Statistical Modelling 2013; **13**(**3**): 223–252

Bayesian semiparametric additive quantile regression

Elisabeth Waldmann[1], Thomas Kneib[1], Yu Ryan Yue[2], Stefan Lang[3] and Claudia Flexeder[4]
[1]Department of Economics, Georg August University Göttingen, Germany
[2]Zicklin School of Business, Baruch College, City University of New York, USA
[3]Department of Statistics, University of Innsbruck, Austria
[4]Institute of Epidemiology I, HelmholtzZentrum München, Germany

Abstract: Quantile regression provides a convenient framework for analyzing the impact of covariates on the complete conditional distribution of a response variable instead of only the mean. While frequentist treatments of quantile regression are typically completely nonparametric, a Bayesian formulation relies on assuming the asymmetric Laplace distribution as auxiliary error distribution that yields posterior modes equivalent to frequentist estimates. In this paper, we utilize a location-scale mixture of normals representation of the asymmetric Laplace distribution to transfer different flexible modelling concepts from Gaussian mean regression to Bayesian semiparametric quantile regression. In particular, we will consider high-dimensional geoadditive models comprising LASSO regularization priors and mixed models with potentially non-normal random effects distribution modeled via a Dirichlet process mixture. These extensions are illustrated using two large-scale applications on net rents in Munich and longitudinal measurements on obesity among children. The impact of the likelihood misspecification that underlies the Bayesian formulation of quantile regression is studied in terms of simulations.

Key words: Quantile Regression; Geodditive Regression; MCMC; LASSO Regularization; Dirichlet Process

Received February 2012; **revised** January 2013; **accepted** February 2013

1 Introduction

Quantile regression allows to determine the influence of covariates on the conditional quantiles of the distribution of a dependent variable. Therefore, one of the main advantages over mean regression is that quantile regression permits to supply detailed information about the complete conditional distribution instead of only the mean. In addition, outliers and extreme data are usually less influential in quantile regression due to the inherent robustness of quantiles.

Address for correspondence: Elisabeth Waldmann, Chair of Statistics, Platz der Göttinger Sieben 5, 37073, Göttinger, Germany. E-mail: ewaldma@uni-goettingen.de

 10.1177/1471082X13480650

For classical linear quantile regression as introduced by Koenker and Bassett (1978), estimation of the quantile-specific regression coefficients $\boldsymbol{\beta}_\tau$ relies on minimizing the sum of asymmetrically weighted absolute deviations (AWADs)

$$\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau),$$

where $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ are the observed response and covariate values for n observations, the check function

$$\rho_\tau(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau) = \begin{cases} \tau|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau| & \text{if } y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}_\tau \\ (1-\tau)|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_\tau| & \text{if } y_i < \boldsymbol{x}_i'\boldsymbol{\beta}_\tau \end{cases}$$

defines asymmetrically weighted absolute residuals and $\tau \in (0, 1)$ is the quantile of interest. This approach is completely nonparametric and does not require the assumption of a specific response distribution. No closed form solution for the minimization problem exists and quantile regression estimates are typically obtained based on linear programming (see Koenker, 2005, for details).

Recent interest in quantile regression has focused on broadening the scope of supported model specifications. For example, additive quantile regression models have gained considerable attention. Oh *et al.* (2011) propose differentiable approximations to the AWAD criterion that allow to employ different types of smoothing approaches while Li *et al.* (2010) and Wu and Liu (2009) show ways to incorporate regularization on fixed effects into linear quantile regression, the first one in the Bayesian context, the second in a frequentist framework. Fenske *et al.* (2011) propose boosting approaches for flexible, additive quantile regression models, where penalized least squares estimates are utilized as base-learners. Koenker *et al.* (1994) added an L_1-norm penalty to the AWAD criterion that allows to still use linear programming techniques in the context of quantile smoothing splines. Koenker and Mizera (2004) extend this approach to surface estimation based on triograms.

In this paper, we will introduce yet more flexible types of quantile regression models motivated by two large-scale applications on rents for flats in the city of Munich and on longitudinal childhood growth measurements. The German tenancy law puts restrictions on the increase of rents and forces landlords to keep the price in a range defined by flats which are comparable in size, location and quality. To make it easier for tenants and owners to assess if the rent is appropriate for a flat, so-called rental guides are derived based on large samples of flats. In the following, we will use data from the 2007 Munich rental guide with about 3000 observations and 250 covariates. Kneib *et al.* (2011) suggested a high-dimensional geoadditive model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + f_1(\text{size}_i) + f_2(\text{year}_i) + f_{\text{spat}}(s_i) + \varepsilon_i \tag{1.1}$$

for analyzing the *expectation* of the net rent per square metre y_i in terms of nonlinear effects f_1 and f_2 of the size of the flat in square metres and the year of construction, a spatial effect f_{spat} based on district information s_i and a high-dimensional

vector of mostly categorical covariates $\boldsymbol{x}_i$ (such as presence of a fridge, attic, garden or balcony) with linear effects $\boldsymbol{\beta}$. While penalized splines and a Gaussian Markov random field have been employed for the nonlinear and spatial effects, respectively, least absolute shrinkage and selection operator (LASSO) and ridge penalization have been applied to the vector $\boldsymbol{\beta}$ to achieve regularization. It turned out that, for mean regression, a geoadditive model with LASSO regularization outperforms a model of moderate dimension resulting from expert knowledge and has slight advantages over a comparable model with ridge regularization. We therefore aim at extending the high-dimensional geoadditive model to quantile regression, to enable a more detailed view on the conditional distribution of the net rents. In particular, quantile regression allows to determine flexible bounds for the net rent based on, for example, the 5% and the 95% quantiles without imposing strong assumptions on the error distribution.

The second application deals with longitudinal measurements on the natural course of growth of children in the LISA (Influences of *L*ife-style factors on the development of the *I*mmune *S*ystem and *A*llergies in East and West Germany) study. The objective of the study is to analyze the variations in individual body mass index (BMI) patterns while simultaneously determining the impact of factors driving the growth of children, such as the breast-feeding behaviour or maternal BMI. Since there is considerable variation in the highly nonlinear individual profiles, Heinzl *et al.* (2012) suggest a flexible additive mixed model

$$y_{it} = f(t) + \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{z}_{it}'\boldsymbol{b}_i + \varepsilon_{it}, \tag{1.2}$$

where $t = 1, \ldots, T_i$, denotes the time, $i = 1, \ldots, n$, the individual, $f(t)$ represents the overall trend in the BMI measurements, $\boldsymbol{x}_{it}'\boldsymbol{\beta}$ contains parametric, fixed effects common to all children and the random effects term $\boldsymbol{z}_{it}'\boldsymbol{b}_i$ contains individual-specific deviations from the overall trend. Since there appear to be different groups of children with specific patterns in their individual-specific deviations and to account for potential non-normality of the random effects distribution, Heinzl *et al.* (2012) utilized a Dirichlet process mixture (DPM) in mean regression as random effects distribution which allows for very flexible random effects distributions and model-based clustering of the random effects. The presented data set was also already treated by Mayr (2010), who introduced a boosting approach to model quantile-based prediction intervals and Fenske *et al.* (2008), who also used boosting for additive quantile models. Random effects in quantile regression were, for example, treated in Kim and Yang (2011). In this paper, we show a combination of the mentioned works, which includes the clustering features of the DPMs as well as the fact that the Bayesian framework renders available credible intervals for the parameters. The choice of additional arguments, such as smoothing parameters, is conducted automatically in the Bayesian context.

In summary, our aim is to make flexible components in semiparametric regression models, such as nonlinear effects, spatial effects, LASSO regularized coefficient blocks or non-normal random effects, applicable in the context of quantile regression. These are typically difficult to combine with linear programming or other direct

maximization approaches. Instead, we rely on a Bayesian formulation of quantile regression based on the asymmetric Laplace distribution as an auxiliary error distribution as suggested in Yu and Moyeed (2001). Therefore, we consider the alternative representation of the quantile regression problem as

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

where $\eta_{i,\tau}$ is the predictor of the τth quantile in the regression model and $\varepsilon_{i,\tau}$ is an appropriate error term. Instead of assuming zero mean for the errors as in mean regression, one then imposes the restriction that the τth-quantile of the error distribution is zero. In the Bayesian framework, we have to assume a specific distribution for the errors (or equivalently the responses) to be able to set up a likelihood. The asymmetric Laplace distribution $y_i \sim \text{ALD}(\eta_{i,\tau}, \delta^2, \tau)$ with location parameter $\eta_{i,\tau}$, precision parameter δ^2, asymmetry τ and density

$$p(y_i|\eta_{i,\tau}, \delta^2, \tau) = \tau(1-\tau)\delta^2 \exp(-\delta^2 \rho_\tau(y_i - \eta_{i,\tau})) \tag{1.3}$$

is particularly useful since it yields posterior mode estimates that are equivalent to the minimizers of the AWAD criterion. Obviously, the assumption of an asymmetric Laplace distribution for the error terms will usually be a misspecification such that we actually abuse the asymmetric Laplace distribution likelihood to make quantile regression accessible in a Bayesian formulation. In addition to the heuristic argument that posterior modes with the asymmetric Laplace distribution coincide with the AWAD minimizers and that, asymptotically, the posterior mean obtained with Markov chain Monte Carlo (MCMC) simulations is equivalent to the posterior mode, we will study the impact of the misspecification in terms of a simulation study. It will turn out that the point estimates obtained from Bayesian quantile regression are usually close to their frequentist analogues and to the true effects while confidence intervals have to be interpreted with care especially for extreme quantiles.

To actually make Bayesian inference for the asymmetric Laplace distribution computationally feasible, Kozumi and Kobayashi (2011) and Reed and Yu (2009) introduced a location-scale mixture representation that allows to rewrite Bayesian quantile regression as a conditionally Gaussian regression with offset and weights. As a consequence, Bayesian inferential schemes developed for Gaussian regression models can then (at least conceptually) be easily transferred to quantile regression. Amongst others Yue and Rue (2011) and Lum and Gelfand (2012) use this reparametrization. However, Yue and Rue (2011) observed severe mixing and convergence problems in their approach to sampling-based Bayesian quantile regression and therefore had to resort to an approximate solution based on integrated nested Laplace approximations. We propose a different updating scheme (based on basis coefficients of nonparametric effects instead of function evaluations) that overcomes the difficulties encountered by Yue and Rue (2011). Moreover, we embed Bayesian semiparametric quantile regression in a generic framework that enables the flexible inclusion of hyperprior structures on the variance (and mean) parameters of the conditional Gaussian priors of the regression effects. Such hyperprior structures can be

used to re-cast extensions of semiparametric regression such as LASSO regularization or DPMs in the context of conditionally Gaussian Bayesian quantile regression. Some work on regularization in quantile regression context has also been done by Alhamzawi *et al.* (2012), Li and Zhu (2008), Wang *et al.* (2007) and Alhamzawi and Yu (2013).

The rest of this paper is organized as follows: In Section 2, we first introduce the location-scale mixture representation of the asymmetric Laplace distribution and present a generic MCMC simulation algorithm for Bayesian quantile regression with conditionally Gaussian priors. In a simulation study, we compare Bayesian additive quantile regression to frequentist total variation penalization splines to assess the impact of the misspecified likelihood in the Bayesian formulation. Afterwards, we introduce different special cases of the generic Bayesian model and the corresponding hyperprior specifications. Section 3 presents the applications based on a high-dimensional geoadditive regression model in case of the Munich rental guide and a nonparametric random effects model for the longitudinal growth measurements.

2 Bayesian semiparametric quantile regression

2.1 Generic Bayesian quantile regression with auxiliary error distribution

While the asymmetric Laplace distribution (1.3) provides a convenient way to express quantile regression in a Bayesian framework based on an auxiliary error distribution, it complicates inference based on MCMC simulations due to the inherent non-differentiability of the check function ρ_τ. We therefore follow Yue and Rue (2011) and utilize a scale mixture of Gaussians representation of the asymmetric Laplace distribution. Let $Z \sim \mathrm{N}(0, 1)$ and $W \sim \mathrm{Exp}(\delta^2)$ be two independent random variables following a standard normal and exponential distribution with rate parameter δ^2, respectively. Then

$$Y = \eta + \xi W + \sigma Z\sqrt{\frac{W}{\delta^2}}$$

with $\xi = \frac{1-2\tau}{\tau(1-\tau)}$ and $\sigma^2 = \frac{2}{\tau(1-\tau)}$ follows the $\mathrm{ALD}(\eta, \delta^2, \tau)$ distribution. As a consequence, the Bayesian quantile regression problem can be reformulated as a conditionally Gaussian regression with offsets ξW and weights $\sigma\sqrt{W/\delta^2}$ after imputing W as a part of the MCMC sampler.

To be more specific, we assume that n independent realizations $y_i \sim \mathrm{ALD}(\eta_i, \delta^2, \tau)$ are given with generic semiparametric predictor

$$\eta_i = \sum_{j=1}^{J} f_j(\boldsymbol{v}_i).$$

The predictor comprises various functions f_j that are defined on the complete vector of covariates $\boldsymbol{v}_i$. For example, specific components may be given by (*i*) *linear functions* $f_j(\boldsymbol{v}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is a subvector of $\boldsymbol{v}_i$, (*ii*) *univariate nonlinear functions* $f_j(\boldsymbol{v}_i) = f(x_i)$ where x_i is a single continuous element of $\boldsymbol{v}_i$, (*iii*) *spatial effects* $f_j(\boldsymbol{v}_i) = f_{\text{spat}}(s_i)$ where s_i is a spatial location variable or (*iv*) *random effects* $f_j(\boldsymbol{v}_i) = x_i \boldsymbol{b}_{c_i}$, where x_i is some covariate (potentially including a constant for random intercepts) and c_i is a cluster variable that groups the observations; see Fahrmeir *et al.* (2004) or Kneib *et al.* (2009) for similar generic model specifications.

In matrix notation, we can always write the generic model as

$$\eta = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \ldots + \boldsymbol{Z}_J\boldsymbol{\gamma}_J$$

where the design matrices $\boldsymbol{Z}_j$ are obtained by suitable basis expansions and $\boldsymbol{\gamma}_j$ contain the corresponding basis coefficients.

Our assumptions imply the observation model

$$\boldsymbol{y} \,|\, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J, \boldsymbol{w}, \delta^2 \sim \mathrm{N}(\boldsymbol{\eta} + \xi\boldsymbol{w}, \sigma^2/\delta^2\boldsymbol{D})$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ are the vectors of response observations and predictors, respectively, $\boldsymbol{w} = (w_1, \ldots, w_n)'$ is the vector of i.i.d. $\mathrm{Exp}(\delta^2)$ distributed weights implied by the scale mixture, and $\boldsymbol{D} = \mathrm{diag}(w_1, \ldots, w_n)$ is a corresponding diagonal matrix of the weights.

In order to enforce specific properties of the basis coefficients such as (e.g., spatial) smoothness, we assume conditionally Gaussian, possibly partially improper priors

$$\boldsymbol{\gamma}_j \,|\, \boldsymbol{m}_j, \boldsymbol{\theta}_j, \delta^2 \propto \exp\left(-\frac{1}{2}(\boldsymbol{\gamma}_j - \boldsymbol{m}_j)'\boldsymbol{K}_j(\boldsymbol{\theta}_j)(\boldsymbol{\gamma}_j - \boldsymbol{m}_j)\right) \tag{2.1}$$

where the prior precision $\boldsymbol{K}_j(\boldsymbol{\theta}_j)$ may depend on a vector of further hyperparameters $\boldsymbol{\theta}_j$ for which additional hyperpriors have to be defined depending on the specific type of effect. A term $\boldsymbol{f} = \boldsymbol{Z}\boldsymbol{\gamma}$ (e.g., P-spline, LASSO component) is then specified by defining

- a design matrix $\boldsymbol{Z}$,
- a precision or penalty matrix $\boldsymbol{K}(\boldsymbol{\theta})$,
- a prior $p(\boldsymbol{m})$ for $\boldsymbol{m}$,
- a prior $p(\boldsymbol{\theta})$ for the hyperparameter(s) $\boldsymbol{\theta}$.

We will give specific examples in Sections 2.2–2.4.

Since the location-scale mixture representation and the conditionally Gaussian prior structure yield a conjugate model hierarchy, the full conditionals for the regression coefficients are again Gaussian $\boldsymbol{\gamma}_j|\cdot \sim \mathrm{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with expectation and

covariance matrix

$$\begin{aligned}\boldsymbol{\mu}_j &= \boldsymbol{\Sigma}_j^{-1}\left(\tfrac{\delta^2}{\sigma^2}\boldsymbol{Z}_j'\boldsymbol{D}^{-1}(\boldsymbol{y}-\xi\boldsymbol{w}-\boldsymbol{\eta}_{-j})+\boldsymbol{K}_j(\boldsymbol{\theta}_j)\boldsymbol{m}_j\right),\\ \boldsymbol{\Sigma}_j &= \left(\boldsymbol{K}_j(\boldsymbol{\theta}_j)+\tfrac{\delta^2}{\sigma^2}\boldsymbol{Z}_j'\boldsymbol{D}^{-1}\boldsymbol{Z}_j\right)^{-1},\end{aligned} \tag{2.2}$$

where $\boldsymbol{\eta}_{-j}=\boldsymbol{\eta}-\boldsymbol{Z}_j\boldsymbol{\gamma}_j$ is the partial predictor without the jth effect. When comparing the full conditionals with those arising from mean regression, where

$$\boldsymbol{\mu}_j=\delta^2\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{Z}_j'(\boldsymbol{y}-\boldsymbol{\eta}_{-j})+\boldsymbol{K}_j(\boldsymbol{\theta}_j)\boldsymbol{m}_j)\qquad\text{and}\qquad\boldsymbol{\Sigma}_j=(\boldsymbol{K}_j(\boldsymbol{\theta}_j)+\delta^2\boldsymbol{Z}_j'\boldsymbol{Z}_j)^{-1}$$

we find only minor differences corresponding basically to the imputed weights and the offset.

Resulting from the scale mixture representation of the asymmetric Laplace distribution, the weights w_i are a priori i.i.d. exponentially distributed given the prior precision δ^2, i.e., $w_i \mid \delta^2 \sim \text{Exp}(\delta^2)$. This implies that the full conditionals for imputing the inverse of the weights are inverse Gaussian:

$$w_i^{-1}|\cdot \sim \text{InvGauss}\left(\sqrt{\frac{\xi^2+2\sigma^2}{(y_i-\eta_i)^2}},\ \frac{\delta^2(\xi^2+2\sigma^2)}{\sigma^2}\right). \tag{2.3}$$

If the prior for the precision parameter δ^2 is chosen to be the conjugate gamma distribution $\text{Ga}(a_0, b_0)$, the resulting full conditional is also gamma:

$$\delta^2|\cdot \sim \text{Ga}\left(a_0+\frac{3n}{2},\ b_0+\frac{1}{2\sigma^2}\sum_{i=1}^{n}w_i^{-1}(y_i-\eta_i-\xi w_i)^2+\sum_{i=1}^{n}w_i\right). \tag{2.4}$$

Since the prior for the weights also depends on δ^2, these are also part of the updated gamma parameters in the full conditional, yielding a slight change compared to the corresponding full conditional in mean regression.

In summary, we obtain the prior structure shown in Figure 1 that induces the following algorithm for generic Bayesian quantile regression:

i. for $j = 1, \ldots, J$ sample $\boldsymbol{\gamma}_j$ from the Gaussian distribution with parameters (2.2),
ii. for $j = 1, \ldots, J$ sample $\boldsymbol{\theta}_j$ and $\boldsymbol{m}_j$ from the corresponding hyper full conditional (as detailled in the following sections),
iii. for $i = 1, \ldots, n$ sample w_i from the inverse Gaussian distribution (2.3),
iv. sample δ^2 from the gamma distribution (2.4).

Note that our MCMC sampler for geoadditive quantile regression differs from the one proposed in Yue and Rue (2011) in the update related to nonparametric effects. While Yue and Rue (2011) update the vector of function evaluations $\boldsymbol{f}_j$, our sampler

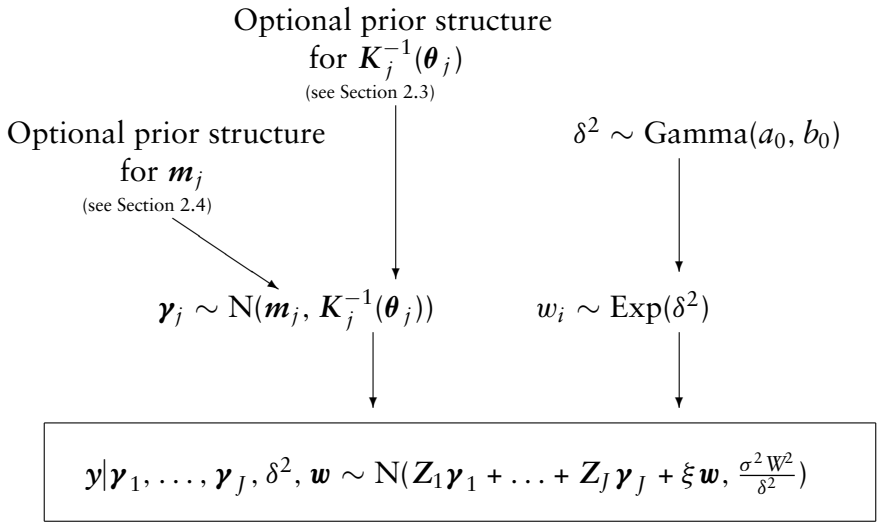

Figure 1 Structure of a simple quantile regression model

is based on the corresponding basis coefficients γ_j. This has two major advantages: On the one hand, the dimensionality of the parameter vector is considerably smaller, inducing a tremendous reduction in computing time. On the other hand, it avoids the severe mixing and convergence problems observed by Yue and Rue (2011) and therefore renders estimation of models with more than one nonparametric effect possible.

In the following sections, we present specific examples for modelling the functions f_j and updating the corresponding hyperparameters θ_j and m_j. For notational simplicity the index j will be suppressed, as there will be always a focus on one special class of effects.

2.2 Geoadditive quantile regression

Continuous covariates. For approximating potentially nonlinear effects, Bayesian P-splines can be used, see Eilers and Marx (1996) and Brezger and Lang (2006) for full details. Here the $n \times K$ design matrix Z is composed of B-spline basis functions evaluated at the observations x_i. Assuming a first or second order random walk for γ, i.e.,

$$\gamma_k \mid \gamma_{k-1}, \theta^2 \sim N\left(\gamma_{k-1}, \frac{1}{\theta^2}\right), \qquad k = 2, \ldots, K$$

or

$$\gamma_k \,|\, \gamma_{k-1}, \gamma_{k-2}, \theta^2 \sim N\left(2\gamma_{k-1} - \gamma_{k-2}, \frac{1}{\theta^2}\right), \qquad k = 3, \ldots, K$$

as smoothness prior with diffuse priors for initial values yields the penalty matrix $\boldsymbol{K}(\boldsymbol{\theta}) = \theta^2 \boldsymbol{R}'\boldsymbol{R}$ where $\boldsymbol{R}$ is a first or second order difference matrix. The prior also implies $\boldsymbol{m} = \boldsymbol{0}$. The vector of additional parameters $\boldsymbol{\theta}$ collapses to a single precision parameter θ^2 that governs the trade off between fidelity to the data and smoothness. The standard prior is $\theta^2 \sim \text{Ga}(a, b)$ implying the full conditional

$$\theta^2 \,|\, \cdot \sim \text{Ga}(a + 0.5\text{rank}(\boldsymbol{K}(\boldsymbol{\theta})), b + 0.5\boldsymbol{\gamma}'\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{\gamma}). \tag{2.5}$$

Spatial effects. For data observed on a regular or irregular lattice as in our case study on rents in Munich, a common approach for the spatial effect is based on Markov random fields (see Rue and Held, 2005). Let $s_i \in \{1, \ldots, K\}$ denote the spatial index or region of the ith observations. Then we assume $f_{\text{spat}}(s_i) = \gamma_{s_i}$, i.e., separate parameters $\gamma_1, \ldots, \gamma_K$ for each region are estimated. The $n \times K$ design matrix $\boldsymbol{Z}$ is an incidence matrix whose entry in the ith row and kth column is equal to one if observation i has been observed at location k and zero otherwise.

The most simple Markov random field prior for the regression coefficients γ_s is defined by

$$\gamma_s \,|\, \gamma_u, u \neq s, \theta \sim N\left(\sum_{u \in \partial_s} \frac{1}{N_s}\gamma_u, \frac{1}{N_s\theta^2}\right),$$

where N_s is the number of adjacent regions of s, and ∂_s denotes the regions which are neighbours of region s. This implies the penalty matrix $\boldsymbol{K}(\boldsymbol{\theta})$ with elements

$$\boldsymbol{K}(\boldsymbol{\theta})[k, u] = \frac{1}{\theta^2}\begin{cases} -1 & k \neq u, k \sim u, \\ 0 & k \neq u, k \nsim u, \\ |N(k)| & k = u \end{cases}$$

and $\boldsymbol{m} = \boldsymbol{0}$. Again the vector $\boldsymbol{\theta}$ of additional parameters collapses to a single precision parameter θ with gamma prior $\theta^2 \sim \text{Ga}(a, b)$ and corresponding full conditional (2.5).

Impact of the likelihood misspecification. As the ALD obviously is a misspecification of the true error structure, we conducted some simulations to analyze the validity of Bayesian credible intervals obtained with the auxiliary ALD error distribution in terms of coverage probabilities in nonlinear effects. The results were compared to the confidence intervals developed in Koenker (2011) for quantile regression smoothing splines with total variation penalization and implemented in function rqss of the R-package quantreg. We considered the following four model set-ups:

- M1: $y = 2 + 5\sin(2/3x) + \varepsilon$, $\varepsilon \sim N(0, .3 + (2x - 1)^2)$
- M2: $y = \sin(2(4x - 2)) + 2\exp((-16^2)(x - .5)^2) + \varepsilon$, $\varepsilon \sim N(0, .3 + (2x - 1)^2)$

- M3: $y = 5\sin(2/3x_1) + 1.5\log(x_2) + \varepsilon$, $\varepsilon \sim N(0, 1.5(x_1 - 1.5)^2 + 0.5)$
- M4: $y = \sin(2(4x - 2)) + 2\exp((-16^2)(x - .5)^2) + \varepsilon$, $\varepsilon \sim \text{Gamma}(4, 2/(3x))$

In all four situations, the error distributions are heteroscedastic. While the first two models represent situations with only one nonlinear effect and either rather low (M1) or high (M2) curvature of the quantile curves, the third model comprises two nonlinear effects with only one covariate affecting the variability. The fourth model has the same predictor as (M2), but the error is assigned a gamma distribution. As the effect of this change was rather small, we only show the results of (M2). Adding more effects made the computational time of **rqss** rise immensely, while for MCMC time rises only linearly with the number of effects.

For each model, we simulated 100 data sets with sample sizes $n = 500$ and $n = 750$. Since **rqss** does not allow to extrapolate outside the convex hull of observed covariate values in prediction, we evaluated the function estimates and their confidence intervals on a 400 point equidistant grid within the intervals of observed covariate values.

The most interesting results of this simulation exercise can be summarized as follows:

- For extreme quantiles in the outer range of the response distribution, confidence intervals obtained from Bayesian quantile regression tend to underestimate uncertainty while **rqss** tends to be right on average (see Figure 2 for an example in M1).
- For more central parts of the response distribution, the coverage rates are much closer to the nominal level, in particular if the functions to be estimated are smooth (see Figure 3). In the simulations, the results for 20% and 80% quantiles have been close enough to the nominal level to consider them a helpful tool in applied analyses.
- In contrast, confidence intervals obtained with **rqss** tend to be much too wide for central parts of the response distribution. In these situations, the confidence intervals are therefore much too conservative and the empirical coverage is very close to 100%. This overestimation of uncertainty is particularly expressed in case of M3 comprising two nonlinear effects (see Figure 4). In model M3, we also experienced some numerical problems when computing the confidence intervals within **rqss** that required manual fine-tuning of several optimization parameters.
- Both approaches have difficulties in reaching the nominal coverage level when functions have abrupt changes or peaks (see Figure 3).
- When considering single function estimates, **rqss** occasionally produced very wiggly confidence intervals (see Figure 2). While the coverage is still right on average, such confidence intervals still seem hard to justify in applied work. Although Figure 2 presents an exceptional example, this kind of estimation occurs from time to time especially for extreme quantiles.

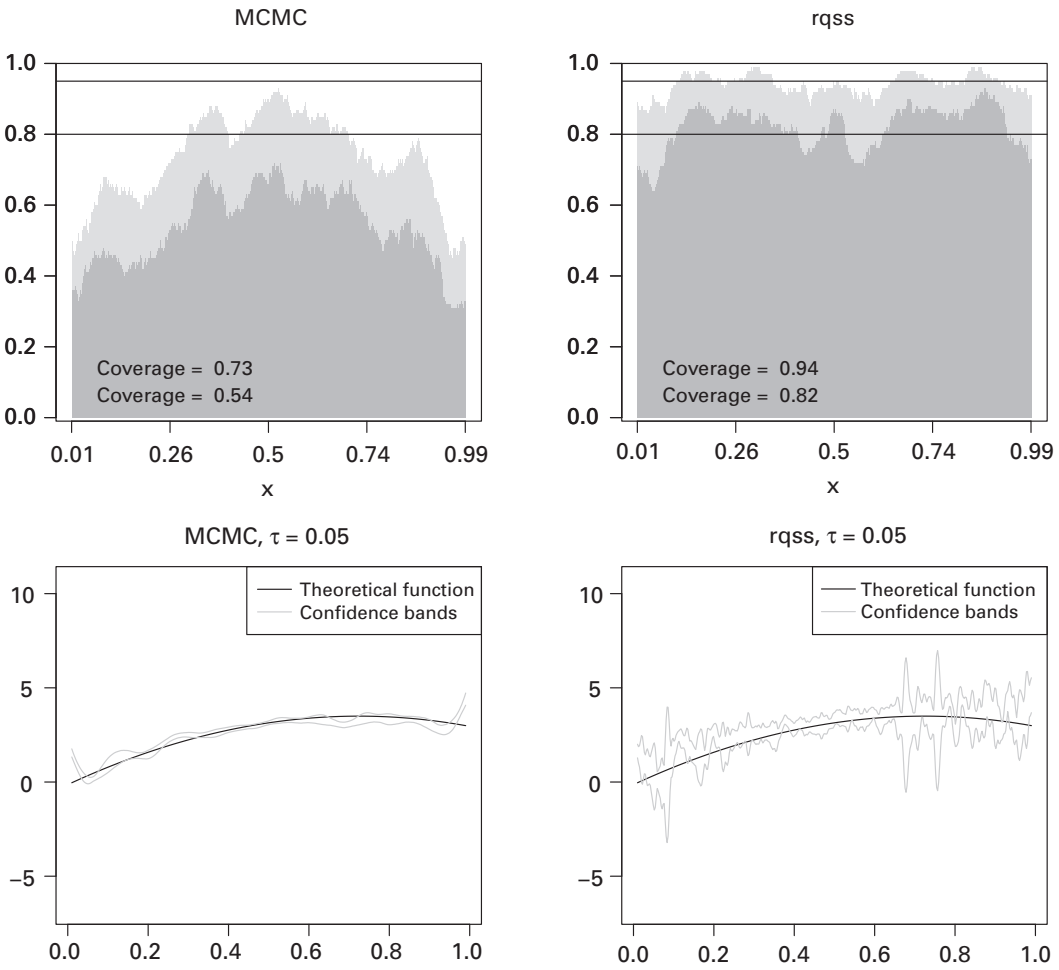

Figure 2 Upper panel: Coverage probabilities of the 80% and 95% intervals for the estimation of the 5% quantile of M1. The horizontal lines display the 80% and the 95% mark, and the dark grey and the light grey vertical lines the coverage over all 100 calculated models in each point x of the 80% and the 95% interval, respectively. Lower panel: Theoretical function with the two different confidence bands, M1

- We also repeated the analyses for Bayesian quantile regression and larger sample sizes (n = 2000) to study the changes due to increased information. However, the impact of varying sample size was rather small with some very moderate tendency to improved performance for the confidence intervals obtained with Bayesian quantile regression so that we do not present details here.
- The point estimates of the regression effects did not really differ for the two methods (see Figure 5).

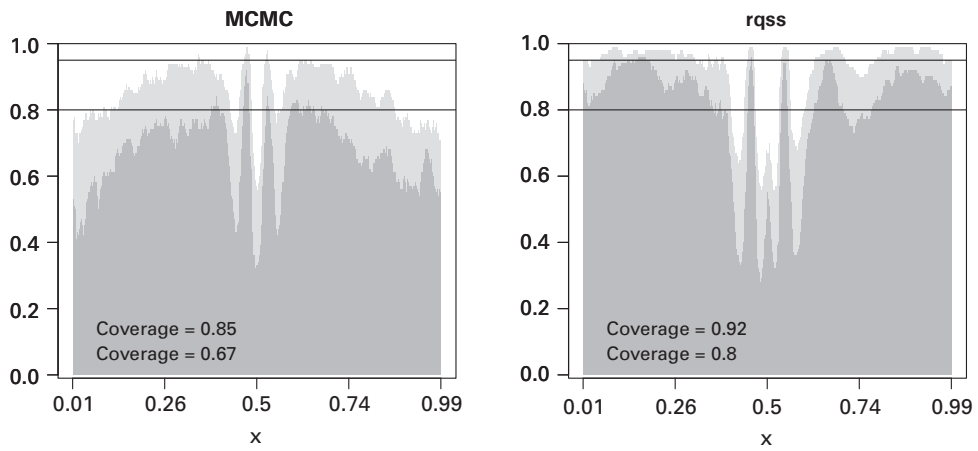

Figure 3 Coverage probabilities of the 80% and 95% intervals for the estimation of the 80% quantile of M2. The horizontal lines display the 80% and the 95% mark, and the dark grey and the light grey vertical lines the coverage over all 100 calculated models in each point x of the 80% and the 95% interval, respectively

We also calculated simultaneous confidence bands in the Bayesian set-up as suggested by Krivobokova *et al.* (2010). These led to broad intervals, close to those estimated by rqss and are therefore not described in detail here.

In summary, although the confidence bands obtained with Bayesian quantile regression are based on a misspecified model, they still provide a reasonable reflection of estimation uncertainty provided that the considered quantiles are not too extreme. One explanation for the deteriorated behaviour for extreme quantiles is the increasing asymmetry of the asymmetric Laplace distribution for very large or small values of the quantile τ. This asymmetry has the consequence that the spike at the corresponding quantile gets larger and larger which will usually be in contradiction to the true data likelihood. In contrast, for central quantiles, the asymmetric Laplace distribution gets more and more symmetric and fits better to the asymptotically expected normal likelihood in case of not too small sample sizes.

2.3 LASSO regularization

In this section, we will show how the concept of the LASSO (Tibshirani, 1996) can be adapted to Bayesian quantile regression by specifying suitable hyperpriors for $\boldsymbol{\theta}$ following the ideas presented in Park and Casella (2008). Suppose that the regression coefficients in a K-dimensional vector $\boldsymbol{\gamma}$ shall be subject to LASSO regularization. Then a prior structure that yields posterior mode estimates which can be interpreted

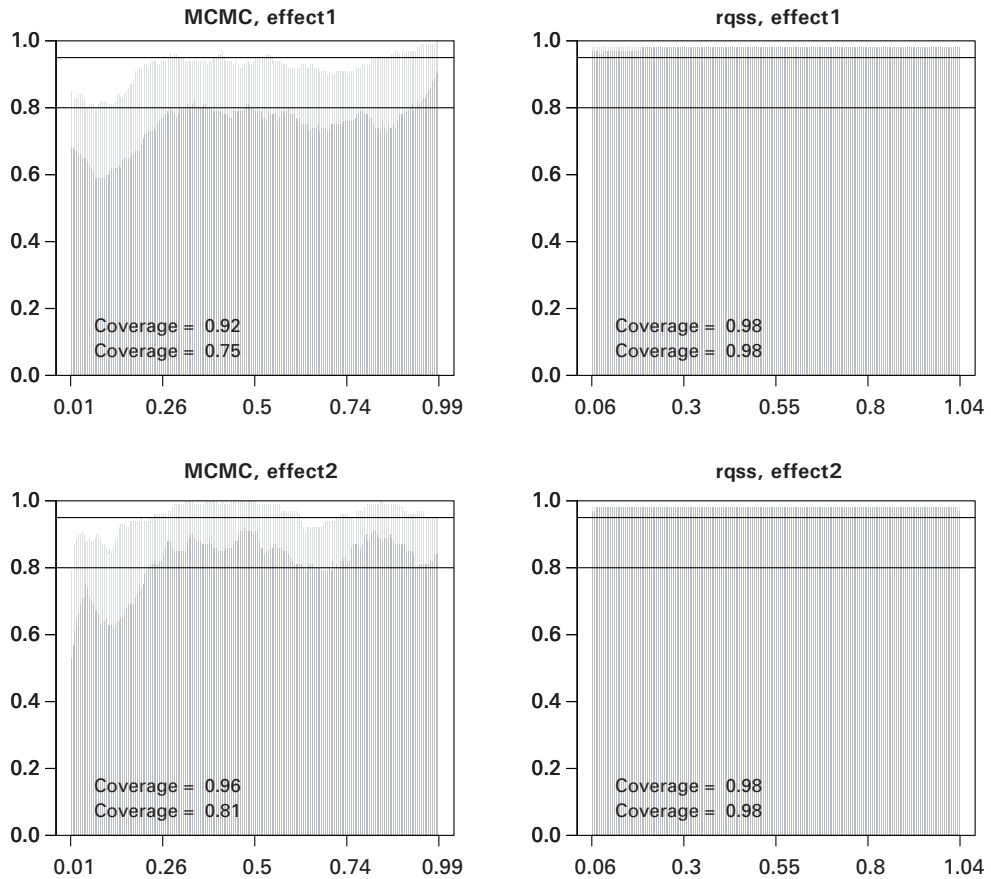

Figure 4 Coverage probabilities of the 80% and the 95% intervals for the estimation of the 50% quantile of the first effect (upper panel) and second effect (lower panel) of the model (M3). The horizontal lines display the 80% and the 95% mark, and the dark grey vertical lines the coverage of the 80% interval and the light grey vertical lines the coverage of the 95% interval over all 100 calculated models in each point x

as the Bayesian analogue to LASSO-regularized penalized maximum likelihood estimates is given by the Laplace prior

$$p(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \lambda^2 \exp\left(-\lambda|\gamma_k|\right),$$

To enable the inclusion of Bayesian LASSO regularization in the basic geoadditive quantile regression sampler outlined in the previous section, we again rely on a scale-mixture representation of the Laplace prior yielding $\boldsymbol{\gamma}|\boldsymbol{\theta} \sim N(0, \boldsymbol{K}^{-1}(\boldsymbol{\theta}) = \text{diag}(1/\theta_1^2, \ldots, 1/\theta_K^2))$ and hyperpriors $\theta_k \sim \text{Exp}(\lambda^2)$, $k = 1, \ldots, K$ and $\lambda^2 \sim \text{Ga}(a, b)$.

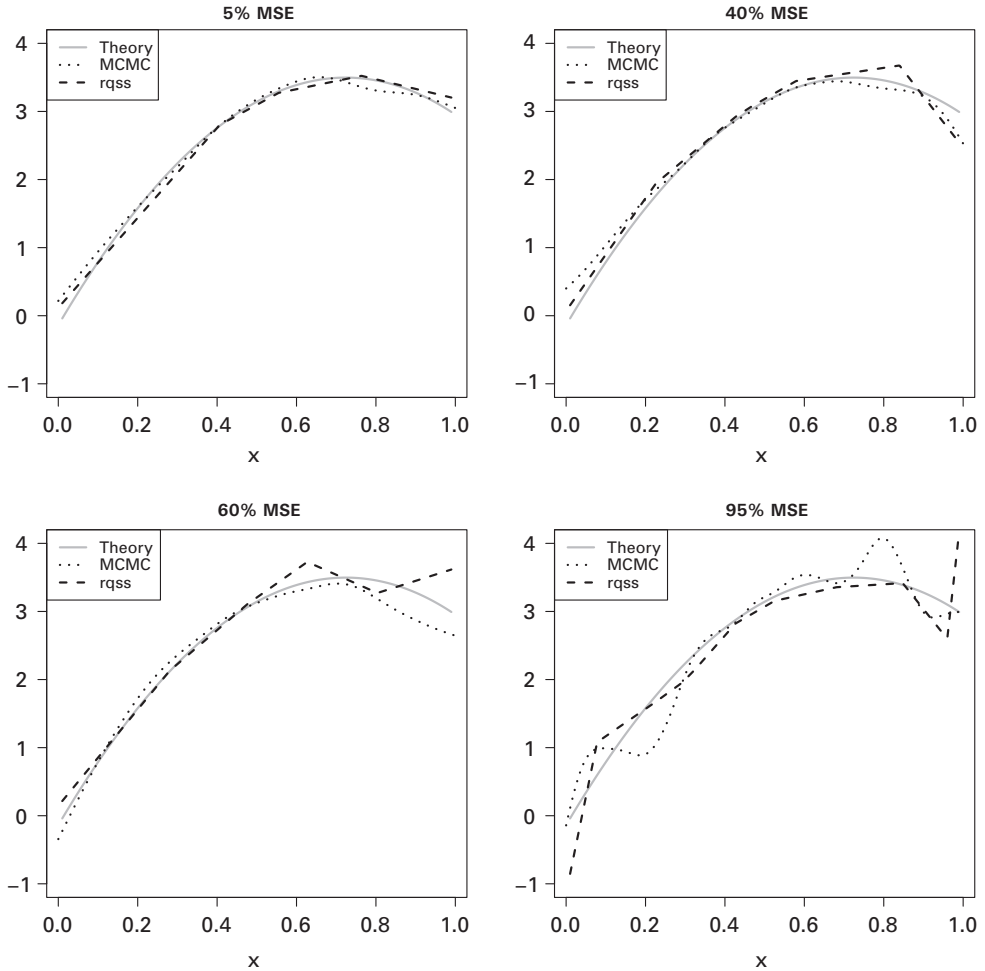

Figure 5 Estimates of the 5% quantile for (M1) chosen according to their quantile in the MSE distribution obtained from the 100 simulation replications

The full conditionals for the LASSO-specific parameters are

$$\theta_k^{-2}|\cdot \sim \text{InvGauss}\left(\frac{|\lambda|}{\gamma_k}, \lambda^2\right) \quad \text{and} \quad \lambda^2|\cdot \sim \text{Ga}\left(a + K, b + 0.5\sum_{k=1}^{K} 1/\theta_k^2\right).$$

According to Park and Casella (2008) the LASSO prior specified so far may result in a multimodal posterior. We can avoid this problem with the scale-dependent prior $\boldsymbol{\gamma}|\boldsymbol{\theta} \sim N(0, \frac{1}{\delta^2}\boldsymbol{K}^{-1}(\boldsymbol{\theta}) = \text{diag}(1/\theta_1^2, \ldots, 1/\theta_K^2))$ where the prior covariance is scaled by

the inverse of δ^2. Then all full conditionals have to be slightly modified accordingly in analogy to Park and Casella (2008).

2.4 Dirichlet process mixtures for random effects

A further extension of the predictor for Bayesian quantile regression results when considering random effects with potentially non-normal random effects distribution specified via a DPM prior (see Ghosh and Ramamoorthi, 2010, for a recent review). The latter allows to specify a hyperprior on the space of all random effects distributions while simultaneously enabling model-based clustering of the random effects to retrieve groups of observations with similar random effects profiles. In the following $\boldsymbol{\gamma}_i$ will denote the random effects vector for individual i, referred to as $\boldsymbol{b}_i$ in equation (1.2).

Our prior specification and implementation of DPMs is based on the stick-breaking representation introduced by Sethuraman (1994). Let G denote the random effects distribution and assume that a Dirichlet hyperprior is specified for G, i.e.,

$$\boldsymbol{\gamma}_i \sim G, \quad G \sim DP(\nu_0, G_0),$$

where G_0 is a base distribution and $\nu_0 > 0$ specifies a concentration parameter that determines a priori expected deviations of G from the base distribution G_0. Then it follows from the stick-breaking representation of Dirichlet processes that

$$G(\cdot) = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\phi}_k}(\cdot),$$

where $\delta_{\boldsymbol{\phi}_k}(\cdot)$ are Dirac measures (i.e., point masses) located at cluster-specific parameter vectors $\boldsymbol{\phi}_k$ drawn from a base distribution G_0, i.e.,

$$\boldsymbol{\phi}_k \overset{\text{i.i.d.}}{\sim} G_0,$$

independently from the (random) weights π_k. The weights are generated through the *stick-breaking process*

$$\pi_1 = v_1, \quad \pi_k = v_k \left(1 - \sum_{j=1}^{k-1}(1 - \pi_j)\right) = v_k \prod_{j=1}^{k-1}(1 - v_j), \quad k = 2, 3, \ldots,$$

with $v_k \overset{\text{i.i.d.}}{\sim} \text{Be}(1, \nu_0)$, where Be denotes the beta distribution. As a consequence, realizations of the Dirichlet process can be constructed as infinite mixtures of point masses at locations generated as i.i.d. draws from a base measure. The weights π_k are generated by first breaking the part $\pi_1 = v_1$ from a stick of the length $1 = \pi_1 + \pi_2 + \ldots$, then breaking off the part $\pi_2 = v_2(1 - \pi_1)$ from the remaining stick of length $1 - \pi_1$, and so on.

The stick-breaking representation of Dirichlet processes enables an intuitive interpretation since it may be considered an infinite extension of finite mixture models. However, it also reveals that realizations from a Dirichlet process are almost surely discrete with all probability mass concentrated on the locations $\boldsymbol{\phi}_k$, $k = 1, 2, \ldots$. To overcome this limitation, we do not specify the Dirichlet process directly for the random effects distribution but for the hyperparameters of this distribution, yielding DPMs with the following prior hierarchy:

$$\begin{aligned} \boldsymbol{\gamma}_i &\overset{\text{ind.}}{\sim} p(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i), \\ \boldsymbol{\phi}_i &\overset{\text{i.i.d.}}{\sim} G, \\ G &\sim DP(\nu_0, G_0). \end{aligned}$$

Here, the random effects are assumed to be realized independently from distributions $p(\boldsymbol{\gamma}_i|\boldsymbol{\phi}_i)$ with individual-specific parameters $\boldsymbol{\phi}_i$. These are generated according to a probability measure obtained from a Dirichlet process with concentration parameter ν_0 and base distribution G_0. Since the realization of the Dirichlet process is almost surely discrete, ties among the individual-specific parameters $\boldsymbol{\phi}_i$ will arise and therefore there will be groups of individuals sharing the same random effects distribution.

Specific choices we make for the prior specification in case of random effects are

$$\begin{aligned} \boldsymbol{\gamma}_i &\overset{\text{ind.}}{\sim} N(\boldsymbol{m}_i, \boldsymbol{K}(\boldsymbol{\theta})^{-1}), \\ \boldsymbol{m}_i &\overset{\text{i.i.d.}}{\sim} G, \\ G &\sim DP(\nu_0, G_0), \end{aligned}$$

i.e., the random effects are independent Gaussian distributed with a common covariance matrix $\boldsymbol{K}(\boldsymbol{\theta})^{-1}$ but differing means $\boldsymbol{m}_i$ following a Dirichlet process prior.

We will choose G_0 to be Gaussian and assign ν_0 a Gamma distribution. The hyper-prior structures of $\boldsymbol{\mu}_{\boldsymbol{\gamma}_i}$ and $\boldsymbol{K}(\boldsymbol{\theta})$ are visualized in Figure 6.

For the random effects $\boldsymbol{\gamma}_i$ we need to take into account the data distribution and the prior with the parameters generated via the DPM. The fact, that we are using DPMs does not change the equation in comparison to what we would get in a normal mixed model approach. The full conditional of $\boldsymbol{m}$ is of the same structure as (2.2).

As the stick-breaking process can obviously not be conducted to infinity, the standard approach is to truncate the process in a certain N—see Ishwaran and James (2001)—and only take into account the first N terms of the sum. In our example, N is chosen to be 100.

A common way to implement the DPM—also Ishwaran and James (2001)—is to introduce a vector of latent classification variables $\boldsymbol{c}$ of the length n, which consists in each iteration of the Gibbs sampler of m different values of $1, \ldots, n$. The subvector $\boldsymbol{c}^*$ denotes the vector comprising only the distinct values corresponding to non-empty

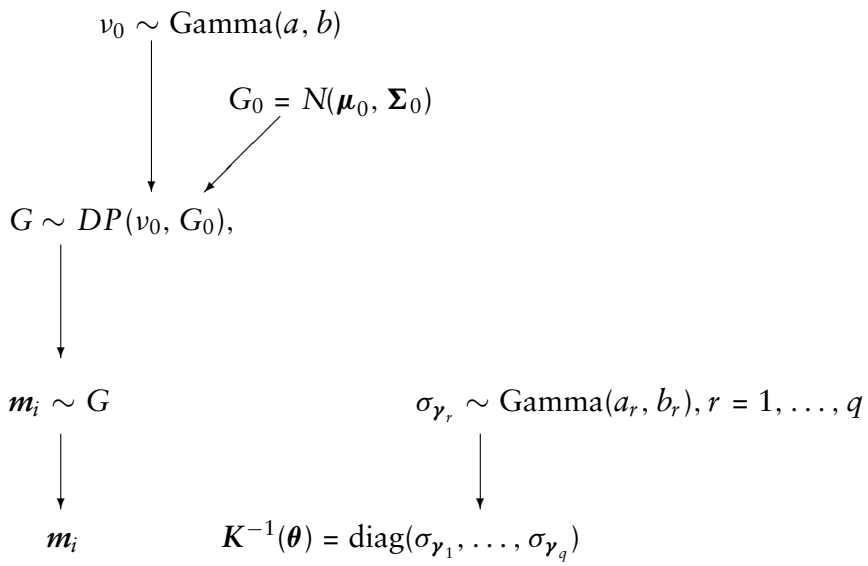

Figure 6 Hyperprior structure for the mean in the DPM context

clusters. The auxiliary variables ϕ_k ($k \in 1, \ldots, N$) are drawn from two different types of distributions depending on the fact if $k \in \boldsymbol{c}$ or not.

For the k which are not one of the different values in the set of c_i, ϕ_k is drawn from the base distribution:

$$\phi_k | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

If $k \in \boldsymbol{c}^*$, ϕ_k is drawn as follows:

$$\phi_k | \sigma^2_{\gamma_r}, \mu_{0_r}, \sigma^2_{0_r}, \boldsymbol{\gamma}, \boldsymbol{c} \sim N(\mu^*_{0_r}, \sigma^{2*}_{0_r}),$$

for the different k

$$\mu^*_{0_r} = \left(\frac{n_h}{\sigma^2_{\gamma_r}} + \frac{1}{\sigma^2_{0_r}} \right)^{-1} \left(\frac{n_h}{\sigma^2_{\gamma_r}} \bar{b}_{r,h} + \frac{\mu_{0_r}}{\sigma^2_{0r}} \right) \quad \text{and} \quad \sigma^{2*}_{0_r} = \left(\frac{n_h}{\sigma^2_{\gamma_r}} + \frac{1}{\sigma^2_{0_r}} \right)^{-1}.$$

The $\sigma^2_{0_r}, r = 1, \ldots, q$ are the diagonal elements of $\boldsymbol{\Sigma}_0$ and their priors are, again, gamma distributions. The latent classification variables are drawn from a mixture of the likelihood for the different ϕ, weighted with different π_k, which are drawn via the stick-breaking representation of the Dirichlet distribution:

$$c_i | \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{\gamma}_i, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \sim \sum_{h=1}^{N} \pi_k p(\boldsymbol{\gamma}_i | \phi_k, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) \delta_k(\cdot)$$

with $\delta_k(\cdot)$ being the Dirac measure in k.

As a next step, the π_k are constructed in a stick-breaking step using an auxiliary variable v_k, which is beta distributed, as already described. The π_k are then obtained through the above-mentioned product

$$\pi_k = v_k \prod_{l<k} (1 - v_l).$$

Finally the precision, which was assigned a gamma distribution as prior, for the DP is drawn from a gamma distribution

$$v_0|\boldsymbol{\pi} \sim \text{Gamma}\left(N - 1 + a_\alpha, b_\alpha - \sum_{h=0}^{N-1} \log(1 - V_k)\right).$$

The $\boldsymbol{m}_i$ themselves are estimated by using $\boldsymbol{m}_i = \phi_{c_i}$. As $\boldsymbol{c}$ contains only m different values (with $m \leq n$) the clustering mechanism follows directly from the construction.

It is obvious that, if this algorithm is repeated many times in an MCMC simulation, due to the randomness of each step we get different values for each individual and furthermore different clusters. Therefore, we will use a nearest neighbour approach to get a clustering which is valid over all iterations.

3 Applications

3.1 High-dimensional geoadditive regression for the Munich rental guide

As a first application of semiparametric Bayesian quantile regression, we consider the high-dimensional geoadditive regression model (1.1) but extended to conditional quantile specifications. We chose to model the 5%, 20%, 50%, 80% and 95% quantiles to give a detailed summary on both the central part of the distribution and the boundaries. This reveals information not only about the expected rent for a flat but also about the span, the rent is supposed to be in.

Due to the high dimensionality of the vector of parametric covariates (238 covariates in total), we will consider Bayesian LASSO regularization for all components (except the intercept). For the nonlinear effects of the size of the flat and the year of construction, we consider cubic P-splines with second order random walk prior and 20 equidistant knots. The spatial effect was calculated with a Markov random field where two subquarters were treated as neighbours if they share a common boundary. For all gamma type priors, we chose hyperparameters $a = b = 0.001$. The number of iterations for the MCMC sampler was fixed at 35000 with a burn in period of 5000 iterations and a thinning parameter of 30 (yielding 1000 samples for determining posterior means). A graphical analysis of mixing and convergence showed no inadequacies.

If we look at the results, we see that in fact most of the effects differ over the quantiles (see, e.g., the nonlinear effects in Figures 7 and 8). The pictures show the

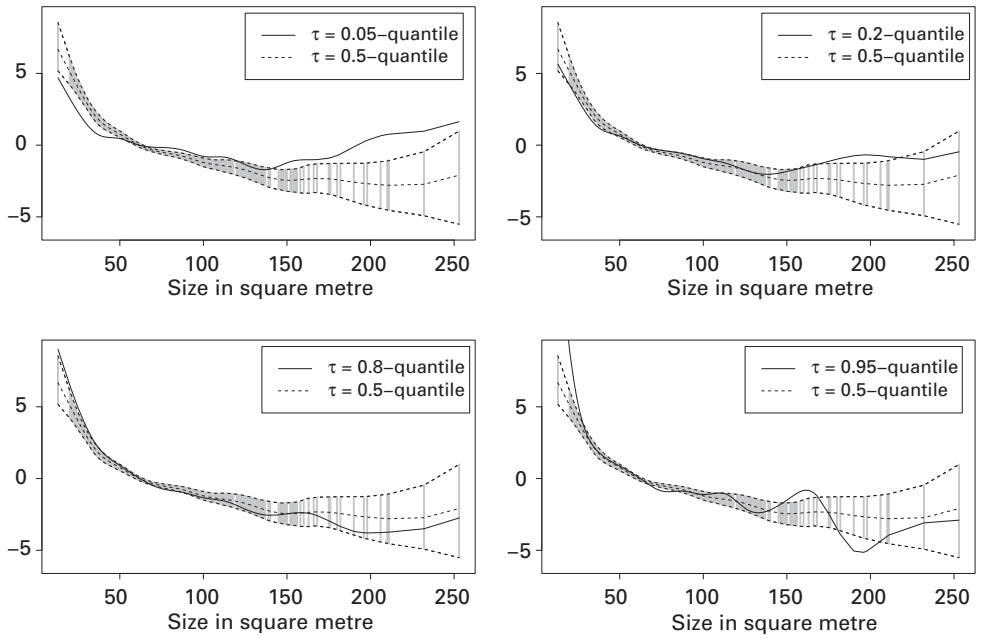

Figure 7 Munich rent index: nonlinear effect of size of the flat. Solid line: effect on the noncentral quantiles, dashed lines: effect on the median and 95%-posterior interval, light grey lines in background: concentration of data

centred curves for the noncentral quantiles in comparison to the posterior interval of the median regression. The grey stripes in the background indicate the concentration of the data for the corresponding values. For reasons of lucidity and because the figures have the aim to compare the effects, the posterior intervals are only plotted for the median regression. The first thing which has to be mentioned is that the posterior interval is broader in the parts with less observations. Another obvious fact is that smoothing works better for the central quantiles than for the extreme ones. This might be caused by the sparsity of data in these areas. As for the differences of the impact on the different quantiles we see a tendency of the functions to tilt over for both effects. The effect of the size of the flat seems to be less expressed for the lower quantiles, as especially in the low price segment the size of the flat does not really have any influence at all and increasing variability is observed as τ grows. It is the other way round for the year of construction: there is no pronounced effect of the age of the building on the highest quantile. Another effect which shows in the latter is that the positive impact of old buildings does not exist for lower quantiles.

In the spatial effect, the most noticeable fact is the higher number of significant effects for the outer quantiles (see Figure 9). While for the 95% quantile, 93 subquarters are selected to be significant, there are only 54 subquarters, which have significant impact on the median. Furthermore, we see a tendency towards more subquarters

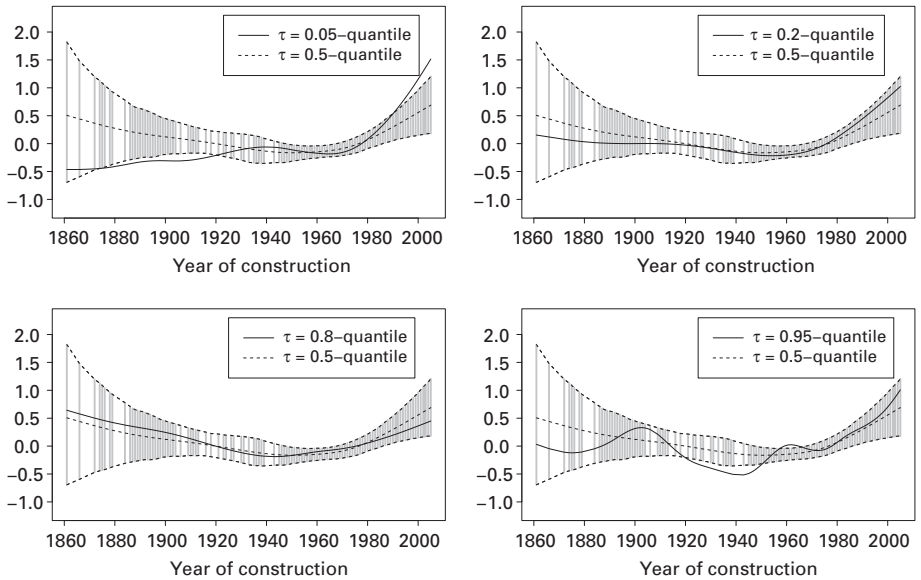

Figure 8 Munich rent index: nonlinear effect of year of construction. Solid line: effect on the noncentral quantiles, dashed lines: effect on the median and 95%-posterior interval, light grey lines in background: concentration of data

which have positive effect on the price in the higher quantiles (50 subquarters with significant positive influence versus 43 with significant negative influence in the 95% quantile) and the opposite effect for lower quantiles (52 subquarters with significant negative influence versus 29 with significant positive influence in the 5% quantile). The second fact we can see in the graphics is that, just as in the nonlinear effects, the estimation for the regression on the outer quantiles is less smooth than for the ones in the middle, indicating more variability in extreme quantiles.

Selection of the LASSO-regularized covariates with parametric effects via 95% posterior credible intervals led to a whole of 103 variables aggregated over all five models. The quantile for which the least effects were recognized as significant was the median with 47, while the highest number was 58 for $\tau = 0.95$. Analyzing posteriors of these parameters, different compartments can be detected. There are covariates, which show very similar effects over all five quantiles, while others are only selected for parts of them. In most of the cases, we can see the same direction of effect, while in some the sign changes over the different quantiles (see Figure 10 for some exemplary effects).

In some cases, the reason for covariates not to be selected is obviously the lack of data (e.g., the fact that *sauna*, just as high class facilities, is only selected to have an influence on the 95% quantile, while *no modernizations at all* only appears in the lowest quantile), while others just seem to be selected due to inner correlation. The

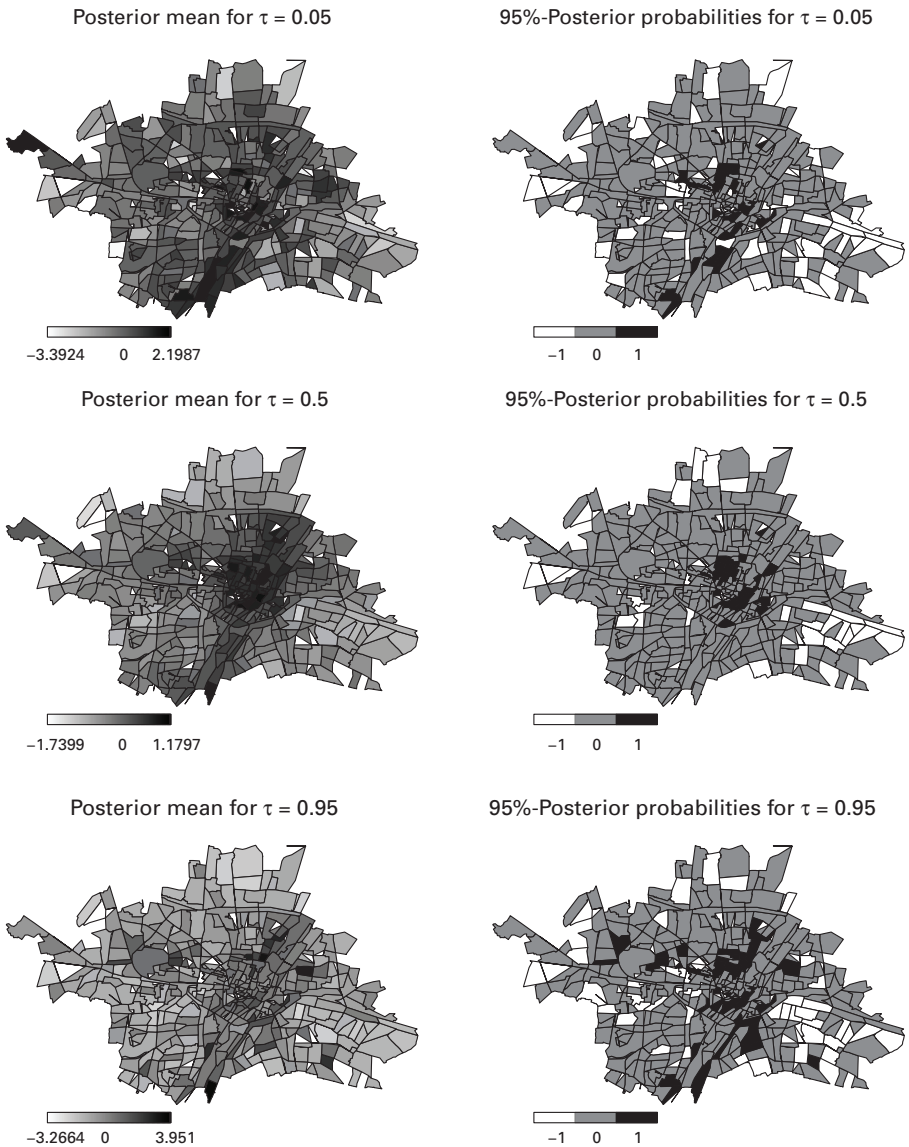

Figure 9 Munich rent index: spatial effects on the median (in the middle) and the outer quantiles (τ = 0.05 on the top and τ = 0.95 at the bottom). On the left: centralized effects, on the right: quarters with significant negative effect in white, quarters with significant positive effect in black and quarters with nonsignificant effects in grey

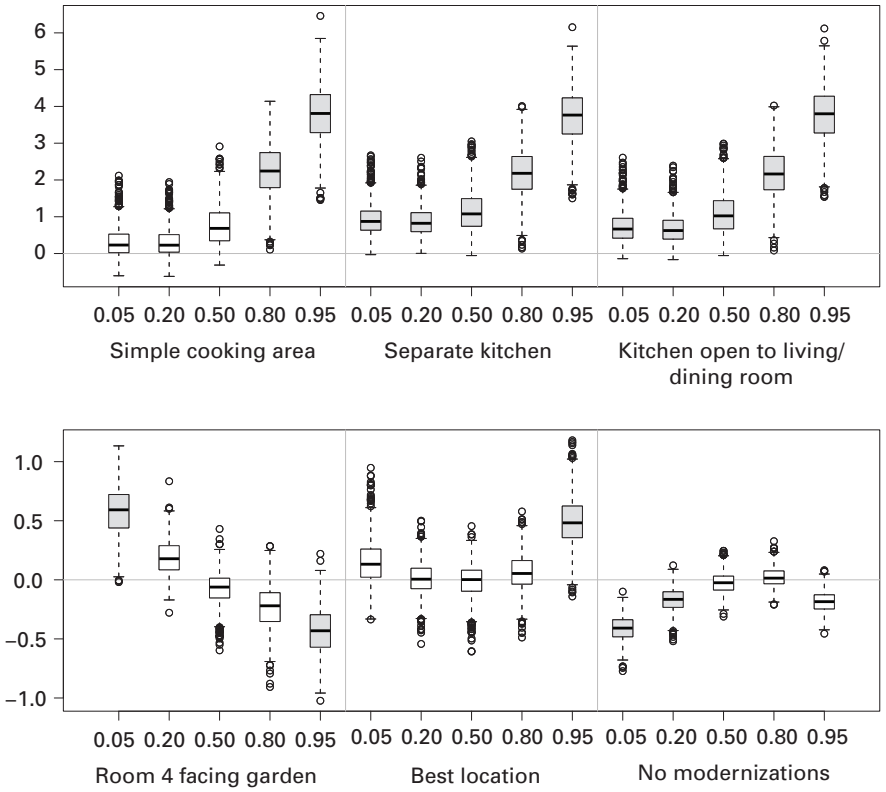

Figure 10 Munich rent index: examples for regularized linear effects with different behaviour in different quantiles; the grey boxes represent significant effects and the white ones nonsignificant effects

latter is not surprising as the LASSO is known for only selecting a few or even one representative for highly correlated covariates.

To compare the results of the high-dimensional geoadditive model with those of an expert model with a restricted set of covariates and with quantiles calculated from a Gaussian mean regression model results, we performed a ten-fold cross validation comparing the empirical risk on the test data. The quantiles for each flat were estimated as the quantiles of the distribution:

$$N(\hat{\eta}_i, \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}),$$

where $\hat{\eta}_i$ denotes the prediction for the ith flat and $\hat{\boldsymbol{\varepsilon}}$ is the vector of the residuals. Thus, the quantiles are obtained by the underlying Gaussian distribution, using the prediction as mean and the squared residuals as variance. The empirical risk is obtained by evaluating the quantile loss function $\rho(y_i - \eta_{i,\tau})$ at the posterior mean estimates in the test set. It turns out that the empirical risk is lower for the LASSO as compared to the expert model for nearly all quantiles; see Table 1 for average risks

Table 1 Munich rent index: mean risk for the high-dimensional geoadditive model, the expert model and the mean regression model, averaged over the ten cross validation folds

Quantile	LASSO Model	Expert Model	Mean Regression
$\tau = 0.05$	0.2285	0.2304	0.2146
$\tau = 0.20$	0.5043	0.5476	0.5516
$\tau = 0.50$	0.7100	0.7317	0.7339
$\tau = 0.80$	0.5005	0.5089	0.5168
$\tau = 0.95$	0.2301	0.2224	0.2039

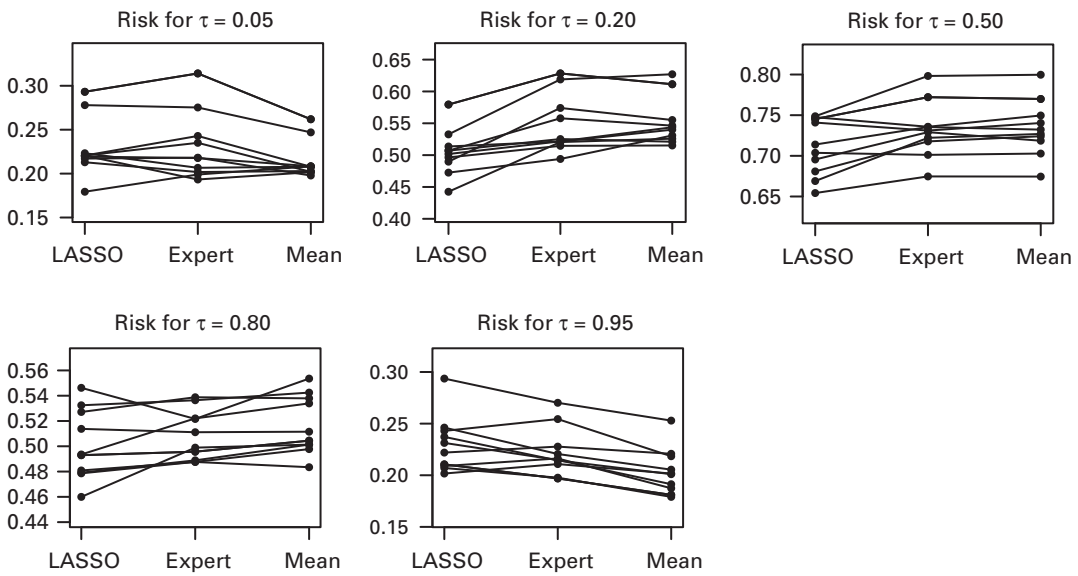

Figure 11 Munich rent index: parallel coordinate plots of risk functions for LASSO, expert and mean regression model

over the folds. In comparison to the mean regression model both quantile regression models, expert and LASSO, perform better. These results are also illustrated in Figure 11 in terms of a parallel coordinate plot.

Geoadditive quantile regression including LASSO regularization is implemented in BayesX (Lang *et al.*, 2005) and will be published in the next version of the software.

3.2 Nonparametric random effects for longitudinal childhood growth study

As an application involving the specification of individual-specific random effects, we consider the childhood growth data described in the introduction. These longitudinal

data were collected for 3097 healthy neonates over 60 months at nine mandatory medical examinations (birth, 2 weeks, 1, 3, 6, 12, 24, 48 and 60 months). We restricted attention to complete cases, yielding a final sample size of 2043 individuals. In order to make the outcomes comparable we used the BMI of the children as the dependent variable, even though it is not a perfect measurement of growth, because of the problems of measuring very young children's height. Collected covariates are *sex* (gender), *diet* (nutrition until the age of 4 months, 0 = bottle-fed or mixture of bottle-fed and breast-fed, 1 = breast-fed only), *mSmoke* (maternal smoking during pregnancy, 0 = no, 1 = yes), *area* (0 = rural: region of Wesel and Bad Honnef, 1 = urban: Munich and Leipzig), *ageY* (age in years), *mBMI* (maternal BMI at the beginning of pregnancy) and *mDiffBMI* (maternal BMI gain during pregnancy). The latter two variables were used in centered form.

We consider a quantile-specific version of model equation (1.2), where the temporal trend is specified as a cubic P-spline with 20 inner knots and second order random walk prior. The random effects comprise subject-specific intercepts, a random slope for a linear time trend and an additional random slope for the nonlinear time transformation $\log(t + 1)/(t + 1)^2$. The prior for the precision of the DPM was a gamma distribution with $a = 0.5$ and $b = 10$; the rest of the priors was handled in the same way as in the analysis of the Munich rental guide. The analysis was performed for the same five quantiles (5%, 20%, 50%, 80% and 95%) as in the Munich rental guide while the number of MCMC iterations had to be higher to achieve satisfactory results for mixing and convergence. Due to high autocorrelations especially in the estimation for the nonlinear trend function, we used a burn in period of 100 000 iterations, a thinning parameter of 200 and did a total of 300 000 iterations in order to obtain posteriors of the size of 1000 samples. Note that the mixing performance of all parameters (in particular the random effects) except the time trend was satisfactory already after a much smaller number of iterations.

To account for correlations arising from the longitudinal arrangement of the data while allowing for potential non-normality of the random effects distribution, we included random effects using the DPMs as explained in Section 2.4. Since we assume quantile-specific random effects, the random effects distribution and therefore the correlations may vary across the different quantiles considered. More specifically, when assuming a common random effect across quantiles, the quantile refers to a population quantile while quantile-specific random effects allow us to study individual-specific quantiles in a conditional model. One particular advantage of the latter approach is that with the DPM prior for random effects the clustering of individuals may vary with the chosen quantile. This is perfectly reasonable in practice since children may be similar in terms of, for example, the upper part of their BMI distribution but may differ with respect to the lower part of the distribution. In our application, the resulting number of clusters varies between 8 and 10 clusters for the different quantiles so that, in fact, children belong to different groups for different parts of the BMI distribution. This fact would not have been detected by using common random effect quantile regression, neither by a DPM mean regression model. The estimated random effects as well as the clustering of the children can be

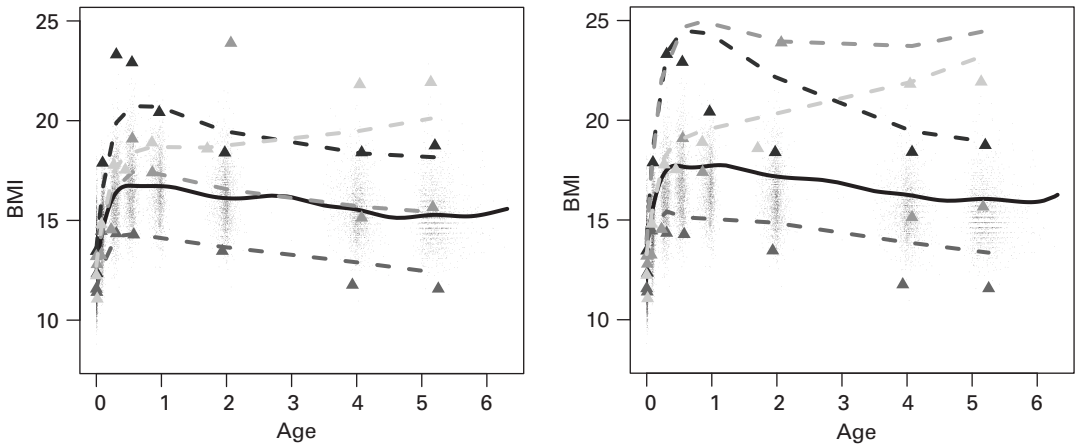

Figure 12 Growth study: effect of the age on the BMI in 50%-quantile (on the left) and 95%-quantile regression (on the right). Triangles: observations of four different individuals (one grey tone for each), dashed lines: estimations for the four different individuals (colours corresponding to the triangles), solid line: overall estimation

interpreted as reflecting the amount and structure of unobserved heterogeneity in the BMI distribution that cannot be captured by the covariates at hand.

The results for median regression are quite similar to those of mean regression obtained by Heinzl *et al*. (2012), with the obvious difference in robustness. Figure 12 shows the BMI conditional on time for four different individuals for median regression (on the left side) and 95%-quantile regression on the right. The observations are depicted by triangles in different tones of grey, the estimations in dashed lines in the corresponding colour and the overall estimation as a solid black line. While the outlier for the children shown in the lighter grey tones are ignored in the median regression, they are obviously taken into account in the curve on the right side. The same effect arises for the clustered version, which shows that the different quantiles as well as the different clusters are only affected by extreme values if the quantile we are looking at is an extreme one and the child with the extreme values belongs to the corresponding cluster.

The linear effects are presented in Table 2 and Figure 13. In the latter, significant effects, i.e., the ones which did not contain zero in their 95% posterior credible interval, are marked in grey. The most interesting result is the observation that breastfeeding have a significant effect on the lowest quantile only.

The calculated models were compared to two simpler models via the deviance information criterion (DIC). The DIC was calculated as

$$\mathrm{DIC} = 2 \sum_{b=1}^{B} \mathrm{dev}(\boldsymbol{\theta}_b) - \mathrm{dev}(\bar{\boldsymbol{\theta}}),$$

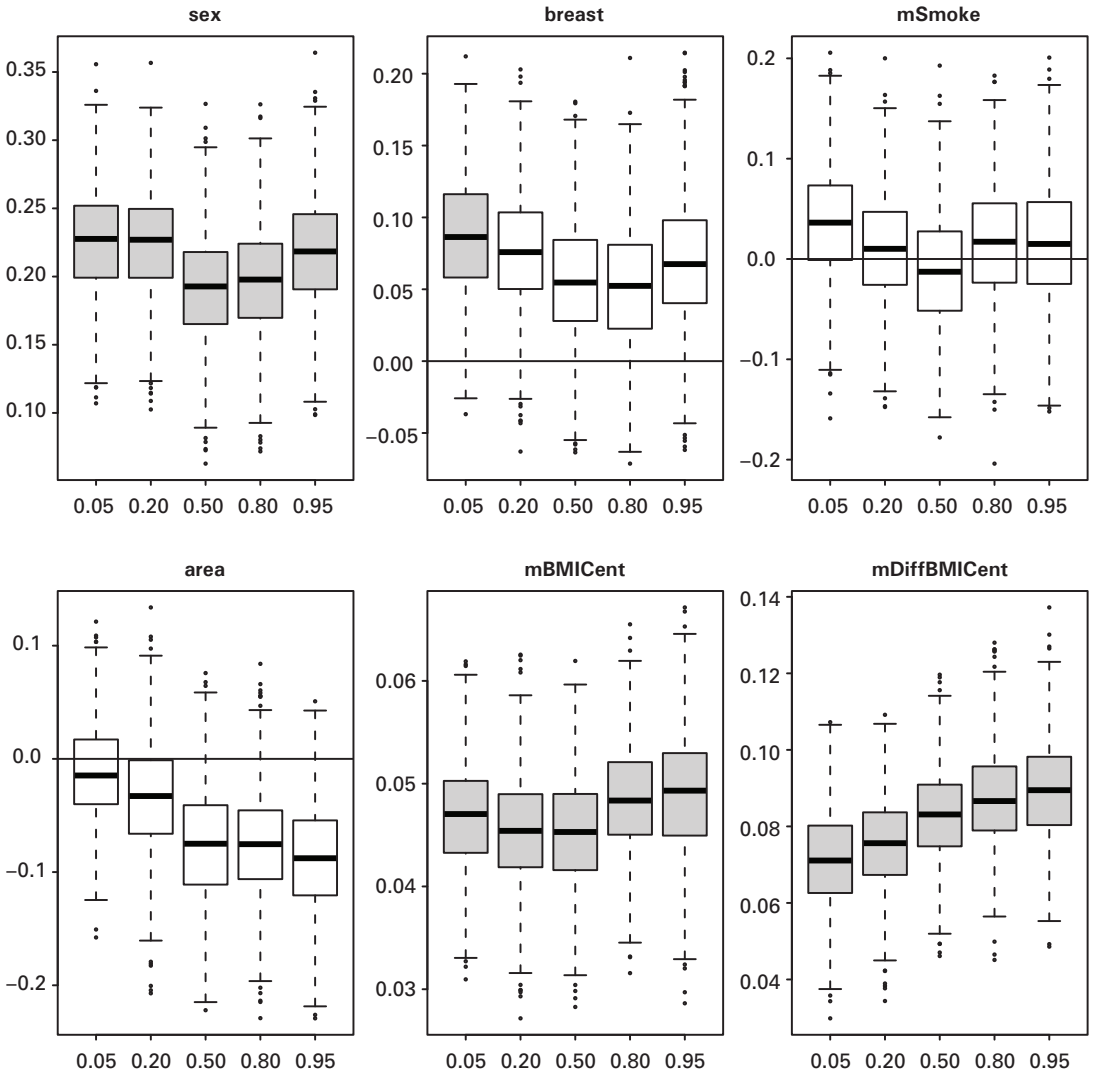

Figure 13 Growth study: boxplots of the different effects on the different quantiles; the grey boxes represent significant effects and the white ones nonsignificant effects

i.e., as the difference of two times the mean over the deviance in each sample and the deviance which is calculated with the sample mean of the parameters where b indexes the MCMC iterations. The DPM-Model outperformed a model without random effects with high difference in all quantiles (the mean difference was 6447.81). Using a normal Gaussian prior for the random effects performed worse than the more complex DPM-Model over all quantiles too (mean difference: 1141.272). For an

Table 2 Growth study: linear effects for the five different quantiles

	0.05	0.2	0.5	0.8	0.95
sex	0.2260	0.2254	0.1915	0.1977	0.2177
diet	0.0874	0.0752	0.0557	0.0526	0.0696
mSmoke	0.0367	0.0097	−0.0123	0.0166	0.0163
area	−0.0120	−0.0341	−0.0757	−0.0752	−0.0868
mBMI	0.0468	0.0454	0.0453	0.0485	0.0491
mDiffBMI	0.0712	0.0755	0.0828	0.0871	0.0895

Table 3 Growth study: DIC for different models

τ	DPM	Gaussian random effects	No random effects
0.05	15296.03	16586.32	23031.73
0.20	32957.19	34246.35	38377.08
0.50	34665.83	35774.21	39813.83
0.80	34379.04	35301.70	40103.93
0.95	16636.18	17732.05	24846.75

overview of the different DICs see Table 3. These results can, however, only be accepted with reservation, since the DIC calculations are based on the auxiliary assumption of the asymmetric Laplace distribution for the error terms.

The MCMC algorithm for the DPM random effects model was implemented in C++ and R. Just like the scheme itself the program was based on work by Heinzl *et al*. (2012).

4 Conclusions and discussion

The presented possibilities of modelling quantiles are useful extensions for the regression toolbox. Obviously the mixture representation of the asymmetric Laplace distribution allows for considerable flexibility and, in particular, other Bayesian approaches from mean regression can easily be incorporated. Especially the combination of different effects is rendered possible and seems to be nearly unlimited when using latent Gaussian model formulations. In future work, this concept could be transferred to variational approximations (Ormerod and Wand, 2010), which have the advantage to avoid simulation-based inference and therefore to reduce computation times.

One principal difficulty with Bayesian quantile regression is that the asymmetric Laplace distribution is only a working model that yields a likelihood misspecification. Therefore significance and uncertainty statements or quantities derived from the samples, like the DIC, have to be interpreted with care. We evaluated the impact of the likelihood misspecification in simulations and found that uncertainty in the estimates

is well represented for central quantiles while for extreme quantiles coverages of confidence intervals tend to be too small. On the other hand, for models of the complexity considered here, no alternative inferential principle is available so far so that we would still consider Bayesian quantile regression to be a valuable approach. In particular, point estimates derived with the misspecified likelihood will usually be very close to the corresponding frequentist estimates.

A possible alternative for applying asymmetric Laplace distribution for the error terms is to include the estimation of the error density in the MCMC algorithm, for example, via mixtures—see, for example, Kottas and Krnjajic (2009), Dunson and Taylor (2005) and Taddy and Kottas (2010). If these mixture approaches are also based on Gaussian mixture components, similar updating schemes as in this paper can be used if the mixture indicators are imputed as additional unknowns in the MCMC algorithm. However, the computational burden is still considerably higher due to the necessity to update the mixture component parameters and the mixture indicators in each iteration of the MCMC sampler.

Acknowledgements

Financial support from the German Research Foundation (DFG), grant KN 922/4-1 is gratefully acknowledged. We also thank Felix Heinzl for providing his code on DPMs in mean regression and Joachim Heinrich for sharing his expertise on the childhood growth study. The comments of two referees, an associated editor and Herwig Friedl led to a considerable improvement of the paper as compared to the original submission.

References

Alhamzawi R and Yu K (2013) Conjugate priors and variable selection for Bayesian quantile regression. *Computational Statistics & Data Analysis*.

Alhamzawi R, Yu K and Benoit D (2012) Bayesian adaptive LASSO quantile regression. *Statistical Modelling*, **12**, 279–97.

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

Dunson T and Taylor J (2005) Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics*, **17**, 385–400.

Eilers PHC and Marx BD (1996) Flexible smoothing using B-splines and penalized likelihood. *Statistical Science*, **11**, 89–121.

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, Fahrmeir L, Rzehak P and Höhle M (2008) Detection of risk factors for obesity in early childhood with quantile regression methods for longitudinal data. Technical Report No. 38, Department of Statistics, LMU München. Available at http://epub.ub.uni-muenchen.de/6260/

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.
Ghosh JK and Ramamoorthi RV (2010) *Bayesian Nonparametrics*. New York: Springer.
Heinzl F, Kneib T, and Fahrmeir L (2012) Additive mixed models with Dirichlet process mixture and P-spline priors. *Advances in Statistical Analysis*, **96**, 47–68.
Ishwaran H and James L (2001) Gibbs sampling methods for stick-breaking priors. *Journal of Computational and Graphical Statistics*, **11**, 508–32.
Kim M and Yang Y (2011) Semiparametric approach to a random effects quantile regression model. *Journal of the American Statistical Association*, **106**, 1405–17.
Kneib T, Hothorn T and Tutz G (2009) Variable selection and model choice in geoadditive regression models. *Biometrics*, **65**, 626–34.
Kneib T, Konrath S and Fahrmeir L (2011) High-dimensional structured additive regression models: Bayesian regularisation, smoothing and predictive performance. *Applied Statistics*, **60**, 51–70.
Koenker R (2005) *Quantile regression*. Econometric Society Monograph Series. Cambridge University Press.
Koenker R (2011) Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, **25**, 239–62.
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**(4), 673–80.
Kottas A and Krnjajic M (2009) Bayesian nonparamteric modeling in quantile regression. *Scandinavian Journal of Statistics*, **36**, 297–319.
Kozumi H and Kobayashi G (2011) Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, **81**, 1565–78.
Krivobokova T, Kneib T and Claeskens G (2010) Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association*, **105**, 852–63.
Lang S, Kneib T and Brezger A (2005) Bayesx: Analyzing Bayesian structural additive regression models. *Journal of Statistical Software*, **14**, 1–22.
Li Q, Xi R and Lin N (2010) Bayesian regularized quantile regression. *Bayesian Analysis*, **5**, 533–56.
Li Y and Zhu J (2008) *l*1-norm quantile regression. *Journal of Computational and Graphical Statistics*, **17**, 163–85.
Lum C and Gelfand A (2012) Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis*, **7**(2), 235–58.
Mayr A (2010) *Prediction inference with ensemble methods*. Master Thesis, Institut for Statistics, LMU, Munich.
Oh HS, Lee TCM and Nychka DW (2011) Fast nonparametric quantile regression with arbitrary smoothing methods. *Journal of Computational and Graphical Statistics*, **20**, 510–26.
Ormerod J and Wand M (2010) Explaining variational approximations. *The American Statistician*, **64**(2), 140–53.
Park T and Casella G (2008) The Bayesian lasso. *Journal of the American Statistical Association*, **103**, 681–86.
Reed C and Yu K (2009) *A partially collapsed gibbs sampler for Bayesian quantile regression*. Technical report, Department of Mathematical Sciences, Brunel University.
Rue H and Held L (2005) *Gaussian markov random fields*. Boca Raton, FL: Chapman & Hall / CRC.

Sethuraman J (1994) A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–50.

Taddy M and Kottas A (2010) A Bayesian nonparamteric approach to inference for quantile regression. *Journal of Business and Economic Statistics*, **28**, 357–69.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, **58**, 267–88.

Wang H, Li G and Jiang G (2007) Robust regression shrinkage and consistent variable selection through the LAD-LASSO. *Journal of Business & Economic Statistics*, **25**, 347–55.

Wu Y and Liu Y (2009) Variable selection in quantile regression. *Statistica Sinica*, **10**, 801–17.

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 and Data Analysis*, **55**, 84–96.