Statistical Modelling 2013; **13**(**4**): 275–303

Beyond mean regression

Thomas Kneib
Chair of Statistics, Georg August University, Göttingen, Germany

Abstract: Usual exponential family regression models focus on only one designated quantity of the response distribution, namely the mean. While this entails easy interpretation of the estimated regression effects, it may often lead to incomplete analyses when more complex relationships are indeed present and also bears the risk of false conclusions about the significance/importance of covariates. We will therefore give an overview on extended types of regression models that allows us to go beyond mean regression. More specifically, we will consider generalized additive models for location, scale and shape as well as semiparametric quantile and expectile regression. We will review the basic properties of all three approaches and compare them with respect to the flexibility in terms of the supported types of predictor specification, the availability of software and the support for different types of inferential procedures. The considered model classes are illustrated using a data set on rents for flats in the City of Munich.

Key words: expectile regression; generalized additive models for location, scale and shape; quantile regression; semiparametric regression

1 Introduction

A common (mis-)perception of statistics in the public opinion equates statistics with means and averages, which has led to the famous quote that 'statisticians are mean lovers' (Friedman *et al*., 2002). While most of us would probably argue that statistics offers much more than means and that the analysis of any data set should usually also comprise the calculation of other summary statistics such as variances, quantiles, ranges, etc., the mean is still omnipresent in statistical analyses. This applies in particular for regression modelling where the framework of generalized linear models has led to a rather high popularity of models relating a regression predictor to the mean of a response via a suitably chosen link function. While considerable work has been done in recent years on extending regression models towards more flexible specifications of the predictor with generalized additive models (Hastie and Tibshirani, 1990; Ruppert *et al*., 2003; Wood, 2006) being the most prominent example, the combination of such flexible, semiparametric predictors with regression

Address for correspondence: Thomas Kneib, Chair of Statistics, Georg August University Göttingen, Platzer der Göttinger Sieben 5, 37073 Göttingen, Germany. E-mail: tkneib@uni-goettingen.de

 10.1177/1471082X13494159

models that go beyond the mean is still challenging. We will therefore review the current state of the art of semiparametric regression beyond the mean and try to guide the reader towards suitable model classes and to provide information on the flexibility of the different model specifications, supported inferential principles and the availability of software for actually fitting such regression models.

To set the scene, consider a regression situation with n observations $(y_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, on a continuous response variable y and covariates $\mathbf{z}$. Then a typical regression model for the mean takes the form

$$y_i = \eta_i + \varepsilon_i, \qquad E(\varepsilon_i) = 0, \qquad \text{Var}(\varepsilon_i) = \sigma^2,$$

with regression predictor η_i specified in dependence of the covariate $\mathbf{z}$ (we will discuss specific choices later). The two assumptions on the error term imply that, on the one hand, the predictor describes the expectation of the response since

$$E(y_i) = \eta_i + E(\varepsilon_i) = \eta_i$$

and, on the other hand, that ordinary least squares estimation can be used due to the homoscedasticity of the errors. Often, the error term will additionally be assumed to follow a normal distribution such that ε_i i.i.d. $N(0, \sigma^2)$ to facilitate inference about the parameters contained in the regression predictor. In this case, the mean regression model does not only imply that the variance does not depend on covariates but also implies that all other distributional characteristics (such as the skewness or the kurtosis) are equal for all observations.

Of course, mean regression models have the advantage of being easy to understand and estimate and to entail easy interpretation of the regression effects contained in the predictor since changes in the covariate values only induce changes in the expected value of the response. However, it is often also too restrictive due to the strong assumptions on the error term. For example, in case of heteroscedasticity also the variance (or the standard deviation) of the response may depend on covariates. This can (at least conceptually) easily be incorporated in the model formulation by modifying the regression equation to the location-scale model

$$y_i = \eta_{i1} + \exp(\eta_{i2})\varepsilon_i, \qquad E(\varepsilon_i) = 0, \qquad \text{Var}(\varepsilon_i) = 1, \tag{1.1}$$

with two predictor structures η_{i1} for the mean and η_{i2} for the standard deviation, respectively, such that

$$E(y_i) = \eta_{i1}, \qquad \text{Var}(y_i) = \exp(\eta_{i2})^2.$$

When the errors additionally follow a normal distribution, this is the simplest example of a generalized additive model for location, scale and shape (GAMLSS, in fact without an effect on the shape) as introduced by Rigby and Stasinopoulos (2005) as a comprehensive class of models where different parameters of the response distribution are related to regression predictors. GAMLSS rely on flexible regression specifications where a predictor is formulated not only for the mean of the response but also for further parameters of the response distribution. This has the advantage that a parametric distribution is kept for the response such that interpretation

remains feasible (at least for not too complex types of distributions) and maximum likelihood estimation can still be used. We will introduce GAMLSS in more detail in Section 4.

While GAMLSS retain the assumption of a parametric distribution for the responses (or equivalently the error terms), it may also be useful to completely drop this assumption and to formulate nonparametric models that still allow us to describe more than the mean of the response. This may in particular be the case if interest is not on identifying covariate effects on specific parameters of the response distribution but on the relation of 'extreme' observations in the tails of the distribution on covariates. This is enabled in quantile and expectile regression models where we go back to the initial model formulation

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau}, \tag{1.2}$$

but modify the assumptions on the error terms appropriately to model observations in the tails as denoted by the asymmetry parameter $\tau \in (0, 1)$ that specifies the desired 'extremeness' and is therefore added to both the predictor and the error terms as a subscript. In quantile regression (as originally proposed in Koenker and Bassett 1978 and comprehensively described in Koenker 2005), we assume that the τ-quantile of the error term is zero, i.e., $F_{\varepsilon_{i\tau}}(0) = \tau$, where $F_{\varepsilon_{i\tau}}(\cdot)$ denotes the cumulative distribution function of the ith error term. This assumption implies that the predictor $\eta_{i\tau}$ specifies the τ-quantile of y_i and, as a consequence, the regression effects can be interpreted on the quantiles of the response distribution. Estimation results for a dense set of quantiles then also allow us to characterize the complete distribution of the responses in terms of covariates, see Section 5 for details. An alternative to quantile regression is expectile regression where basically the assumptions on quantiles are replaced with expectiles which provide an alternative way of describing the tails of distributions in terms of a generalization of the mean instead of a generalization of the median as in quantile regression. We will introduce expectiles in more detail in Section 6.

Note that in fact any regression model relying on an explicit distributional assumption for the responses also implicitly defines a quantile regression model. For example, in case of a simple mean regression model with homoscedastic normal errors, the τ-quantile of response y_i is given by $\eta_i + \sigma z_\tau$, where z_τ denotes the τ-quantile of the standard normal distribution, and as a consequence the simple mean regression structure implies parallel quantile curves. This can be overcome in the location-scale model (1.1), where the quantiles are determined as $\eta_{i1} + \exp(\eta_{i2})z_\tau$, but flexibility is still limited as compared to 'real' quantile regression models.

To illustrate some of the basic concepts introduced so far, we analyse data from the Munich rental guide 1999 where the response variable of interest is the net rent paid for a specific flat. In addition to point predictions corresponding to expected or average rents for flats with specific characteristics, interval boundaries for flats comprising, for example, two-thirds of usual rents with given characteristics are of high relevance to guide tenants and landlords. In the following, we consider only very simple models where either the size of the flat ('living area') or the year of

construction are treated as covariates. For the former, we use a linear model specification while a quadratic polynomial is used for the latter. A simple, homoscedastic model yields the estimated effects visualized in the top row of Figure 1. Each of the lines corresponds to a quantile curve from the set of quantiles specified by $\tau = 0.01, 0.1, 0.2, \ldots, 0.8, 0.9, 0.99$. As discussed above, the resulting quantile curves are all parallel since homoscedasticity is assumed. Especially for the effect of living area, this assumption seems highly questionable and the estimated quantile curves do not fit well with the visual impression of the distribution of the data, at least for small and large values of the living area. For the year of construction, the effect is less drastic but there seems to be some skewness in the error distribution that is not reflected adequately.

As a second try, we consider the location-scale model (1.1) where again only a single covariate is included in linear (living area) or quadratic (year of construction) functional form. The results are shown in the second row of Figure 1 and indicate an improved fit to the data especially for the living area where heteroscedasticity is captured satisfactorily. Note that the estimated quantile curves are now nonlinear despite the linear predictors specified for the living area since the predicted quantiles are given by $\eta_{i1} + \exp(\eta_{i2})z_\tau$, where the predictor for the standard deviation enters nonlinearly. While the location-scale model seems to fit the heteroscedasticity for living area, it still does not capture the skewness for the year of construction since the normal distribution assumed for the errors is symmetric.

Finally, the third row of Figure 1 shows estimates for quantile regression model (1.2). Now the estimated quantile curves are again linear for the living area since the (linear) predictor acts directly on the quantile of interest. Obviously, quantile regression allows to fit both heteroscedasticity and skewness satisfactorily in this example and would therefore probably be the model of choice here. However, an extended model including a further predictor for the skewness (utilizing for example a Box-Cox power model or a power exponential model) may also yield a satisfactory fit.

While this simple example already illustrates some of the advantages of going beyond mean regression, there are several further areas of application where similar model types are needed. Some examples, we have worked on or are currently working on include

- childhood malnutrition in developing countries, where the impact of covariates on extreme forms of malnutrition is of higher relevance than models for the average nutritional status (Fenske *et al.*, 2011).
- efficiency estimation in agricultural production, where we are particularly interested in covariates impacting above-average performance of farms.
- modelling gas flow networks, where the behaviour of the network in high demand situations shall be studied.

In fact, once starting to think about regression models beyond the mean, they seem to appear basically everywhere and it seems more and more questionable to restrict attention solely to the mean of a response.

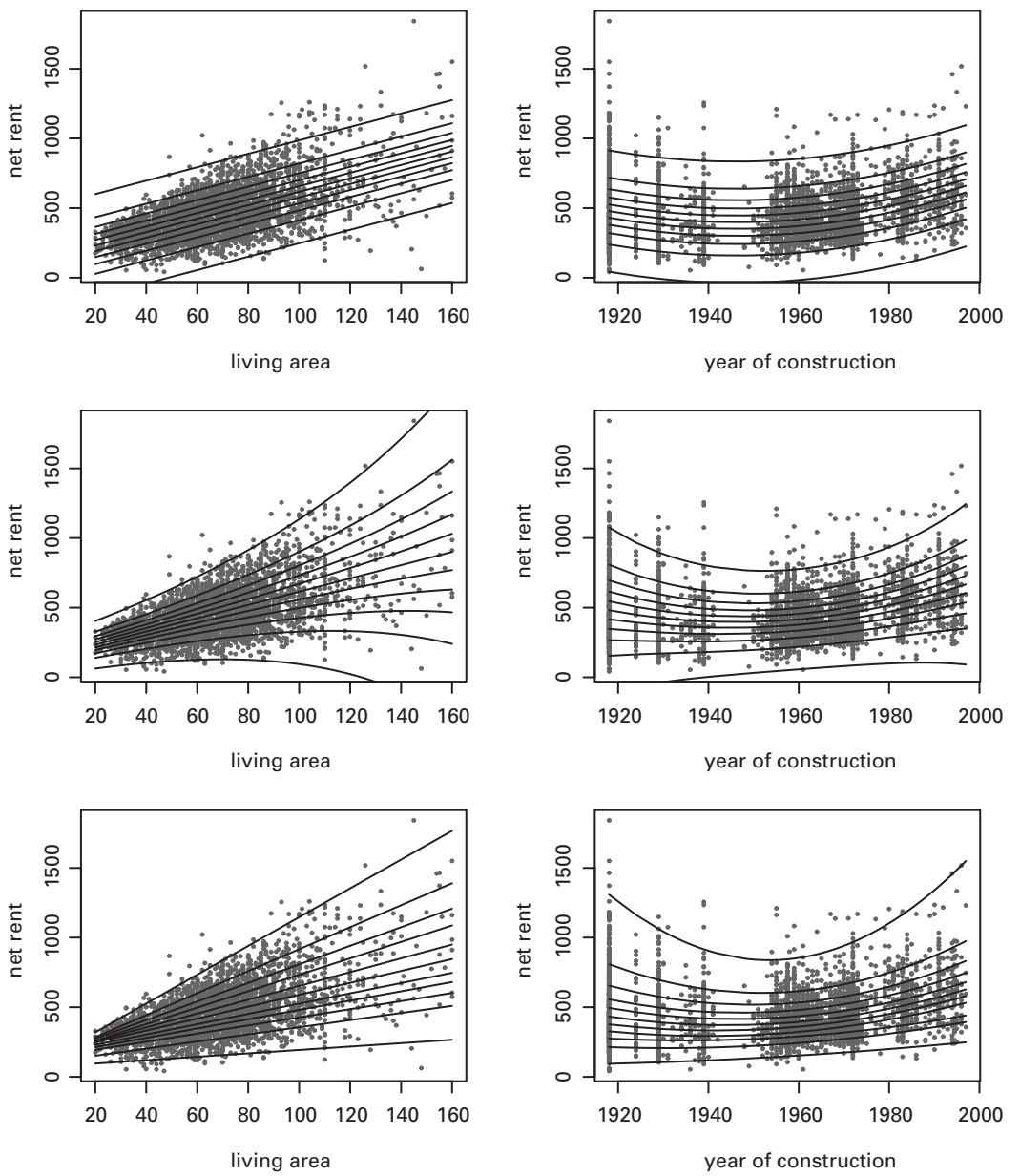

Figure 1 Parametric models for the net rent in the Munich rental guide example. The top panel shows estimates from a homoscedastic normal model, the second row estimates from a heteroscedastic location-scale normal model and the third row estimates from quantile regression

In all of the specified examples (and in fact also for the Munich rental guide), simple linear regression specifications are not sufficient, but the data call for more general, semiparametric predictor specifications similar in spirit to generalized additive models with predictor

$$\eta_i = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}),$$

where β_0 is an intercept while $f_1, \ldots, f_p$ are smooth functions of the continuous covariates $x_1, \ldots, x_p$. In this review, we will consider an even broader model class provided by structured additive regression (Fahrmeir *et al.*, 2004; Kneib *et al.*, 2009) that allows us to additionally include spatial effects, varying coefficient terms, interaction surfaces, random effects and a number of further extensions. Such complex predictor structures are nowadays often required in applied regression modelling to adequately reflect the complexity of the data collected. However, the combination of flexible predictor structures and regression models beyond the mean is still challenging and not all desired combinations are yet supported.

The main aims of this paper can now be summarized as follows:

- Provide a brief introduction to the basic modelling concepts of GAMLSS, quantile and expectile regression.
- Review semiparametric regression specifications and how they fit into the different modelling concepts.
- Discuss pros and cons of different inferential procedures and their suitability for the different model classes.
- Discuss advantages and disadvantages of the model classes and provide guidance for first time users of these models also in terms of software availability.

The view taken in this paper is notoriously subjective and not everyone will agree with the pros and cons discussed for the different methods. This is in fact intended and I am the first to admit that some of the statements may be debatable. Anyway, I consider them to be worth this debate and would be happy if this paper helps in stimulating discussion about semiparametric regression models beyond the mean.

In the remainder of this paper, we will first introduce the class of semiparametric predictor structures we are interested in (Section 2), followed by a discussion of inferential procedures that could be used for fitting semiparametric regression models (Section 3). Sections 4, 5 and 6 then contain material on GAMLSS, quantile regression and expectile regression. Each section will introduce the basic concepts associated with the model specification, the suitability of the previously introduced inferential approaches and the advantages and disadvantages of the model class. Section 7 will provide some summarizing comments as well as directions towards missing material for future work.

2 Semiparametric regression models

In the following, we will describe a generic structure for semiparametric regression predictors as a general framework for any of the predictor components arising in GAMLSS, quantile regression or expectile regression. For simplicity, we will drop any indices indicating the specific purpose of the predictor to suppress complex notation. For the moment, the predictor may also be considered the usual predictor in mean regression representing the expectation of the response.

Instead of restricting our attention to regression models with linear predictors $\eta_i = \boldsymbol{z}_i'\boldsymbol{\beta}$, we are interested in semiparametric regression models with structured additive predictors (Fahrmeir *et al*., 2004; Kneib *et al*., 2009) of the generic form

$$\eta_i = \beta_0 + \sum_{j=1}^{p} f_j(\boldsymbol{z}_i), \tag{2.1}$$

where β_0 is an intercept describing the overall level of the predictor and the generic functions $f_j(\boldsymbol{z}_i)$ reflect different types of regression effects depending on (subsets of) the complete covariate vector $\boldsymbol{z}_i$. Associated with each function is a penalty term $\lambda_j \text{pen}(f_j)$ that enforces specific properties of the function such as smoothness or sparsity and $\lambda_j \geq 0$ are the corresponding smoothing parameters that govern the impact of the penalty.

A broad and flexible class of function types is obtained with the following assumptions:

- The functions f_j are approximated in terms of (possibly non-standard) basis function expansions

$$f_j(\boldsymbol{z}) = \sum_{k=1}^{K} \beta_{jk} B_k(\boldsymbol{z}),$$

  where $B_k(\boldsymbol{z})$ are the basis functions and β_{jk} denote the corresponding basis coefficients.
- The penalty is quadratic in the vector of basis coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})'$, i.e.

$$\text{pen}(f_j) = \boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j$$

  with penalty matrix $\boldsymbol{K}_j$ chosen such that the desired regularization properties are achieved. In a Bayesian formulation, we would equivalently assume that the regression coefficients are assigned a normal prior $\boldsymbol{\beta}_j \sim N(\boldsymbol{0}, \delta_j^2 \boldsymbol{K}_j^-)$, where the variance δ_j^2 represents an inverse smoothing parameter, the penalty matrix $\boldsymbol{K}_j$ defines the precision of the normal distribution and $\boldsymbol{K}_j^-$ is the generalized inverse. Note that $\boldsymbol{K}_j$ is not necessarily of full rank such that the normal

distribution may be partially improper with density

$$p(\boldsymbol{\beta}_j|\delta_j^2) \propto \left(\frac{1}{\delta_j^2}\right)^{\frac{\text{rank}(\boldsymbol{K}_j)}{2}} \exp\left(-\frac{1}{2\delta_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right) \tag{2.2}$$

that cannot be normalized to integrate to one.

This framework comprises several well-known special cases such as generalized additive models (Hastie and Tibshirani, 1990; Wood, 2006), varying coefficient models (Hastie and Tibshirani, 1993) or geoadditive models (Kamman and Wand, 2003) and has been introduced in full generality as structured additive regression in Fahrmeir *et al.* (2004) for Bayesian and penalized likelihood inference and (Kneib *et al.*, 2009) for functional gradient descent boosting.

The basis function expansion allows us to rewrite the generic predictor in matrix notation as

$$\boldsymbol{\eta} = \beta_0 1 + \boldsymbol{B}_1\boldsymbol{\beta}_1 + \ldots + \boldsymbol{B}_p\boldsymbol{\beta}_p,$$

where 1 is a vector of ones while the design matrices $\boldsymbol{B}_j$ are obtained from the evaluations of the basis functions, i.e. $\boldsymbol{B}[i,k] = B_k(z_i)$.

Some special predictor components arising for specific choices of the basis functions and penalties are as follows:

- Penalized splines for nonlinear effects $f(z) = f(x)$ of a single continuous covariate x: The basis functions are B-spline bases $B_k^{(l)}(x)$ of fixed degree l defined upon a set of equidistant knots while the penalty matrix is given by $\boldsymbol{K} = \boldsymbol{D}'\boldsymbol{D}$, where $\boldsymbol{D}$ is a dth order difference matrix. Cubic penalized splines with second order difference penalty can be considered a low rank approximation to smoothing splines based on an integrated squared second derivative penalty.
- Varying coefficient terms $f(z) = x_1 f(x_2)$, where the effect of covariate x_1 varies smoothly with respect to the continuous covariate x_2: The function $f(x_2)$ is again represented using penalized splines and, as a consequence, the elements of the design matrix are given by $\boldsymbol{B}[i,k] = x_{i1}B_k^{(l)}(x_{i2})$. The penalty remains the same as with usual penalized splines.
- Markov random fields $f(z) = f(s)$, where $s \in \{1,\ldots,S\}$ denotes a discrete spatial location index (e.g., regions, counties, ...): The basis functions are indicator functions for the different regions, i.e., $B_s(s_i) = I(s_i = s)$ is equal to one if observation i is collected in region s and zero otherwise. To enforce spatial smoothness, the penalty matrix is an adjacency matrix with elements

$$\boldsymbol{K}[s,r] = \begin{cases} -1, & \text{if regions } s \text{ and } r \text{ are neighbors,} \\ N_s, & \text{if } s = r \text{ and} \\ 0, & \text{otherwise,} \end{cases}$$

where N_s denotes the number of neighbors for region s.

- Tensor product interaction surfaces $f(\boldsymbol{z}) = f(x_1, x_2)$ of two continuous covariates: Based on penalized spline bases $B_k^{(1)}(x_1)$ and $B_l^{(2)}(x_2)$ with penalty matrices $\boldsymbol{K}_1$ and $\boldsymbol{K}_2$, the basis functions for the interaction surface are constructed by considering all tensor product basis functions $B_{kl}(x_1, x_2) = B_k^{(1)}(x_1) \cdot B_l^{(2)}(x_2)$ and forming the kronecker penalty $\boldsymbol{K} = (\boldsymbol{K}_1 \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{K}_2)$.
- Radial basis functions $f(\boldsymbol{z}) = f(\boldsymbol{x})$ for a vector of continuous covariates $\boldsymbol{x}$: Within the framework of reproducing kernel Hilbert spaces, the basis functions are given by kernel functions $k(\cdot, \cdot)$ centered at the observed covariate values $\boldsymbol{x}_k$, i.e., $B_k(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}_k)$, while the entries of the penalty matrix are given by $\boldsymbol{K}[k, l] = k(\boldsymbol{x}_k, \boldsymbol{x}_l)$.
- Cluster-specific random effects $f(\boldsymbol{z}) = \beta_c$ depending on the cluster index $c \in \{1, \ldots, C\}$: As for regional spatial effects, the basis functions are indicator functions for the different clusters, i.e. $B_c(c_i) = I(s_i = c)$ is equal to one if observation i belongs to cluster c and zero otherwise. For i.i.d. random effects, the penalty matrix is simply the identity matrix, i.e., $\boldsymbol{K} = \boldsymbol{I}$.

3 Inferential procedures

To fit semiparametric regression models, several inferential procedures have been proposed so far. In the following, we will mostly concentrate on three potential avenues to adapt inference in semiparametric mean regression to regression models beyond the mean:

- Direct optimization of a lack of fit criterion,
- Bayesian inference and
- functional gradient descent boosting.

These three areas have been chosen since they reflect the current state of the art in semiparametric regression utilizing the predictor structure discussed in the previous section.

Direct optimization relies on the description of the estimation task in terms of a lack of fit criterion $l(\boldsymbol{y}, \boldsymbol{\eta})$ that describes the discrepancy between a candidate set of predictors $\boldsymbol{\eta}$ and the observed responses $\boldsymbol{y}$. In parametric settings, the lack of fit may be described by the negative log-likelihood while in more general settings any type of a loss function may be employed. Estimation is then achieved by minimizing the penalized lack of fit criterion

$$l(\boldsymbol{y}, \boldsymbol{\eta}) + \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j$$

using some numerical optimization procedure such as penalized Fisher scoring if the loss function is twice continuously differentiable with respect to the basis coefficients

of the structured additive predictor. The selection of smoothing parameters is then typically based on some external criterion such as a generalized cross validation score or Akaike's information criterion. An increasingly popular alternative that allows to simultaneously estimate basis coefficients and smoothing parameters employs the connection between penalized smoothing and mixed models (see Green, 1987; Speed, 1991, for early references and Ruppert *et al*., 2003; Fahrmeir and Kneib, 2011 for comprehensive introductions). Here, the basis coefficients $\boldsymbol{\beta}_j$ of a specific term in the predictor are treated as random effects $\boldsymbol{\beta}_j \sim \mathrm{N}(\mathbf{0}, \delta_j^2 \boldsymbol{K}_j^-)$ such that the smoothing variance δ_j^2 (and therefore also the smoothing parameter λ_j) can be estimated using restricted maximum likelihood principles.

Bayesian inference relies on the interpretation of penalties $\mathrm{pen}(\boldsymbol{\beta}_j)$ as prior distributions (2.2) and requires the specification of an observation model $p(\boldsymbol{y}|\boldsymbol{\eta})$ for the responses corresponding to the likelihood in a frequentist setting. Then the posterior can in principle be assessed via the proportionality

$$p(\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\eta}) p(\boldsymbol{\beta}_0) p(\boldsymbol{\beta}_1) \cdot \ldots \cdot p(\boldsymbol{\beta}_p)$$

although in most cases of practical relevance the normalizing constant for the right hand side cannot be determined analytically such that the posterior is only known upon proportionality. Still, maximization of the posterior for given smoothing parameters is possible using the same numerical optimization schemes as for direct optimization. The resulting posterior mode estimates then coincide with penalized likelihood estimates. However, in most cases, Bayesian inference will rely on Markov chain Monte Carlo (MCMC) simulation techniques that do not only provide the posterior mode but also give access to the full posterior distribution. In particular, if the observation model is Gaussian or has a latent Gaussian representation (as for example in case of the probit model), MCMC simulations can be conveniently implemented since the normal prior (2.2) is conjugate and will lead to simple Gibbs sample updates. For more general types of distributions, iteratively weighted least squares proposals based on a quadratic approximation of the full conditional are generally a good alternative that automatically adapts to the form of the full conditional without requiring the manual tuning of proposal hyperparameters (Gamerman, 1997; Brezger and Lang, 2006). Inference based on MCMC also naturally incorporates estimation of the smoothing variance δ_j^2 by assigning an additional hyperprior such as an inverse gamma prior $\delta_j^2 \sim \mathrm{IG}(a, b)$ that is conjugate to the multivariate normal prior of the basis coefficients. A recent alternative to Bayesian inference utilizing MCMC is provided by nested integrated Laplace approximations (Rue *et al*., 2009) that allow to construct precise approximate solutions for marginal posterior distributions of interest.

A third alternative to perform inference in structured additive regression models is given by functional gradient descent boosting, an approach that originated from the machine learning community for optimization in complex problems (see Bühlmann and Hothorn 2007 for an introduction from a statistical perspective). The

basic idea is to fit simple base-learning procedures to iteratively updated gradients of an optimization problem to achieve a final estimate. When using a component-wise boosting approach, where separate base-learners are specified for each of the model components in the semiparametric predictor (2.1), boosting has the particular advantage that it combines model estimation (including data-driven determination of the appropriate amount of smoothness) with automatic variable selection and model choice.

In case of structured additive regression with the quadratic penalties discussed in the previous section, a suitable base-learner class is given by penalized least squares fits as characterized by the hat matrix

$$\boldsymbol{H}_j = \boldsymbol{B}_j(\boldsymbol{B}_j'\boldsymbol{B}_j + \lambda_j \boldsymbol{K}_j)^{-1}\boldsymbol{B}_j'$$

that projects the current gradients to the fitted values (see Kneib *et al.*, 2009 for details). Note that in this case, the smoothing parameter λ_j is not a hyperparameter of the model but only a tuning constant of the base-learner. In fact, the exact value of λ_j is not of that much importance as long as all base-learners are of comparable complexity (see Hofner *et al.*, 2012).

A compact description of functional gradient descent boosting is provided by the following algorithm:

1. Initialize the predictor $\hat{\boldsymbol{\eta}}^{[0]} \equiv$ offset and the functions $\hat{\boldsymbol{f}}_j^{[0]} \equiv \boldsymbol{0}$; set $m = 0$.
2. Increase m by 1. Compute the negative gradients ('residuals')

$$u_i = -\frac{\partial}{\partial \eta} l(y_i, \eta)\big|_{\eta=\hat{\eta}^{[m-1]}}, \quad i = 1, \ldots, n.$$

3. Fit the base-learners to the negative gradient vector $\boldsymbol{u} = (u_1, \ldots, u_n)'$, yielding

$$\hat{\boldsymbol{u}}_j = \boldsymbol{B}_j(\boldsymbol{B}_j'\boldsymbol{B}_j + \lambda_j \boldsymbol{K}_j)^{-1}\boldsymbol{B}_j\boldsymbol{u}.$$

4. Find the best-fitting base-learner

$$j^* = \arg\min_j \sum_{i=1}^{n} (u_i - \hat{u}_{ij})^2.$$

5. Update $\hat{\boldsymbol{f}}_{j^*}^{[m]} = \hat{\boldsymbol{f}}_{j^*}^{[m-1]} + \nu \cdot \hat{\boldsymbol{u}}_{j^*}$ and keep all other effects constant, i.e. $\hat{\boldsymbol{f}}_j^{[m]} = \hat{\boldsymbol{f}}^{[m-1]}$, $j \neq j^*$.
6. Iterate steps 2 to 4 until $m = m_{\text{stop}}$.

The step length factor ν is applied in the update step to prevent the algorithm from overfitting the current base-learner. This is particularly advantageous in case of correlated covariates where boosting gives all covariates a chance to enter the final estimate while using a full fit in the update would usually only include one

representative out of the correlated set of covariates. Usual values for the step length are $\nu = 0.1$ or $\nu = 0.01$. The most crucial quantity for the boosting algorithm is the number of boosting iterations m_{stop}. When using a rather larger value of m_{stop}, this will yield estimates very close to the minimum of the loss function. However, in complex models with a large number of terms, some kind of regularization is usually desired. In boosting, this regularization is achieved implicitly by stopping early, more precisely stopping in an iteration that yields results that are well generalizable to new data sets. Hence, m_{stop} is often determined by cross validation techniques. Another advantage of early stopping in componentwise boosting is that covariates that carry only very few information on the responses will effectively drop out of the model since they will not be selected in the early iterations of the boosting algorithm.

Before actually relating the inferential procedures to the three types of regression models we want to consider, we will now give some information on their advantages and disadvantages.

Direct optimization has the advantage that, unlike MCMC inference and partially also boosting, it does not depend on hyperprior choices, sampling performance or the choice of tuning constants such as step length and smoothing parameters for the base-learners. It is therefore also often advantageous for theoretical considerations since the estimates can be characterized as roots of the (quasi-) score function of the lack of fit criterion. Finally, the connection to mixed models for smoothing parameter selection does not only allow for a routine determination of an adequate amount of smoothness but also bridges the gap between frequentist and Bayesian interpretation of semiparametric regression when the basis coefficients are interpreted as random effects. On the other hand, direct optimization is not very modular and therefore also smaller modifications in the model structure often require the re-development of numerical implementations. For example, utilizing an L_1 penalty for some of the basis coefficients not only affects estimation of these basis coefficients but would induce a combination of L_1 and L_2 penalties that makes inference challenging. Moreover, the connection to mixed models is usually only useful with quadratic penalty terms since for the corresponding case of Gaussian random effects distributions, estimation of random effects variances is well developed. Finally, measures of uncertainty and hypothesis tests will usually rely on asymptotic arguments when using direct optimization techniques.

Bayesian inference based on MCMC simulations has the distinct advantage of being very flexible since it decomposes the estimation of the complete model into smaller blocks that are treated separately in a modular fashion using the corresponding full conditionals. As a consequence, the hierarchical model formulation in structured additive regression can be fully exploited to derive both numerically efficient sampling techniques and to build complex models from simple building blocks. In particular, when modifying the prior for one basis coefficient block, this does not affect the full conditional for the remaining basis coefficients. The decomposition into parameter blocks also has the advantage that the estimation complexity only grows linearly with the number of model terms while direct optimization will usually grow with quadratic or cubic order. Finally, MCMC provides access to the full posterior

distribution and therefore enables exact uncertainty assessments also for complex functions of the unknown parameters without the need of asymptotic arguments. On the downside, MCMC requires the specification of a likelihood and is therefore not directly applicable in quasi-likelihood or distribution-free approaches (although we will abuse the Bayesian machinery later on in quantile regression nevertheless). Moreover, the suitable choice of a proposal density as well as sensitivity with respect to prior choices including hyperparameters may lead to debates about the validity of the results obtained with MCMC (although this may not be that much of an issue in situations with enough informative data and IWLS proposals). Another technical difficulty is the question whether the posterior is actually proper although some of the basis coefficient priors (2.2) are (partially) improper (see Fahrmeir and Kneib, 2009, for some results in this regard).

For boosting, the main advantage is its automatic ability to perform variable selection and model choice by early stopping when choosing the number of boosting iterations appropriately. It also offers considerable flexibility with respect to the considered optimization problem since basically any loss function can be plugged into the generic algorithm described above. This in particular enables the estimation of robust regression models or quasi-likelihood regression. In addition, the decomposition of the model into base-learners that are fitted separately provides a similar kind of modularity as in case of MCMC although the flexibility is more limited for boosting. On the other hand, boosting does only provide point estimates for the predictor terms while no measures of estimation uncertainty are directly available.

In summary, there is no generally favourable approach to estimate structured additive regression models but all approaches have their characteristic properties. The discussion above will hopefully give some guidance on which approach may be most suitable for a given problem.

4 Generalized additive models for location, scale and shape (GAMLSS)

GAMLSS provide a unified framework for estimating semiparametric regression models when assuming that the responses y_i follow distributions depending on up to four parameters $(\mu_i, \sigma_i, \nu_i, \xi_i)$, where usually μ_i and σ_i are a location and a scale parameter, respectively, while ν_i and ξ_i correspond to shape parameters such as skewness or kurtosis. The limitation to four parameters is only chosen for convenience since common distributions rarely have more than four distributional parameters and because interpretation becomes quite messy in more complex cases. Each of the distributional parameters is related to a predictor via a suitable link function, i.e.

$$\mu = g_1(\eta_1), \quad \sigma = g_2(\eta_2), \quad \nu = g_4(\eta_4) \ldots$$

The class of distributions covered by GAMLSS is very broad and comprises, at the moment, more than 50 different distributions. For continuous responses, the most prominent examples are the normal distribution (with up to two parameters and either the standard deviation or the variance as scale parameter), the power

exponential distribution (with up to three parameters), the gamma distribution (with up to three parameters), the t-distribution (with up to three parameters) or the Box-Cox power exponential distribution (with four parameters). Discrete distributions such as zero-inflated Poisson or zero-inflated negative binomial are also supported but are not considered here since they do not fit well in the framework where we aim at comparing GAMLSS with quantile and expectile regression.

Estimation in GAMLSS usually relies on likelihood principles and requires the (at least numerical) availability of first (and optimally second) derivatives to facilitate optimization via Fisher-scoring type algorithms. Penalized maximum likelihood inference for GAMLSS including the automatic determination of smoothing parameters is included in the R add-on package gamlss (Stasinopoulos and Rigby, 2011).

Bayesian inference in GAMLSS is also conceptually straightforward but the development of suitable proposal densities for the basis coefficient blocks can become quite challenging depending on the type of distribution considered. Some first attempts in this direction are available in Cottet *et al.* (2008) and we are currently in the process of developing a general strategy based on iteratively weighted least squares proposals for Bayesian GAMLSS that shall be included in the software package BayesX (Belitz *et al.*, 2012).

The derivation of boosting algorithms is somewhat more challenging since a GAMLSS comprises not only one but also several predictors. As a consequence, the optimal stopping iteration may be different for each of the predictors such that algorithmic fine-tuning is required. Such a proposal is made in Mayr *et al.* (2012) and implemented in the R add-on package gamboostLSS (Hofner *et al.*, 2011).

The major advantage of GAMLSS is that the predictors act directly on interpretable response quantities and therefore facilitate the understanding of the estimated regression effects. As an example, we estimate a location-scale model

$$y_i \sim \mathrm{N}(\mu_i, \sigma_i^2),$$

with

$$\mu_i = \eta_{i1} \quad \text{and} \quad \sigma_i = \exp(\eta_{i2})$$

based on the normal distribution for the Munich rental guide data, where both predictors for the mean and the standard deviation are additively composed of nonlinear effects of living area and year of construction. For both nonlinear effects, a cubic penalized spline with 20 inner knots and second difference penalty has been chosen and estimation was carried out using the penalized maximum likelihood approach implemented in gamlss. The estimation results are visualized in Figure 2 and basically confirm our initial findings from the introduction. More specifically, we find an almost linear increase in the expected net rent with increasing living area and also an increasing net rent for newer buildings. The latter effect is almost absent for very old buildings but only starts for flats built after 1950. For very new buildings, there seems to be a small decline in net rents but the associated uncertainty does not allow to make a strong statement about that decline. For the standard deviation, we also find a linear increase with the living area that confirms our exploratory detection of heterogeneous variances for this covariate. For the year of construction, the effect

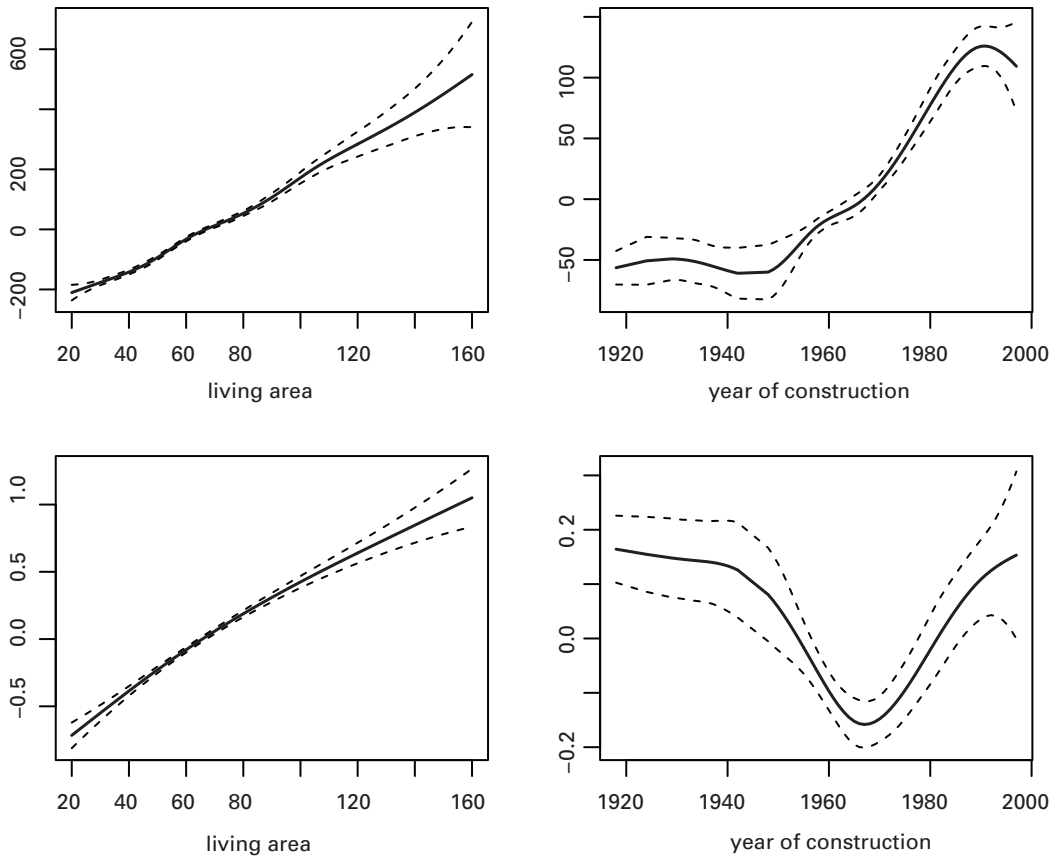

Figure 2 Additive model fits obtained with a location-scale normal model. The top panel shows results for the mean and the lower panel results for the standard deviation. The solid line indicated the penalized maximum likelihood estimate, the dashed lines indicate pointwise 95% confidence intervals

on the standard deviation is much smaller in magnitude but gives some indication of reduced variability during the 1960s and 1970s. Although these results are in line with the ones obtained in the introduction based on parametric models, the GAMLSS specification based on splines has two important advantages: It considers both living area and year of construction in one joint additive model and determines the amount of smoothness and nonlinearity from the data instead of imposing a parametric form for the effects.

Another advantage of GAMLSS in addition to easy interpretability is that we formulate one coherent model for the response distribution. This also implies that quantile curves derived from the estimated model will never cross (which will be an issue in quantile regression, where separate curves are estimated per quantile,

see the next section). Finally, GAMLSS can be interpreted and estimated in both a frequentist and Bayesian context without any conceptual difficulties.

On the downside, GAMLSS bear the risk of mis-specifying the response model. For example, in case of the rental guide, we found some evidence for varying skewness with year of construction in the introduction. With the above location-scale model, we will not be able to detect or capture such effects. Of course, utilizing a more complex model with more parameters allows to address the skewness problem, but naturally the maximum flexibility is limited and any distributional choice bears the risk of deciding for the wrong model. An exploratory tool for assessing the model fit and also for comparing different model specifications are quantile residuals as suggested in Dunn and Smyth (1996). Another disadvantage of GAMLSS is the fact that no direct estimates of quantiles of the response distribution are available. If the ultimate aim of an analysis is the determination of specific quantiles, then it may be more adequate to formulate a model directly for these quantiles (as detailed in the following section).

5 Quantile regression

Quantile regression for the τ-quantile starts from the model

$$y_i = \eta_{i\tau} + \varepsilon_{i\tau}, \quad F_{\varepsilon_{i\tau}}(0) = \tau,$$

where $F_{\varepsilon_{i\tau}}(\cdot)$ denotes the cumulative distribution function of $\varepsilon_{i\tau}$. This defines a specification that is similar to usual mean regression but replaces the assumption of zero means for the error terms with the assumption of zero τ-quantiles. As a consequence, the predictor $\eta_{i\tau}$ is the τ-quantile of the response y_i since

$$\tau = F_{\varepsilon_{i\tau}}(0) = P(\varepsilon_{i\tau} \leq 0) = P(\eta_{i\tau} + \varepsilon_{i\tau} \leq \eta_{i\tau}) = P(y_i \leq \eta_{i\tau}) = F_{y_i}(\eta_{i\tau}).$$

Note that no further assumptions (apart from independence) are made on the error terms and therefore quantile regression is also applicable in situations with heteroscedastic error terms. In fact, quantile regression is only of interest in such situations, where not only the mean depends on covariates but also other properties of the response distribution.

Classical estimation in quantile regression relies on optimizing an asymmetrically weighted absolute error criterion. Recall that empirical quantiles $\hat{q}_\tau$ based on an i.i.d. sample of observations $y_1, \ldots, y_n$ can be estimated as

$$\hat{q}_\tau = \arg\min_q \sum_{i=1}^{n} w_\tau(y_i, q)|y_i - q|$$

with asymmetric weights

$$w_\tau(y_i, q) = \begin{cases} 1-\tau, & y_i < q, \\ 0, & y_i = q, \\ \tau, & y_i > q, \end{cases}$$

that basically weight observations below and above the quantile of interest differently to shift the estimate to upper or lower parts of the sample. Regression quantiles can then in analogy be determined by replacing the common quantile q with the predictor $\eta_{i\tau}$ of the semiparametric regression model and augmenting the penalty terms, yielding

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})|y_i - \eta_{i\tau}| + \sum_{j=1}^{p} \lambda_j \,\mathrm{pen}(f_j). \tag{5.1}$$

For quantile regression, linear programming is the standard approach in parametric model specifications that allows for routine and fast optimization of the asymmetrically weighted L_1-loss function (see Koenker, 2005, for details). This approach can still be used in combination with L_1 penalty terms arising for example from the LASSO or in total variation penalization for spline regression (Koenker *et al.*, 1994) where

$$\mathrm{pen}(f_j) = \int |f_j''(x)|\mathrm{d}x.$$

However, the class of quadratic penalties introduced in Section 2 does not fit in the L_1 framework and renders linear programming inappropriate. This is particularly problematic when moving from purely additive quantile regression to extended models that also comprise spatial or random effects. In addition, the simultaneous estimation of basis coefficients and smoothing parameters is still challenging and largely unsolved in the linear programming framework.

While Bayesian inference is obviously an alternative for estimation in GAMLSS, it seems to be more complicated to relate Bayesian approaches to the nonparametric formulation of quantile regression that does not involve an explicit specification of the observation model. However, due to the formal equivalence between penalized estimates and posterior modes based on suitable auxiliary error distributions, such a connection is indeed possible (see Yu and Moyeed, 2001 and Yue and Rue, 2011, for references on Bayesian quantile regression). Assuming the asymmetric Laplace distribution $\mathrm{ALD}(0, \sigma^2, \tau)$ with density

$$p_{\varepsilon_{i\tau}}(\varepsilon_i) = \frac{\tau(1-\tau)}{\sigma^2} \exp\left(-w_\tau(\varepsilon_i, 0)\frac{|\varepsilon_i|}{\sigma^2}\right).$$

for the error terms induces the likelihood

$$\exp\left(-\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})\frac{|y_i - \eta_{i\tau}|}{\sigma^2}\right)$$

and therefore maximizing the corresponding penalized likelihood is equivalent to minimizing (5.1).

The main advantage of the connection between quantile regression and the asymmetric Laplace distribution is that the latter has a representation as a location-scale mixture of normals and therefore enables the construction of efficient Gibbs sampling algorithms based on this latent Gaussian formulation. More specifically, if $z_i \,|\, \sigma^2 \sim \text{Exp}(1/\sigma^2)$ and

$$y_i \,|\, z_i, \boldsymbol{\eta}_\tau, \sigma^2 \sim N(\boldsymbol{\eta}_\tau + \xi z_i, \sigma^2/w_i)$$

with

$$\xi = \frac{1-2\tau}{\tau(1-\tau)}, \qquad w_i = \frac{1}{\delta^2 z_i}, \qquad \delta^2 = \frac{2}{\tau(1-\tau)},$$

then y_i is marginally $\text{ALD}(\eta_{i\tau}, \sigma^2, \tau)$ distributed. It therefore turns out that, after imputing z_i as additional unknowns in the Bayesian algorithm, y_i can be treated as conditionally Gaussian with the same regression predictor as the quantile regression problem of interest if an additional offset term ξz_i and weights w_i are included. As a consequence, Gibbs sampling updates result for all unknown quantities (including the latent variables z_i) and Bayesian inference becomes feasible even for rather complex predictor structures and including the automatic determination of smoothing parameters (see Yue and Rue, 2011, and Waldmann *et al*., 2013, for details on the location scale representation of the ALD and its application in Bayesian semiparametric quantile regression). An alternative, approximate solution to the optimization of Bayesian quantile regression within the framework of integrated nested Laplace approximations is provided by Yue and Rue (2011).

In the following, we discuss a geoadditive extension of the Munich rental guide analysis, where the quantile-specific predictor is given by

$$\eta_\tau = \beta_0 + f_1(\text{living area}) + f_2(\text{year of construction}) + f_3(\text{subquarter}),$$

where f_1 and f_2 are Bayesian cubic penalized splines with 20 inner knots and second order difference penalty for living area and year of construction and f_3 corresponds to a Markov random field defined upon the roughly 450 subquarters of the City of Munich where two subquarters are treated as neighbors if they share a common boundary. Estimation is based on a Gibbs sampler implemented in BayesX (Belitz *et al*., 2012). Figure 3 shows selected posterior mean estimates for the 0.1, 0.5 and 0.9 quantiles together with pointwise 95% credible intervals. For the living area, we obtain close to linear estimates for all quantiles but the slope of the estimated effect seems to increase with larger values for the quantile τ (as already seen in the introduction). For the year of construction, there is less variation across the different

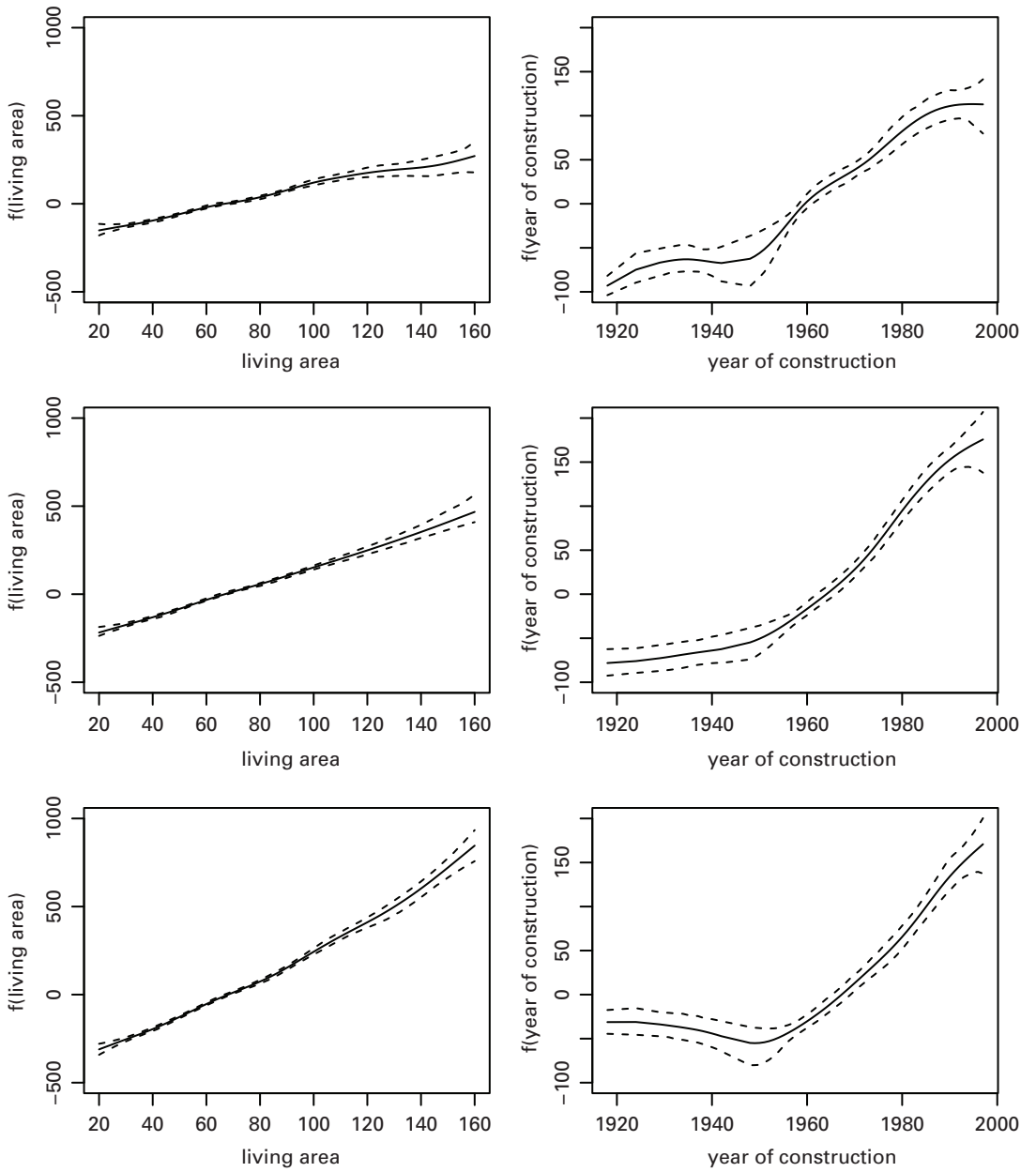

Figure 3 Estimated nonlinear effects in a Bayesian geoadditive quantile regression model for the Munich rental guide. The top row shows results for $\tau = 0.1$, the middle row for $\tau = 0.5$ and the bottom row for $\tau = 0.9$

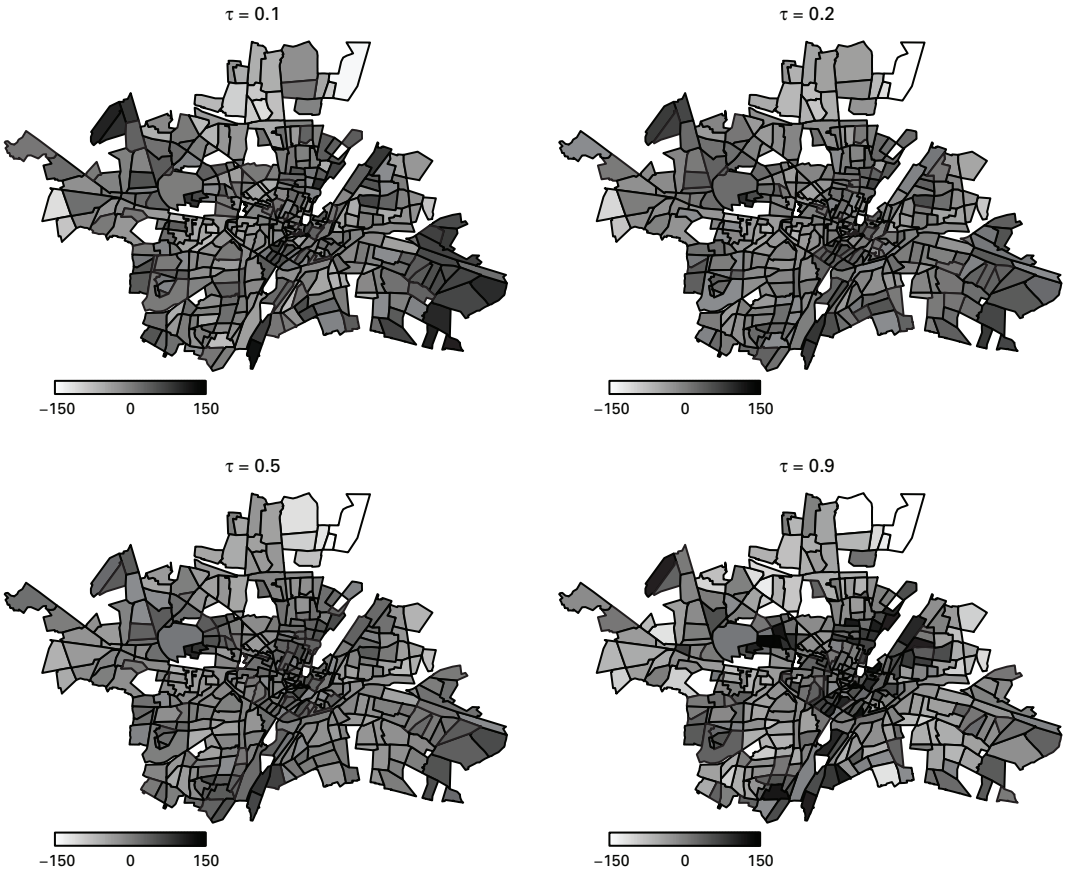

Figure 4 Estimated spatial effects in a Bayesian geoadditive quantile regression model for the Munich rental guide

quantiles. Figure 4 shows estimated spatial effects for four different quantiles. There seems to be a somewhat larger variation in the spatial effect for more extreme quantiles, i.e., either cheap or expensive flats seem to be more heterogeneous across the subquarters of Munich. In addition, there is a tendency for larger effects in the center where even the 10% quantile is higher than on average.

Obviously, Bayesian quantile regression abuses the likelihood for the asymmetric Laplace distribution based on the formal equivalence of posterior modes and penalized maximum likelihood inference. This may seem questionable since of course the data will usually not follow the asymmetric Laplace distribution and moreover different distributional specifications are used for each quantile, so that there is no coherent supermodel combining the separate specification. In addition, we obtain posterior mean estimates and also estimate the smoothing parameters along with

the rest of the parameters which makes the connection between the original optimization criterion and the Bayesian formulation even weaker. In summary, it is not automatically clear that either the point estimates or the credible bands obtained from our Bayesian analyses can be interpreted in a meaningful way. This problem is investigated in more detail in Waldmann *et al.* (2013) who study Bayesian quantile regression in simulations and complex case studies and also compare the results with those from a frequentist analysis using quantile smoothing splines (Koenker *et al.*, 1994). Their findings indicate that the point estimates of Bayesian quantile regression are typically very close to the true model. The credible bands are usually too narrow in particular for extreme quantiles but still reasonably reflect the uncertainty attached to the estimated effects.

As a consequence, alternative avenues for Bayesian quantile regression have also been explored in the literature. Basically, these approaches refrain from using the asymmetrically weighted absolute error criterion underlying fequentist quantile regression and instead aim at modelling the error distribution F_ε in a flexible, data-driven way while still incorporating the quantile restriction $F_\varepsilon(\tau) = 0$ such that the covariates affect the quantile of interest. Examples include Kottas and Krnjajic (2009) based on different Dirichlet process mixtures for the quantile-specific error distribution, Reich *et al.* (2010) utilising a joint location-scale model for all quantiles with linear predictors for mean and standard deviation and Dirichlet process mixture for the error density, and Taddy and Kottas (2010) utilising a joint Dirichlet process mixture for covariates and responses.

An alternative to a Bayesian treatment of quantile regression is provided by a special instance of the boosting algorithm described in Section 3 (see Fenske *et al.*, 2011 for details). This algorithm requires the gradients of the loss function which, in case of quantile regression, are given by

$$u_i = \begin{cases} \tau, & y_i > \hat{\eta}_i^{[m-1]}, \\ 0, & y_i = \hat{\eta}_i^{[m-1]}, \\ \tau - 1, & y_i < \hat{\eta}_i^{[m-1]}. \end{cases}$$

Note that the case $y_i = \hat{\eta}_i^{[m-1]}$ only appears with zero probability and therefore the definition in this point is basically arbitrary. Based on these gradients, the boosting algorithm can be applied without any further changes.

Concerning software for quantile regression, the usual starting point will be the R-package quantreg (Koenker, 2011) that collects both standard parametric quantile regression and a number of extensions comprising quantile smoothing splines in additive models. Most of the optimization in this package is based on linear programming techniques. Bayesian structured additive quantile regression is available in BayesX (Belitz *et al.*, 2012) while the boosting approach is implemented in the R-package mboost (Hothorn *et al.*, 2011).

The main advantage of classical quantile regression is that it allows to analyse regression data in a completely distribution-free approach that avoids restrictive assumptions about the error terms and only requires independence. It also facilitates

easy interpretation of the estimated effects that impact the conditional quantiles of the response distribution. While flexibility of the possible predictor structures is basically limited to additive model specifications (possibly comprising bivariate surfaces based on triograms; Koenker and Mizera, 2004), both the Bayesian formulation and boosting enable the application of the full potential of structured additive regression and also allow for the data-driven determination of smoothing parameters.

A disadvantage of the Bayesian approach already discussed above is that it relies on a misspecified likelihood. As a consequence, formal inferences about the estimated effects are not possible although simulation evidence suggests that these inferences can be considered to be relatively reliable at least for not too extreme quantiles.

A theoretical disadvantage of quantile regression in general is that the estimated cumulative distribution function for the responses is a step function (similar as the empirical cumulative distribution function that can be considered the special case of quantile regression with only an intercept term) while the theoretical cumulative distribution function is usually assumed to be continuous. Of course, this problem is not too important for larger data sets where the steps will be rather small. Another disadvantage of quantile regression is that estimates for each quantile are determined separately. As a consequence, crossing quantile curves where $\hat{\eta}_{\tau_1} > \hat{\eta}_{\tau_2}$ for some $\tau_1 < \tau_2$ are frequently observed especially when considering a dense set of quantiles. Possibilities to circumvent this problem for example based on simultaneous estimation of all quantiles in quantile sheets (Schnabel and Eilers, 2013a) or based on non-decreasing rearrangements (Dette and Volgushev, 2008) have been proposed in the literature but these always require additional efforts and are not always applicable with additive predictor specifications.

6 Expectile regression

An alternative to quantile regression is obtained when replacing the asymmetric absolute deviations with asymmetric quadratic deviations, yielding the optimization criterion

$$\sum_{i=1}^{n} w_\tau(y_i, \eta_{i\tau})(y_i - \eta_{i\tau})^2 + \sum_{j=1}^{p} \lambda_j \,\mathrm{pen}(f_j). \tag{6.1}$$

This has been originally proposed by Newey and Powell (1987) under the term expectile regression (in a linear framework without penalization) and has recently regained considerable attention due to its computational advantages (Schnabel and Eilers, 2009; Sobotka and Kneib, 2012; Sobotka *et al.*, 2013a). The expectile regression criterion exhibits a closer connection to ordinary least squares estimation and in fact usual mean regression appears as a special case with $\tau = 0.5$ where (6.1) reduces to a scaled least squares criterion.

While it is obvious that expectiles provide a counterpart to quantiles, they do not enjoy the easy interpretability of quantiles. It is therefore of interest to not only

define an optimization criterion but also gain a deeper understanding of what an expectile is. Empirical expectiles for i.i.d. samples are obtained as a special case of (6.1) when considering a constant predictor $\eta_{i\tau} = e_\tau$ (and dropping the penalty) yielding a weighted average of the responses, i.e.

$$\hat{e}_\tau = \sum_{i=1}^{n} w_\tau(y_i, \hat{e}_\tau) y_i.$$

Note, however, that the weights depend on the solution so that the solution cannot be determined analytically but has to be computed iteratively. One can also define theoretical expectiles e_τ for a random variable y by replacing the empirical risk in (6.1) with the expectation, i.e.

$$e_\tau = \arg\min_e E(w_\tau(y, e)(y - e)^2).$$

It can then be shown that the solution can also be characterized via

$$\tau = \frac{\int_{-\infty}^{e_\tau} |y - e_\tau| \, f_y(y) \mathrm{d}y}{\int_{-\infty}^{\infty} |y - e_\tau| \, f_y(y) \mathrm{d}y} = \frac{G_y(e_\tau) - e_\tau F_y(e_\tau)}{2(G_y(e_\tau) - e_\tau F_y(e_\tau)) + (e_\tau - \mu)},$$

where $f_y(\cdot)$ and $F_y(\cdot)$ denote the density and cumulative distribution function of y, $G_y(e) = \int_{-\infty}^{e} y f_y(y) \mathrm{d}y$ is the partial moment function of y and $G_y(\infty) = \mu$ is the expectation of y. When not considering only one expectile for given τ but a dense set of asymmetries τ, the complete distribution of y can be characterized by the expectile function similar as with the quantile function. This also offers the possibility to determine quantiles from expectiles as shown in Schnabel and Eilers (2011b) and Waltrup *et al.* (2012). Waltrup *et al.* (2012) also compare expectiles and quantiles based on theoretical investigations and simulations and find that expectiles may be more efficient in estimating quantiles than the direct calculation of quantiles for a number of distributions and also show a smaller probability to obtain crossing expectile curves than in the direct estimation of quantiles. Still, non-crossing expectiles can be enforced with similar approaches as discussed for quantile regression. Another point in favour of expectiles is their relation to commonly applied risk measures in finance such as the expected shortfall, see Taylor (2008) or Kuan *et al.* (2008).

However, the main advantage of expectile regression is that estimates can be derived fairly easily based on iteratively weighted least squares iterations

$$\hat{\boldsymbol{\beta}}_{j\tau}^{[t+1]} = (\boldsymbol{B}_j' \boldsymbol{W}_\tau^{[t]} \boldsymbol{B}_j + \lambda_j \boldsymbol{K}_j)^{-1} \boldsymbol{B}_j' \boldsymbol{W}_\tau^{[t]} \boldsymbol{y},$$

where $\boldsymbol{B}_j$ is the design matrix associated with the jth model term, $\boldsymbol{y}$ is the vector of responses and $\boldsymbol{W}_\tau = \mathrm{diag}(w_\tau(y_1, \eta_{1\tau}), \ldots, w_\tau(y_n, \eta_{n\tau}))$ is a diagonal matrix containing the weights. Since the weights also depend on the current estimates, an iteration loop is required to obtain the final estimates. Still, the approach already shows that the asymmetrically weighted quadratic loss fits well with the class of quadratic penalties

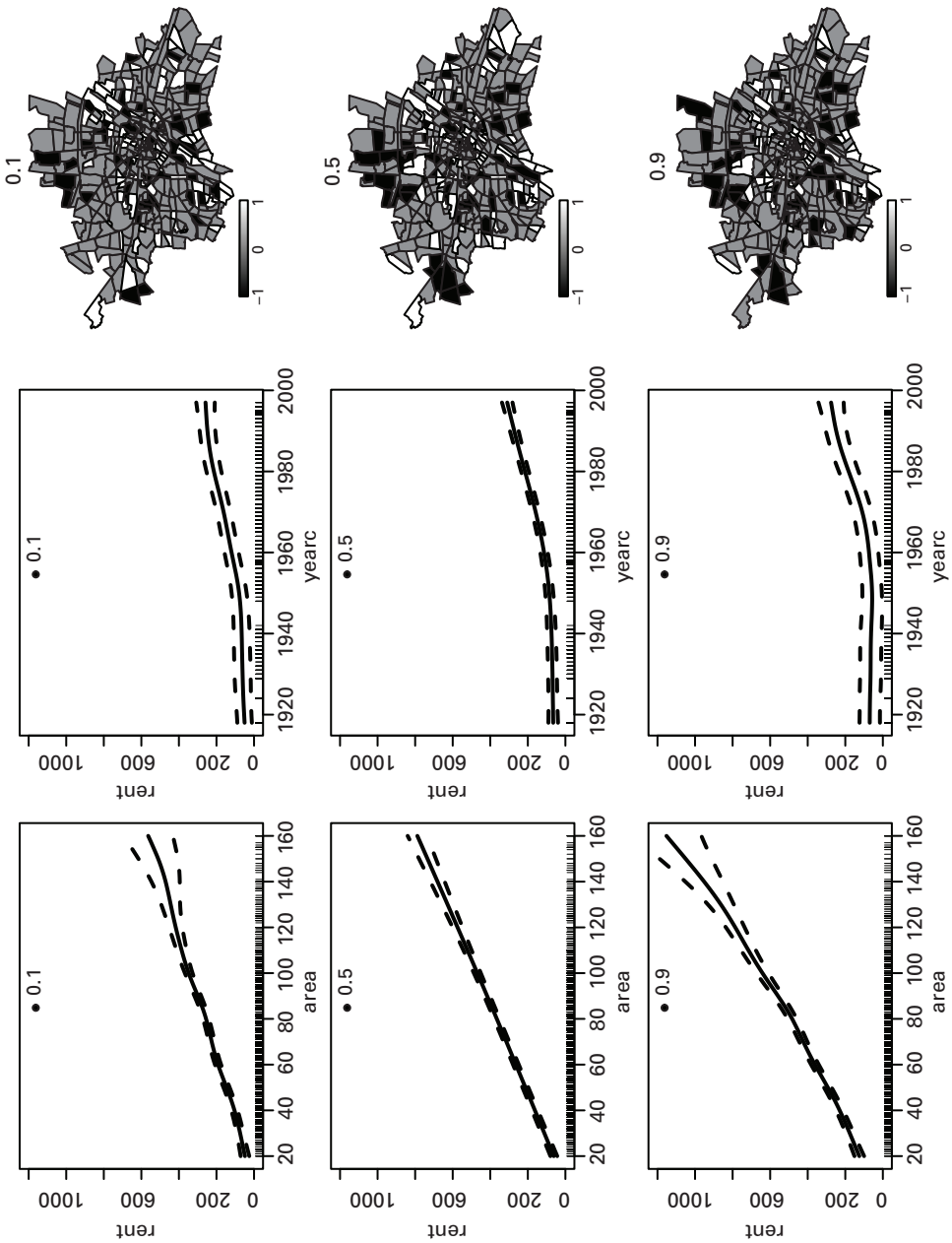

Figure 5 Estimated nonparametric and spatial effects in a geoadditive expectile regression model for the Munich rental guide with asymmetries $\tau = 0.1$ (top panel), $\tau = 0.5$ (middle panel) and $\tau = 0.9$ (bottom panel). For the estimated nonparametric effects, 95% pointwise confidence intervals have been included. For the estimated spatial effect, significance maps code the negative or positive significance of region-wise effects

we are considering. Expectile regression also enables the incorporation of smoothing parameter selection, for example by making use of the mixed model representation of penalized regression, see Sobotka and Kneib (2012) and Schnabel and Eilers (2009) for details.

Similarly as with Bayesian quantile regression, an asymmetric normal distribution can be considered in Bayesian expectile regression but this avenue has not yet been explored in detail (most probably since estimation via iteratively weighted least squares already allows for data-driven determination of smoothing parameters). Still it would have the advantage that some flexible model extensions that can be easily included in Bayesian inference (such as LASSO regularization or Dirichlet process priors for random effects) could then also be employed in expectile regression.

A boosting approach for expectile regression, on the other hand, is very easy to derive since in this case the gradients are given by

$$u_i = \begin{cases} \tau |y_i - \hat{\eta}_i^{[m-1]}|, & y_i > \hat{\eta}_i^{[m-1]}, \\ 0, & y_i = \hat{\eta}_i^{[m-1]}, \\ (\tau - 1)|y_i - \hat{\eta}_i^{[m-1]}|, & y_i < \hat{\eta}_i^{[m-1]}. \end{cases}$$

Both direct optimization and boosting for expectile regression are available in the R-package **expectreg** (Sobotka *et al.*, 2012b).

We utilize the direct optimization approach to re-estimate the geoadditive model for the rental guide example that we have studied in the context of Bayesian quantile regression in the last section. Figure 5 shows estimated effects for three selected asymmetries τ reflecting the lower, central and upper part of the rent distribution. The results are basically in line with those from the Bayesian quantile regression, indicating again that similar information can be acquired from expectile regression as with quantile regression.

7 Summary and conclusions

This paper should convey two basic messages: (i) there is more than mean regression and (ii) for models beyond mean regression there are more flexible alternatives than simple parametric predictor specifications. Actually, both extensions of the classical regression situation fit together very nicely in many situations and provide a flexible, convenient framework for applied analyses of complex regression data.

When choosing a specific model type to work with, the basic distinction is between 'complete distribution models' provided by GAMLSS that fully specify the distribution of the response and the 'distribution free' approaches provided by quantile regression and expectile regression without relying on specific distributional assumptions. To decide between these two branches, one may consider the following question: Does the main interest in the analysis lie on specific quantiles of the response? In this case, quantile and expectile regression may be more attractive since they directly target the quantity of interest and avoid restrictive assumptions for the

responses/error terms. If, on the other hand, one is mainly interested in understanding changes in the complete distribution of the response given covariates, the fully integrated models provided by GAMLSS will usually be more attractive since they provide a coherent, comprehensive description of the response distribution. Moreover, GAMLSS may be advantageous in less informative situations, e.g., in case of comparably small sample sizes.

The decision between quantile and expectile regression is less clear-cut. Most people would probably argue that quantile regression is preferable due to the easier interpretation of quantiles. However, as Waltrup *et al.* (2012) have shown, expectiles can be easily transformed to calculate quantiles and may then also be more efficient. In addition, inference for expectiles in models with complex semiparametric predictors is at present better developed than for classical quantile regression. Still, the connection between the asymmetric Laplace distribution and quantile regression (and in particular the location-scale mixture representation) makes Bayesian quantile regression a promising alternative.

One further type of regression models that goes beyond mean regression and has purposely been left out of the comparison is modal regression that estimates conditional modes of the response distribution. Einbeck and Tutz (2006) provide an example for nonparametric estimation in the bivariate scatterplot smoothing setup relying on kernel smoothing. While modal regression is interesting from an applied perspective as an alternative to mixture models for multi-modal response distributions, inference is at the moment limited to rather simple regression specifications and therefore not yet readily combined with the full complexity of semiparametric regression.

Acknowledgements

This paper summarizes joint work with Nora Fenske, Claudia Flexeder, Benjamin Hofner, Torsten Hothorn, Göran Kauermann, Stefan Lang, Andreas Mayr, Matthias Schmid, Linda Schulze Waltrup, Fabian Sobotka, Elisabeth Waldmann and Yu Ryan Yue. It has been a real pleasure working with all of you! The title for the paper is borrowed from Ludwig Fahrmeir who is also responsible for stimulating my interest in regression beyond the mean. Ludwig Fahrmeir and Fabian Sobotka read early versions of this paper and gave me a lot of constructive feedback. I am very grateful to Brian Marx and Jeff Simonoff who initially suggested to turn a presentation from the International Workshop on Statistical Modelling in Prague 2012 into this discussion paper. Financial support by the German Research Foundation (DFG), grant KN 922/4-1, is gratefully acknowledged.

References

Belitz C, Brezger A, Kneib T, Lang S and Umlauf N (2012) *BayesX - Software for Bayesian inference in structured additive regression models. Version 2.1* http://www.bayesx.org/.

Brezger A and Lang S (2006) Generalized additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, **50**, 967–91.

Bühlmann P and Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Cottet R, Kohn RJ and Nott DJ (2008) Variable selection and model averaging in semiparametric overdispersed generalized linear models. *Journal of American Statistical Association*, **103**, 661–71.

Dette H and Volgushev S (2008) Non-crossing non-parametric estimates of quantile curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 609–27.

Dunn PK and Smyth GK (1996) Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–44.

Einbeck J and Tutz G (2006) Modelling beyond regression functions: an application of multimodal regression to speedflow data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **55**, 461–75.

Fahrmeir L and Kneib T (2009) Propriety of posteriors in structured additive regression models: theory and empirical evidence. *Journal of Statistical Planning and Inference*, **39**, 843–59.

Fahrmeir L and Kneib T (2011) *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford: Oxford University Press.

Fahrmeir L, Kneib T and Lang S (2004) Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731–61.

Fenske N, Kneib T and Hothorn T (2011) Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, **106**, 494–510.

Friedman HH, Friedman LW and Amoo T (2002) Using humor in the introductory statistics course. *Journal of Statistics Education*, **10**.

Gamerman D (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57–68.

Green PJ (1987) Penalized likelihood for general semi-parametric regression models. *International Statistical Review*, **55**, 245–59.

Hastie T and Tibshirani R (1993) Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, **55**, 757–96.

Hastie TJ and Tibshirani RJ (1990) *Generalized additive models*. Boca Raton: Chapman & Hall / CRC.

Hofner B, Mayr A, Fenske N and Schmid M (2011) *gamboostLSS: Boosting Methods for GAMLSS Models* http://CRAN.R-project.org/package=gamboostLSS. R package version 1.0-1.

Hofner B, Hothorn T, Schmid M and Kneib T (2012) A framework for unbiased model selection based on boosting. *Journal of Computational and Graphical Statistics*, **20**, 956–71.

Hothorn T, Buehlmann P, Kneib T, Schmid M and Hofner B (2011) *mboost: Model-Based Boosting* http://CRAN.R-project.org/package=mboost. R package version 2.0-11.

Kamman EE and Wand MP (2003) Geoadditive models. *Applied Statistics*, **52**, 1–18.

Kneib T, Hothorn T and Tutz G (2009) Variable selection and model choice in

geoadditive regression. *Biometrics*, **65**, 626–34.
Koenker R (2005) *Quantile regression*. Economic Society Monographs. Cambridge University Press, New York.
Koenker R (2011) *quantreg: Quantile Regression* http://CRAN.R-project.org/package=quantreg. R package version 4.71.
Koenker R and Bassett G (1978) Regression quantiles. *Econometrica*, **46**, 33–50.
Koenker R and Mizera I (2004) Penalized triograms: Total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B*, **66**, 145–63.
Koenker R, Ng P and Portnoy S (1994) Quantile smoothing splines. *Biometrika*, **81**, 673–80.
Kottas A and Krnjajic M (2009) Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, **36**, 297–19.
Kuan C-M, Yeh J-H and Hsu Y-C (2008) Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics*, **150**, 261–70.
Mayr A, Fenske N, Hofner B, Kneib T and Schmid M (2012) GAMLSS for high-dimensional data–A flexible approach based on boosting. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **61**, 403–27.
Newey WK and Powell JL (1987) Asymmetric least squares estimation and testing. *Econometrica*, **55**, 819–47.
Reich BJ, Bondell HD and Wang H (2010) Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, **11**, 337–52.
Rigby RA and Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–54.
Rue H, Martino S and Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, **71**, 319–92.
Ruppert D, Wand MP and Carroll RJ (2003) *Semiparametric regression*. New York: Cambridge University Press.
Schnabel SK and Eilers P (2009) Optimal expectile smoothing. *Computational Statistics & Data Analysis*, **53**, 4168–77.
Schnabel SK and Eilers P (2013) Simultaneous estimation of quantile curves using quantile sheets. *AStA Advances in Statistical Analysis*, **97**, 77–87.
Schnabel SK and Eilers P (2011) Expectile sheets for joint estimation of expectile curves. *Technical Report*.
Sobotka F and Kneib T (2012) Geoadditive expectile regression. *Computational Statistics and Data Analysis*, **56**, 755–67.
Sobotka F, Kauermann G, Schulze Waltrup L and Kneib T (2013) On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, **23**, 135–48.
Sobotka F, Schnabel S and Schulze Waltrup L (2012) *expectreg: Expectile and Quantile Regression* http://CRAN.R-project.org/package=expectreg. R package version 0.35.
Speed T (1991) Comment on Robinson (1991) "That BLUP is a good thing: The estimation of random effects". *Statistical Science*, **6**, 42–44.
Stasinopoulos M and Rigby B (2011) *gamlss: Generalized Additive Models for Location Scale and Shape* http://www.gamlss.org. R package version 4.0-8.
Taddy MA and Kottas A (2010) A bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics*, **28**, 357–69.
Taylor JW (2008) Estimating value at risk and expected shortfall using expectiles.

Journal of Financial Econometrics, **6**, 231–52.

Waldmann E, Kneib T, Yue YR, Lang S and Flexeder C (2013) Bayesian semiparametric additive quantile regression. *Statistical Modelling*, **13**, 223–52.

Schulze Waltrup L, Sobotka F, Kneib T and Kauermann G (2012) Quantile or expectile regression—is there a favorite? *Technical report*.

Wood SN (2006) *Generalized additive models: An introduction with R*. Boca Raton: Chapman & Hall/CRC.

Yu K and Moyeed RA (2001) Bayesian quantile regression. *Statistics & Probability Letters*, **54**, 437–47.

Yue Y and Rue H (2011) Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis*, **55**, 84–96.