Statistical Modelling 2018; **18(3–4)**: 199–202

Editorial 'Bridging the gap between methodology and applications: Tutorials on semiparametric regression'

Statistical modelling, by definition, aims at simplifying the true relationships between variables in order to make unknown complex systems accessible for empirical research. Or as George Box formulated in his famous aphorism: 'All models are wrong, but some are useful'. This special issue aims to achieve something similar on a different level: to simplify the presentation of the methodology behind semiparametric modelling, in order to make it accessible for applied researchers. In semiparametric regression models, the usual linear predictor is replaced by an additive composition of different types of effects where the functional form of the covariate effects is to be determined from the data rather than via structural specifications by the analyst. This enables considerably more freedom in terms of model specification and arguably reduces the risk of misspecifying important relationships in the data to be analysed.

Semiparametric regression specifications have been the subject of considerable interest among statisticians for several years, in particular since 1990 when Hastie and Tibshirani published their classical textbook on generalized additive models. During the workshop 'The Future of Semiparametric Regression' organized by the Research Training Group 1644 'Scaling Problems in Statistics' (funded by the German Research Foundation, DFG) held on 28–29 September 2016 in Göttingen, many extensions and future directions of flexible statistical modelling were presented and discussed. As a key challenge for the future, the participants of this workshop identified the acceptance of semiparametric regression in applied research areas. Many approaches that are already considered as standard tools in the statistical community appear only occasionally in actual applications published in subject-matter journals. Still the feeling of the workshop participants was that semiparametric regression techniques should become more popular in applied research and their wider application has the potential to enable new and interesting insights.

As a consequence, *Statistical Modelling* agreed to publish a special issue with tutorial papers that provide an easy entry point for researchers that are concerned with practical issues and are not necessarily interested in the complete underlying statistical methodology. Ultimately, our goal is to foster the application of semiparametric regression. Accordingly, the contributors to this special issue were encouraged to abide the following principles when setting up their tutorial papers:

10.1177/1471082X18761252

- Accessibility: All tutorials should be accessible for applied researchers who are not familiar with semiparametric regression to allow for a widespread appreciation of the methods presented in the tutorial.
- Intuition over mathematical rigour: Technical details and formulae should be avoided wherever possible. The tutorials should rather focus on introducing the main modelling ideas in an intuitive and applied manner.
- Practical relevance: The tutorials should provide clear evidence on the practical relevance of the models/methods discussed. This includes a thorough discussion of the advantages and limitations of the presented models/methods compared to standard approaches and alternatives. Furthermore, the tutorials should emphasize practical aspects such as interpretability, model tuning and evaluation.
- Reproducibility: The tutorials should incorporate a reproducible worked-out application with code and data provided as electronic supplements. The tutorials themselves should in general not serve as introductions to specific software packages but to the general approach under discussion.

This special issue collects eight contributions that present quite different yet related approaches for semiparametric regression modelling.

In her article, Waldmann (2018) motivates how and why to apply quantile regression, and illustrates it with an analysis of data on stunting scores for childhood malnutrition in India. The basic idea of this approach is to model the influence of explanatory variables not on the expected value of an assumed distribution but directly on the quantiles of interest. An obvious advantage of this procedure is of course the lack of distributional assumptions but also the robustness inherited from the definition of quantiles. For the illustrative analyses included in the tutorial paper, Waldmann chose a gradient boosting approach (which is in itself the topic of another contribution in this special issue by Mayr and Hofner) to estimate the different models. While quantile regression can be applied without the specification of a distribution, a disadvantage is that the different quantile models are estimated separately, making it possible that quantile curves might cross.

In the model class of Generalized Additive Models for Location Scale and Shape (GAMLSS, Stasinopoulos et al., 2018), the idea is to specify a parametric conditional distribution and to model not only the location (e.g., the mean) but all corresponding distributional parameters (e.g., scale and shape parameters). These models are often used to estimate centile curves for growth charts; due to the underlying distribution, those quantiles derived from a parametric distribution cannot cross, however, they might heavily depend on the correctness of the assumed distribution. The authors illustrate their tutorial with the construction of such centile curves for a continuous response and the analysis of discrete count data. They apply the R package `gamlss`, giving detailed advice regarding what choices and decisions researcher face in practice to find a suitable model.

The tutorial article by Umlauf and Kneib (2018) basically focuses on the same GAMLSS model class—which is also often referred to as distributional regression—but rather from a Bayesian perspective. They present an extremely

flexible modular implementation, which can in practice be also a valuable tool for simpler model classes like classical regression of the mean. The presented R add-on package `bamlss` allows to incorporate different covariate effects, for example non-linear, spatial and random effects in a straightforward manner for various model classes. The authors illustrate how to select models and how to monitor convergence of Markov Chains based on the analysis of large weather datasets from Germany, modelling the conditional distribution of the daily temperature at the country's highest mountain and extreme precipitation events across different regionally distributed weather stations.

Hothorn (2018) illustrates how transformation analysis can be exploited as an alternative approach to model response distributions. Instead of adding complexity to simple models, he uses stepwise complexity reduction to help identify simpler and better-interpretable models. As an example, body mass index distributions in Switzerland are modelled by means of transformation models to understand the impact of sex, age, smoking and other lifestyle factors on a person's body mass index. In doing so, a compromise between model fit and model interpretability is found. The models investigated range from classical linear regression models to novel models with flexible conditional distribution functions, such as transformation trees and forests.

Another area addressed by two contributions is time-to-event data. Bender et al. (2018) demonstrate how continuous time-to-event data can be flexibly modelled by taking advantage of advanced inference methods recently developed for generalized additive mixed models. They describe necessary data pre-processing steps and show how a variety of effects, including a smooth non-linear baseline hazard and potentially non-linear and non-linearly time-varying effects can be estimated and interpreted. Furthermore, graphical tools for model evaluation and interpretation of the estimated effects are provided. The tutorial is accompanied by an `R package` called `pammtools` implementing these tools.

Berger and Schmid (2018) on the contrary focus on the analysis of time-to-event outcomes that are either intrinsically discrete or grouped versions of continuous event times. As a variety of regression methods exists in the literature for such data, their tutorial aims to provide an introduction as to how these models can be applied using open source statistical software. In particular, they consider semiparametric extensions comprising the use of smooth non-linear functions and tree-based methods. The methods are illustrated by data on the duration of unemployment of US citizens.

Function-on-scalar regression models are considered in Bauer et al. (2018). While the covariates are scalars, these models feature different types of functions as the response, such as collections of time series, as well as 2D or 3D images. A hands-on introduction for a flexible semiparametric approach for function-on-scalar regression is provided, using spatially referenced time series of ground velocity measurements from large-scale simulated earthquake data as a running example. Furthermore, important practical considerations and challenges in the modelling process are discussed and the approach is also complemented by comprehensive R code.

Our last contribution is devoted to boosting, a technique that was originally developed for machine learning but was later adapted to estimate statistical models—and offers various practical advantages such as automated variable selection and implicit regularization of effect estimates. Mayr and Hofner (2018) highlight how boosting can be used for semiparametric modelling, what practical implications follow from the design of the algorithm and what kind of drawbacks data analysts should expect. They illustrate the application of boosting via the R add-on package `mboost` with the development of a biomarker for the occurrence of metastases in breast cancer patients based on a high-dimensional dataset of tumour-DNA and in the analysis of a stunting score from children in India.

In summary, the tutorial articles in this special issue provide a step forward towards bridging the gap between methodology and application: We hope that you find that not only do they present a good overview on the current state of the art in semiparametric modelling but also provide an easy entry into this fascinating world, one where almost any model is possible.

Andreas Groll

Thomas Kneib

Andreas Mayr