      Bayesian regression analysis: sampling the posterior distribution using R (or S+)
Day 3 afternoon
      Simple hierarchical models: Bayesian meta-analyses, disease mapping

Day 4 morning
   Simple hierarchical models: Bayesian meta-analyses, disease mapping (continued)
   Markov Chain Monte Carlo methods: Gibbs sampling
Day 4 afternoon
   First examples with WINBUGS: Bayesian meta-analyses, disease mapping, the
   use of historical data in an epidemiological study

Day 5 morning
   Markov Chain Monte Carlo methods: Metropolis-Hastings method
   Markov Chain Monte Carlo methods: Gibbs sampling
Day 5 afternoon
   More advanced analyses with WINBUGS: correction of scoring errors in a dental
   epidemiological study; repeated measurements studies; the estimation of prevalence
   using probabilistic constraints; survival analyses; etc.

Longitudinal and Incomplete Data

Geert Verbeke

I-BioStat
Katholieke Universiteit Leuven
Kapucijnenvoer 35
B-3000 Leuven
Belgium
Email: geert.verbeke@med.kuleuven.ac.be
Tel: +32-16-336891
Sec: +32-16-336892,
Fax: 32-16-337015
Web: http://www.kuleuven.ac.be/biostat

Geert Molenberghs

I-BioStat
Universiteit Hasselt & Katholieke Universiteit Leuven
Agoralaan 1
B-3590 Diepenbeek
Belgium
Email: geert.molenberghs@uhasselt.be
Tel: +32-11-268238
Sec : +32-11-268202
Fax : +32-11-268299
Web : http://www.uhasselt.be/censtat

Abstract

Based on Verbeke and Molenberghs (Springer, 1997, 2000, 2005), a general introduction to longitudinal data and the linear mixed model for continuous responses will be pre-sented. The topic will be approached from the modeller's and practitioner's points of view. Emphasis will be on model formulation, parameter estimation, and hypothesis testing, as well as on the distinction between the random-effects (hierarchical) model and the implied marginal model. Illustrations will be given based on the SAS procedure MIXED.

When the response of interest is categorical, the linear mixed model concepts can be extended towards generalized linear mixed models. An alternative approach is the use of generalized estimating equations (GEE). A lot of emphasis will be put on the fact that the regression parameters in both types of models have different interpretations. Advantages and disadvantages of both procedures will be discussed and compared in detail, and illustrations will be based on the SAS procedures GENMOD, GLIMMIX, and NLMIXED.

Finally, when analysing longitudinal data, one is often confronted with missing observations, i.e., scheduled measurements have not been made, due to a variety of (known or unknown) reasons. It will be shown that, if no appropriate measures are

taken, missing data can cause seriously biased results, and interpretational difficulties.

Throughout the course, it will be assumed that the participants are familiar with basic statistical modelling, including linear models (regression and analysis of variance), as well as generalized linear models (logistic and Poisson regression). Moreover, prerequisite knowledge should also include general estimation and testing theory (maximum likelihood, likelihood ratio).

Format: Several options are possible:
- One day, only theory
- Two days, only theory
- Two days, with practical sessions discussing outputs of SAS analyses
- Two days, with practical sessions in PC lab
- More than two days, contents to be discussed.
-

Targeted audience: Applied statisticians and biomedical researchers in industry, public health organizations, contract research organizations, and academia.

Learning outcomes

As a result of the course, participants should be able to perform a basic analysis for a particular longitudinal data set at hand. Based on a selection of exploratory tools, the nature of the data, and the research questions to be answered in the analyses, they should be able to construct an appropriate statistical model, to fot the model within the SAS framework, and to interpret the obtained results. Further, participants should be aware not only of the possibilities and strengths of a particular selected approach, but also of its drawbacks in comparison to other methods.
The course will be explanatory rather than mathematically rigorous. Emphasis is on giving sufficient detail in order for participants to have a general overview of frequently used approaches, with their advantages and disadvantages, while giving reference to other sources where more detailed information is available. Also, it will be explained in detail how the different approaches can be implemented in the SAS package, and how the resulting outputs should be interpreted.

Presenters

Geert Verbeke is Professor in Biostatistics at the Biostatistical Centre of the Katholieke Universiteit Leuven in Belgium. He received the B.S. degree in mathematics (1989) from the Katholieke Universiteit Leuven, the M.S. in biostatistics (1992) from the Limburgs Universitair Centrum, and earned a Ph.D. in biostatistics (1995) from the Katholieke Universiteit Leuven. Geert Verbeke has published extensively on longitudinal data analyses. He has held visiting positions at the Gerontology Research Center and the Johns Hopkins University (Baltimore, MD). Geert Verbeke is Past President of the Belgian Region of the International Biometric Society, International Program Chair for the International Biometric Conference in Montreal (2006), Board Member of the American Statistical Association. He is past Joint Editor of the Journal of the Royal Statistical Society, Series A (2005--2008). He is the director of the Leuven Center for Biostatistics and statistical Bioinformatics (L-BioStat), and vice-director of the Interuniversity Institute for Biostatistics and

statistical Bioinformatics (I-BioStat), a joint initiative of the Hasselt and Leuven universities in Belgium.

Geert Molenberghs is Professor of Biostatistics at Universiteit Hasselt and Katholieke Universiteit Leuven in Belgium. He received the B.S. degree in mathematics (1988) and a Ph.D. in biostatistics (1993) from Universiteit Antwerpen. He published on surrogate markers in clinical trials, and on categorical, longitudinal, and incomplete data. He was Joint Editor of Applied Statistics (2001-2004) and Co-Editor of Biometrics (2007-2009). He was President of the International Biometric Society (2004-2005), received the Guy Medal in Bronze from the Royal Statistical Society and the Myrto Lefkopoulou award from the Harvard School of Public Health. Geert Molenberghs is founding director of the Center for Statistics. He is also the director of the Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat).

Both authors wrote several books on the use of linear mixed models for the analysis of longitudinal and incomplete data and taught numerous short and longer courses on the topic in universities as well as industry, in Europe, North America, Latin America, and Australia. They repeatedly received the American Statistical Association's Excellence in Continuing Education Award (2002, 2004, 2005, 2008). Both of them are elected Fellow of the American Statistical Association and elected member of the International Statistical Institute.

Course Materials

Copies of the transparencies used in the course

Textbook: Verbeke G. & Molenberghs G. (2000) *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York, 568 pages, ISBN 0-387-95027-3.

Textbook: Molenberghs G. and Verbeke G. (2005) `Models for discrete longitudinal data,' Springer Series in Statistics, Springer-Verlag, New-York, 683 pages.
ISBN 0-387-25144-8.