Computational Statistics and Data Analysis 55 (2011) 84–96

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian inference for additive mixed quantile regression models

Yu Ryan Yue [a,*], Håvard Rue [b]

[a] Department of Statistics and CIS, Baruch College, City University of New York, New York, NY 10010, United States

[b] Norwegian University for Science and Technology, Norway

ARTICLE INFO

Article history:
Received 9 August 2009
Received in revised form 7 May 2010
Accepted 8 May 2010
Available online 21 May 2010

Keywords:
Additive mixed models
Asymmetric Laplace distribution
Gaussian Markov random fields
Gibbs sampling
Integrated nested Laplace approximations
Quantile regression

ABSTRACT

Quantile regression problems in practice may require flexible semiparametric forms of the predictor for modeling the dependence of responses on covariates. Furthermore, it is often necessary to add random effects accounting for overdispersion caused by unobserved heterogeneity or for correlation in longitudinal data. We present a unified approach for Bayesian quantile inference on continuous response via Markov chain Monte Carlo (MCMC) simulation and approximate inference using integrated nested Laplace approximations (INLA) in additive mixed models. Different types of covariate are all treated within the same general framework by assigning appropriate Gaussian Markov random field (GMRF) priors with different forms and degrees of smoothness. We applied the approach to extensive simulation studies and a Munich rental dataset, showing that the methods are also computationally efficient in problems with many covariates and large datasets.

1. Introduction

Quantile regression, a completely distribution-free approach, has emerged as a useful supplement to ordinary mean regression. Its value has been demonstrated by rapidly expanding literatures in econometrics, social sciences, and biomedical studies; see Koenker (2005) for a nice comprehensive review. Given a fixed quantile $\tau \in (0, 1)$, the simple linear quantile regression model is

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau + \varepsilon_{\tau i}, \quad \varepsilon_{\tau i} \sim F_{\tau i} \text{ subject to } F_{\tau i}(0|\boldsymbol{x}_i) = \tau, \tag{1}$$

where y_i denotes the ith observation and $\boldsymbol{x}_i$ is the corresponding covariate vector (including an intercept). The quantile specific linear effects are given by $\boldsymbol{\beta}_\tau$. The random error $\varepsilon_{\tau i}$ follows an unspecified cumulative distribution function $F_{\tau i}$ and its τth quantile conditional on $\boldsymbol{x}_i$ equals zero. We note that model (1) amounts to assuming that

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau, \tag{2}$$

where $Q_{Y_i}(\tau|\boldsymbol{x}_i) = \inf\{y : F_{Y_i}(y|\boldsymbol{x}_i) \geq \tau\}$ is the τth conditional quantile of y_i given the covariate $\boldsymbol{x}_i$. An alternative criterion to fit the linear quantile regression model can be written as the following minimization problem:

$$\arg\min_{\boldsymbol{\beta}_\tau} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau), \quad \text{where } \rho_\tau(u) = \begin{cases} u\tau & u \geq 0 \\ u(\tau - 1) & u < 0, \end{cases} \tag{3}$$

is the so-called 'check function' of Koenker and Bassett (1978). The minimization problem in (3) can be formulated as a set of linear constraints; therefore the estimation of $\boldsymbol{\beta}_\tau$ can be conducted by linear programing and leads to the τth quantiles of

* Corresponding author.
E-mail address: yu.yue@baruch.cuny.edu (Y.R. Yue).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.006

the response variable; see Koenker (2005). Thus, the check function is the appropriate loss function for quantile regression problems regarded from a decision-theoretical point of view.

More flexible models of quantile regression, however, are needed in cases where nonlinear relationships between covariates and quantiles of the response variable occur. Additionally, it is often necessary to add random effects accounting for overdispersion caused by unobserved heterogeneity or for correlation in longitudinal data. Therefore, we consider an additive mixed quantile regression model with quantile function

$$Q_{Y_i}(\tau|\boldsymbol{x}_i, \boldsymbol{u}_i) = \eta_{\tau i} = \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau + \sum_{j=1}^{q} f_{\tau j}(u_{ij}) + b_{\tau g_i}, \tag{4}$$

where the predictor $\eta_{\tau i}$ is composed of a linear term $\boldsymbol{x}_i^T \boldsymbol{\beta}_\tau$, a group-specific random effect term b_{g_i} with $b_{g_i} = b_g$ if unit i is in group g for $g = 1, \ldots, G$, and a sum of nonlinear terms, where $f_{\tau j}$, for $j = 1, \ldots, q$, denote smooth functions of continuous covariates u_{ij} which are assumed to relate in a nonlinear way to the response's quantile function, and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{iq})^T$. Note that those nonlinear terms comprise not only usual nonlinear effects, but also time trends, seasonal effects, two-dimensional surfaces, interactions, and spatial effects. Clearly, the underlying assumption on the error term remains the same as in (1).

If we only make inference about the conditional mean of the response y_i with the additive mixed predictor in (4), we obtain so-called structured additive regression (STAR) models. Thanks to the various forms that the nonlinear functions can take, the STAR models have a wealth of applications (see, e.g., Fahrmeir and Tutz, 2001; Fahrmeir and Lang, 2001). However, it might be the case that the upper or lower quantiles of y_i may depend on the covariates quite differently from the center. The inference on conditional quantiles can therefore provide a more complete description of functional changes than focusing solely on the mean. The work in this paper can be regarded as an extension of the STAR models to the quantile regression context. We here term the proposed model (4) a structured additive quantile regression (STAQ) model.

To estimate the STAQ model using criterion (3), we develop a fully Bayesian inferential framework. The method is based on assuming that the response has i.i.d. (identically independent distributed) asymmetric Laplace distributions and taking appropriate Gaussian-type priors on the additive components in the predictor. Two different tools are proposed for such Bayesian quantile inference. The first approach is carried out by Markov chain Monte Carlo (MCMC) simulations while the second one is a relatively new technique using integrated nested Laplace approximation (INLA), introduced in Rue et al. (2009). Both approaches are fairly efficient and provide reliable quantile inference, as we shall demonstrate via extensive simulation studies and a Munich rental data analysis.

From the frequentist point of view, there is extensive literature on additive quantile regression, a special case of our STAQ model. To review a few, Koenker et al. (1994) and Koenker and Mizera (2004) used a total variation regularization approach to estimate possible univariate and bivariate nonparametric terms; De Gooijer and Zerom (2003), Yu and Lu (2004), Horowitz and Lee (2005), and Cai and Xu (2008) based their estimation on local polynomial fitting; Takeuchi et al. (2006) and Li et al. (2007) explored reproducing kernel Hilbert space (RKHS) norms for nonparametric quantile estimation; Fenske et al. (2009) used a boosting algorithm to allow for data-driven determination of the amount of smoothness required for the nonlinear effects and combined model selection with an automatic variable selection property.

Compared to its frequentist counterparts, our STAQ model offers advantages in the following aspects. First, it is much more flexible. Besides ordinary linear and nonlinear effects, we can easily account for time trends, for two-dimensional surfaces, interactions and spatial effects, for overdispersion caused by unobserved heterogeneity, or for correlation in longitudinal data. Second, the posterior captures all the uncertainty in an estimate, including the uncertainty that would come from selecting a smoothing parameter. Finally, the computation is fairly efficient for large datasets, and a standard R package is available for easy implementation.

The remainder of this paper is organized as follows. Section 2 provides necessary information about asymmetric Laplace distributions, which will be used as the likelihood in our Bayesian quantile regression models. Section 3 specifies all the priors assigned for the latent parameters in model (4). The MCMC and INLA inference tools are both introduced in Section 4. Simulated and real data examples are provided in Section 5, followed by a general discussion in Section 6. A sample R code to implement our method is provided in the Appendix.

2. Asymmetric Laplace distribution

Bayesian inference requires a likelihood. We need an assumption on the data distribution for Bayesian quantile inference because classical quantile regression has no such restriction. A possible parametric link between the minimization problem (3) and the maximum likelihood theory is the asymmetric Laplace density. This skewed distribution appeared in Koenker et al. (1999) and Yu and Moyeed (2001), among others. We say that a random variable Y has an asymmetric Laplace distribution, denoted by ALD(η, δ_0, τ), if the corresponding probability density function is

$$[y|\eta, \delta_0, \tau] = \tau(1-\tau)\delta_0 \exp\{-\delta_0 \rho_\tau (y-\eta)\}, \quad -\infty < \eta < \infty, \tag{5}$$

where $\eta \in \mathbb{R}$ is a location parameter, $\delta_0 \in \mathbb{R}^+$ is a scale parameter, and $0 < \tau < 1$ is a skewness parameter. Note that the check function ρ_τ assigns weight τ or $1-\tau$ to observations greater or less than η, respectively, and $P(Y \leq \eta) = \tau$ regardless

of the value of δ_0. Therefore, the distribution splits along the scale parameter into two parts, one with probability τ to the left, and one with probability $1-\tau$ to the right. See Yu and Zhang (2005) for further properties of this distribution.

Assuming $y_i \overset{\text{i.i.d.}}{\sim} \text{ALD}(\eta_i, \delta_0, \tau)$, the likelihood for n independent observations is given by

$$\mathcal{L}(\boldsymbol{\eta}, \delta_0; \boldsymbol{y}, \tau) \propto \delta_0^n \exp\left\{-\delta_0 \sum_{i=1}^{n} \rho_\tau(y_i - \eta_i)\right\}. \tag{6}$$

When $\delta_0 = 1$ and $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau$, it is easy to see that maximizing the likelihood in (6) with respect to η_i is equivalent to minimizing the loss function in (3). Following this standard asymmetric Laplace error distribution, Yu and Moyeed (2001), among others, implemented Bayesian inference for (parametric) quantile regression and Yuan and Yin (2010) studied it for longitudinal data. In this paper, we, however, choose to take δ_0 as another parameter in our method for the following reasons. First, due to the form of the check function, the contribution of each observation to the quantile estimation only depends on the sign of the residual $y_i - \eta_i$. As a result, the value of δ_0 is immaterial with regard to estimating correct quantiles by optimizing (6). Second, compared to fixing it to be a constant, using δ_0 as a scale parameter clearly makes asymmetric Laplace distributions more flexible to capture the unknown error distributions. Such parameterization has also been successfully used in both Bayesian and frequentist quantile regression work; see, e.g., Tsionas (2003), Kottas and Krnjajić (2009) and Geraci and Bottai (2007).

Another noteworthy feature about the asymmetric Laplace distribution is that it can be represented as a scale mixture of normals (Tsionas, 2003; Rue and Held, 2005):

$$Y \overset{\mathcal{D}}{=} \eta + \xi W + \sigma Z \sqrt{\delta_0^{-1} W}, \tag{7}$$

where $\xi = \frac{1-2\tau}{\tau(1-\tau)}$, $\sigma^2 = \frac{2}{\tau(1-\tau)}$ are two scalars depending on τ. The random variables $W > 0$ and Z are independent and have exponential distribution with mean δ_0^{-1} and standard normal distribution, respectively. This mixture representation provides an easy way to construct efficient Gibbs sampling for MCMC inference, as shown in Section 4.1.

There exist other approaches of specifying error distributions for Bayesian quantile regression. To avoid the restrictive parametric assumption, Kottas and Krnjajić (2009) and Taddy and Kottas (in press) constructed general classes of semiparametric and nonparametric distributions for the likelihood using Dirichlet process mixture models. Reich et al. (2010) assumed a flexible infinite mixture of normals for errors and used a stick-breaking construction for specifying priors. In addition, Dunson and Taylor (2005) used an approximate method based on the substitution likelihood for quantiles. The reasons we chose the asymmetric Laplace distribution in this work are two-fold. First, it facilitates Bayesian inference and computation in the complicated additive models we consider here. Second, it is flexible enough to accommodate different kinds of error distribution (e.g., symmetric, heavy tailed, or skewed), as we shall show in Section 5.1.

3. Prior distributions

3.1. Smoothness priors for functions

To facilitate the description of our method, we will suppress the subscript τ (quantile parameter) in the following. The priors for the unknown functions $f_j(\cdot), j = 1, \ldots, q$, belong to the class of Gaussian Markov random fields (GMRFs). Let $\boldsymbol{f} = (f(u_1), \ldots, f(u_n))^T$ be the vector of the responses at $u_i, i = 1, \ldots, n$. We say that $\boldsymbol{f}$ is a GMRF with mean $\boldsymbol{\mu}$ and precision (i.e., inverse covariance) matrix $\delta\boldsymbol{Q}$ if and only if it has density of the form

$$[\boldsymbol{f} \mid \delta] \propto \delta^{(n-m)/2} \exp\left(-\frac{\delta}{2}(\boldsymbol{f} - \boldsymbol{\mu})'\boldsymbol{Q}(\boldsymbol{f} - \boldsymbol{\mu})\right), \tag{8}$$

where $\boldsymbol{Q}$ is a semidefinite matrix of constants with rank $n - m$ ($m \geq 0$). When $m > 0$, the prior (8) is improper Gaussian with Markov properties: for some $i \neq j$, $f(u_i)$ and $f(u_j)$ are independent conditional on $\{f(u_k) : k \neq i, k \neq j\}$. These Markov properties are conveniently encoded in the structure matrix $\boldsymbol{Q}$: $\boldsymbol{Q}_{ij} = 0$ if and only if $f(u_i)$ and $f(u_j)$ are independent conditional on $\{f(u_k) : k \neq i, k \neq j\}$. In most cases, only $\mathcal{O}(n)$ of the n^2 entries of $\boldsymbol{Q}$ are non-zero, so $\boldsymbol{Q}$ is sparse. This allows efficient algorithms for sampling $\boldsymbol{f}$ (see Section 4). The GMRFs differ from each other with respect to $\boldsymbol{Q}$, whose specific form depends on the covariate type and prior belief about the smoothness of $\boldsymbol{f}$. The GMRFs have been successfully used in many applications; see Rue and Held (2005) for a comprehensive review.

In this paper we use two particular GMRFs: continuous random walk (CRW) models (see, e.g. Wecker and Ansley, 1983) for nonlinear effects, and intrinsic autoregressive models (see, e.g. Besag and Kooperberg, 1995) for spatial effects. Compared to other commonly used GMRFs (i.e., random walk models), the CRW models are consistent with respect to the choice of the locations and the resolution of the covariates. The intrinsic autoregressive models have flexible neighbor structures, which are especially useful to analyze the areal data used in this paper. In addition, both GMRFs we use here have attractive Markov properties that make the $\boldsymbol{Q}$ matrix sparse so we can do efficient computations. More details on comparisons among different GMRFs can be found in Rue and Held (2005)

Another large class of smoothness priors for functions is Bayesian P-splines (Lang and Brezger, 2004; Brezger and Lang, 2006). As shown by Brezger and Lang (2006), some GMRFs can even be treated as special examples of Bayesian P-splines. We use GMRF priors because they are more intuitive to model timescale and metric covariates, such as *year* and *floor size* variables in the Munich rental dataset in Section 5.2. Also, we wish to analyze spatially correlated effects of the districts in the rental data, where the GMRF in use coincides with one of the Bayesian P-spline priors in Brezger and Lang (2006). More importantly, the inference methods (both MCMC and INLA) developed in Section 4 only work with GMRF priors.

3.1.1. Continuous random walk models

Let $u_1 < u_2 < \cdots < u_n$ be the set of continuous locations and $z_i = f(u_i)$ be the function evaluations at u_i, for $i = 1, \ldots, n$. The construction of a CRW model is based on a discretely observed continuous time process $z(u)$ that is a realization of an $(m-1)$-fold integrated Wiener process given by

$$z(u) = \int_0^u \frac{(u-h)^{m-1}}{(m-1)!} \mathrm{d}W(h), \tag{9}$$

where $W(h)$ is a standard Wiener process. One can show that the density of $z(u)$ is Gaussian with zero mean and a completely dense matrix $\boldsymbol{Q}$. Since factorizing an $n \times n$ dense matrix is at cost $\mathcal{O}(n^3)$, such a matrix $\boldsymbol{Q}$ would make Bayesian computation cumbersome for large datasets. Fortunately, the solution of (9) does have a Markov property on an augmented space

$$\tilde{\boldsymbol{z}} = \left[z_1, z_1^{(1)}, \ldots, z_1^{(m-1)}, \ldots, z_n, z_n^{(1)}, \ldots, z_n^{(m-1)}\right]^T,$$

where the derivatives (up to mth order) at u_i are also included. The mn-dimensional vector $\tilde{\boldsymbol{z}}$ is also (singular) Gaussian with mean zero but a banded precision $\boldsymbol{Q}$ of bandwidth $2m-1$ (see Rue and Held, 2005, Chapter 3.5 for details). For inference using the MCMC technique, we will first use efficient algorithms for sparse matrices (Rue and Held, 2005) to simulate the full vector $\tilde{\boldsymbol{z}}$ at cost $\mathcal{O}(n)$, and then simply ignore the derivatives in the analysis. The CRW models have been successfully used as priors for smoothing splines in Wahba (1990) and for varying-coefficient models in Eubank et al. (2004), who instead used recursions based on the Kalman filter.

We now present the important case $m = 2$. Define $\kappa_i = u_{i+1} - u_i$, and define the matrices

$$\boldsymbol{A}_i = \begin{pmatrix} 12/\kappa_i^3 & 6/\kappa_i^2 \\ 6/\kappa_i^2 & 4/\kappa_i \end{pmatrix}, \qquad \boldsymbol{B}_i = \begin{pmatrix} -12/\kappa_i^3 & 6/\kappa_i^2 \\ -6/\kappa_i^2 & 2/\kappa_i \end{pmatrix}, \qquad \boldsymbol{C}_i = \begin{pmatrix} 12/\kappa_i^3 & -6/\kappa_i^2 \\ -6/\kappa_i^2 & 4/\kappa_i \end{pmatrix},$$

for $i = 1, \ldots, n-1$. Then the vector $\tilde{\boldsymbol{z}} = \left[z_1, z_1^{(1)}, \ldots, z_n, z_n^{(1)}\right]^T$ has

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 & & & & \\ \boldsymbol{B}_1' & \boldsymbol{A}_2 + \boldsymbol{C}_1 & \boldsymbol{B}_2 & & & \\ & \boldsymbol{B}_2' & \boldsymbol{A}_3 + \boldsymbol{C}_2 & \boldsymbol{B}_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \boldsymbol{B}_{n-1} & \\ & & & \boldsymbol{B}_{n-1}' & \boldsymbol{C}_{n-1} & \end{pmatrix}. \tag{10}$$

Here we have used diffuse priors for z_1 and $z_1^{(1)}$. The null space of $\boldsymbol{Q}$ is spanned by two vectors, $(u_1, 1, u_2, 1, \ldots, u_n, 1)'$ and $(1, 0, 1, 0, \ldots, 1, 0)'$. When we go to higher-order models, the matrix $\boldsymbol{Q}$ has the same structure as (10), but the matrices $\boldsymbol{A}_i$, $\boldsymbol{B}_i$, and $\boldsymbol{C}_i$ will differ (see Rue and Held, 2005, for more information).

3.1.2. Intrinsic autoregressive spatial models

Let us now turn our attention to a spatial covariate u, where the values of u represent the location or site in connected geographical regions. A common way to deal with spatial covariates is based on a set of predefined neighbors for each u_i. For geographical data as considered in this paper, we usually assume that two sites u_i and u_j are neighbors if they share a common boundary. Letting n_i denote the number of neighbors of site u_i, we assume the following spatial smoothness prior for the function evaluations $f(u_i)$, $i = 1, \ldots, n$:

$$f(u_i) \mid \{f(u_j) j \neq i\}, \qquad \delta \sim \mathrm{N}\left(\frac{1}{n_i}\sum_{j:j\sim i} f(u_j), \frac{1}{n_i \delta}\right), \tag{11}$$

where $j \sim i$ denotes that site u_i and u_j are neighbors. Thus the conditional mean of $f(u_i)$ is an unweighted average of evaluations of neighboring sites. The density of vector $\boldsymbol{f}$ in (11) has the same expression as in (8) with mean zero and matrix $\boldsymbol{Q}$ such that

$$Q_{ij} = \begin{cases} n_i & i = j, \\ -1 & i \sim j, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

It is easy to see that the sparse matrix $\boldsymbol{Q}$ has rank $n-1$.

A more general prior including expression (11) as a special case is given by

$$f(u_i) \mid \{f(u_j) j \neq i\}, \qquad \delta \sim \mathrm{N}\left(\frac{\sum_{j:j\sim i} w_{ij} f(u_j)}{\sum_{j:j\sim i} w_{ij}}, \frac{1}{\delta \sum_{j:j\sim i} w_{ij}}\right),$$

where the w_{ij} are known weights, e.g., the inverse Euclidian distance between the sites of u_i and u_j. The corresponding matrix $\boldsymbol{Q}$ has the same structure as in (12) but different entries (see Rue and Held, 2005, page 103). Other options for modeling spatial effects like conditional autoregressive (CAR) models and stationary Gaussian random field (Kriging) models are also available with the same general form (see Rue and Held, 2005).

Let $\boldsymbol{f}_j, j = 1, \ldots, q$, be a vector of kth function components in (4). Since they are of the same form, we assume that the priors on functions $\boldsymbol{f}_j$ are

$$[\boldsymbol{f}_j \mid \delta_j] \propto \delta_j^{(n-m_j)/2} \exp\left(-\frac{\delta_j}{2} \boldsymbol{f}_j' \boldsymbol{Q}_j \boldsymbol{f}_j\right),$$

where $\boldsymbol{Q}_j$ has rank $n - m_j$ and its structure depends on the type of covariates.

3.2. Further prior assumptions

For a fully Bayesian analysis, hyperpriors for all precisions $(\delta_0, \delta_1, \ldots, \delta_q)$ are introduced in a further stage of the hierarchy. Common choices are gamma priors with parameters a and c so that the mean is a/c and the variances is a/c^2. We make gamma priors proper and highly dispersed. A sensitivity analysis (not shown here) indicates that the performance of our model is fairly robust to the choice of those priors.

For the fixed effect parameters $\{\beta_j\}$, we shall assume a weakly informative Gaussian prior $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \phi^{-1}\boldsymbol{I})$ with small precision ϕ. If $\boldsymbol{\beta}$ is a high-dimensional vector, one may consider using the Bayesian regularization priors developed in Kneib et al. (in press), where conditionally Gaussian priors are assigned with suitable hyperprior assumptions on the variances inducing the desired shrinkage and sparseness on coefficient estimates. These priors can be easily utilized in the models proposed here with small effort on modifying the algorithms.

For the group-specific random effect vector $\boldsymbol{b} = (b_1, \ldots, b_G)'$, we make the usual assumption that the b_g are i.i.d. Gaussian with precision ν, i.e.,

$$b_g \mid \nu \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \nu^{-1}), \quad g = 1, \ldots, G,$$

and define again a highly dispersed gamma prior for ν.

4. Posterior inference

4.1. Markov chain Monte Carlo approaches

Recall that we assume the likelihood to be an asymmetric Laplace distribution, which has a representation as given by (7). Letting $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_q)'$, the posterior distribution can be obtained through

$$[\boldsymbol{\eta} \mid \boldsymbol{y}] \propto \int [y \mid \boldsymbol{\eta}, \delta_0, \boldsymbol{w}][\boldsymbol{w} \mid \delta_0][\boldsymbol{\eta} \mid \boldsymbol{\delta}, \nu][\boldsymbol{\delta}][\nu][\delta_0] \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}\boldsymbol{\delta} \, \mathrm{d}\nu \, \mathrm{d}\delta_0, \tag{13}$$

where $y \mid \boldsymbol{\eta}, \boldsymbol{w}, \delta_0 \sim N(\boldsymbol{\eta} + \xi\boldsymbol{w}, \sigma^2 \boldsymbol{w} \delta_0^{-1})$ with the priors and hyperpriors defined in Section 3. The expression (13) yields a tractable and efficient Gibbs sampler that works as follows:

1. Simulate $w_i^{-1} \mid \cdot \sim$ Inverse Gaussian $(\mu', \lambda'), i = 1, \ldots, n$, where

$$\mu' = \sqrt{\frac{\xi^2 + 2\sigma^2}{(y_i - \eta_i)^2}} \quad \text{and} \quad \lambda' = \frac{\delta_0(\xi^2 + 2\sigma^2)}{\sigma^2},$$

in the parameterization of the inverse Gaussian density given by

$$f(x) = \sqrt{\frac{\lambda'}{2\pi}} x^{-3/2} \exp\left\{-\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x}\right\}, \quad x > 0;$$

see, e.g., Chhikara and Folks (1989).

2. Simulate the fixed effect parameters $\boldsymbol{\beta} \mid \cdot \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2\delta_0^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}})$, where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{D}_w^{-1}(\boldsymbol{y} - \boldsymbol{\eta} - \xi\boldsymbol{w} + \boldsymbol{X}\boldsymbol{\beta}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{D}_w^{-1}\boldsymbol{X}\right)^{-1},$$

and $\boldsymbol{D}_w = \mathrm{diag}(w_1, \ldots, w_n)$.

3. Simulate the function vectors $\boldsymbol{f}_k \mid \cdot \sim N(\boldsymbol{\mu}_k, \sigma^2\delta_0^{-1}\boldsymbol{\Sigma}_k)$, $k = 1, \ldots, n_f$, where

$$\boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k\boldsymbol{D}_{\boldsymbol{w}}^{-1}(\boldsymbol{y} - \boldsymbol{\eta} - \xi\boldsymbol{w} + \boldsymbol{f}_k), \quad \boldsymbol{\Sigma}_k = (\boldsymbol{D}_{\boldsymbol{w}}^{-1} + \sigma^2\lambda_k\boldsymbol{Q}_k)^{-1},$$

and $\lambda_k = \delta_k/\delta_0$. For identifiability, we add a sum-to-zero constraint to each $\boldsymbol{f}_k$ by computing $\boldsymbol{f}_k^* = \boldsymbol{f}_k - \boldsymbol{\Sigma}_k\boldsymbol{1}(\boldsymbol{1}'\boldsymbol{\Sigma}_k\boldsymbol{1})^{-1}\boldsymbol{1}'\boldsymbol{f}_k$ (see Rue and Held, 2005, Section 2.3.3).

4. Simulate the random effect $\boldsymbol{b} \mid \cdot \sim N(\boldsymbol{\mu}_{\boldsymbol{b}}, \sigma^2\delta_0^{-1}\boldsymbol{\Sigma}_{\boldsymbol{b}})$, where

$$\boldsymbol{\mu}_{\boldsymbol{b}} = \boldsymbol{\Sigma}_{\boldsymbol{b}}\boldsymbol{X}_G^T\boldsymbol{D}_w^{-1}(\boldsymbol{y} - \boldsymbol{\eta} - \xi\boldsymbol{w} + \boldsymbol{X}_G\boldsymbol{b}), \quad \boldsymbol{\Sigma}_{\boldsymbol{b}} = \left(\boldsymbol{X}_G^T\boldsymbol{D}_w^{-1}\boldsymbol{X}_G + \sigma^2\delta_0^{-1}\nu\boldsymbol{I}_n\right)^{-1},$$

and $\boldsymbol{X}_G$ is $n \times G$ incidence matrix for $\boldsymbol{b}$.

5. Simulate the error precision:

$$\delta_0 \mid \cdot \sim \mathrm{Gamma}\left(a_0 + \frac{3n}{2}, c_0 + \frac{1}{2\sigma^2}\sum_{i=1}^{n} w_i^{-1}(y_i - \eta_i - \xi w_i)^2 + \sum_{i=1}^{n} w_i\right).$$

6. Simulate the smoothing precision δ_j, $j = 1, \ldots, q$:

$$\delta_j \mid \cdot \sim \mathrm{Gamma}\left(a_j + \frac{n - m_j}{2}, c_j + \frac{1}{2}\boldsymbol{f}_j'\boldsymbol{Q}_j\boldsymbol{f}_j\right).$$

7. Simulate the random effect precision:

$$\nu \mid \cdot \sim \mathrm{Gamma}\left(a_\nu + \frac{G}{2}, b_\nu + \frac{1}{2}\boldsymbol{b}'\boldsymbol{b}\right).$$

Drawing samples iteratively from these regular full conditional distributions is straightforward, without any further need for tuning. Still, some of the updating steps require additional comments. For the high-dimensional vector $\boldsymbol{\beta}$, it will typically be necessary to apply a block update since a complete update is extremely expensive due to the high-dimensional matrix computation. For the smooth and spatial effects, one can take advantage of the sparse structures possessed by the precision matrices of the full conditionals. Using the Cholesky decomposition for sparse matrices given in Rue (2001) and Rue and Held (2005), the simulation of the high-dimensional Gaussian effects becomes feasible, as demonstrated by the examples in this paper. The notorious correlation between a GMRF and its precision parameter tends to make MCMC methods exhibit poor performance. One may use the one-block approach (see Rue and Held, 2005, Chapter 4) to improve Markov chain mixing, or employ a conceptually different inference tool in the next section to completely avoid using the MCMC technique.

4.2. Inference by integrated nested Laplace approximations

Integrated nested Laplace approximations (INLA) is a new approach to statistical inference for latent Gaussian models introduced by Rue and Martino (2007) and Rue et al. (2009). It provides a fast, deterministic alternative to the MCMC approach, which is the standard tool for inference in such models. The main advantage of the INLA approach over the MCMC approach is in computation; it gives answers in seconds and minutes whereas the MCMC approach requires hours and even days. In addition, INLA inference does not suffer from the slow convergence and poor mixing issues that MCMC methods often have for the models concerned here.

In short, the INLA approach provides a recipe for fast Bayesian inference using accurate approximations to the marginal posterior density for the hyperparameters and the posterior marginal densities for the latent variables. The approximate posterior marginals can then be used to compute summary statistics of interest, such as posterior means, variances or quantiles. The theory behind INLA is thoroughly described in Rue et al. (2009) and will not be repeated here.

Unfortunately, we cannot directly implement the INLA method to our models since the asymmetric Laplace distribution is assumed to be the likelihood and the second-order derivative of the log likelihood is zero. However, such log likelihood has an "overall curvature". We therefore choose to first approximate the check function using

$$\tilde{\rho}_{\tau,\gamma}(u) = \begin{cases} \gamma^{-1}\log(\cosh(\tau\gamma|u|)) & \text{if } u \geq 0 \\ \gamma^{-1}\log(\cosh((1-\tau)\gamma|u|)) & \text{if } u < 0, \end{cases}$$

which has second-order derivatives everywhere, and then use INLA to make quantile inference based on the approximated Laplace distribution. Green (1990) used a similar type of approximation in Bayesian image analysis. The parameter $\gamma > 0$ is fixed and precision of the approximation increases as $\gamma \to \infty$. When estimating mild (not extreme) quantiles, one may simply select a value of γ to obtain a suitable approximated Laplace distribution as the likelihood. In our experience, γ between 2 and 5 is sufficient for INLA to provide satisfactory quantile inference in most applications. When estimating

more extreme quantiles with little data support, it is possible that the INLA algorithm converges to wrong estimates. In that case we need to tune the parameters in the likelihood so it can 'squeeze' itself into the correct likelihood, starting from an approximation with small γ which is more rough but more numerically stable, and then, at each run, restart from the previous found estimates of the hyperparameters with larger γ until it shows convergence. By doing this squeezing procedure, INLA estimation can much be more stable.

Consequently, using INLA to make quantile inference has two sources of error: one is from approximating the check function and the other is from approximating posterior marginals. However, these approximation errors appear to be negligible in practice; see Section 5.1.

5. Examples

5.1. Simulation studies

There are three purposes in this section. First, to verify the adequacy of using the asymmetric Laplace distribution as the likelihood by considering several differently distributed errors. Second, to compare the MCMC and INLA approaches, two different inference tools for the STAQ model, regarding both accuracy and computational efficiency. Third, to compare the proposed method with an alternative approach, boosting quantile regression models in Fenske et al. (2009), which have been shown to have at least equally fine performance compared to the state-of-art quantile smoothing splines via total variation regularization in Koenker et al. (1994) and Koenker and Mizera (2004).

The datasets are simulated by the following two additive models:

$$\begin{aligned}&\text{Model 1: } y_i = 0.4u_i + 0.5\sin(2.7u_i) + 1.1/(1+u_i^2) + \varepsilon_i, \quad u_i \in [-3, 3],\\ &\text{Model 2: } y_i = 2 + \sin(2u_i/3) + 0.5(1+(u_i-3)^2)\varepsilon_i, \quad u_i \in [0, 6],\end{aligned} \tag{14}$$

where the u_i are equally spaced and the error terms ε_i are i.i.d. sampled from distribution F. The resulting quantile functions have nonlinear predictor structures given by

$$\begin{aligned}&\text{Model 1: } Q_1(\tau|u_i) = 0.4u_i + 0.5\sin(2.7u_i) + 1.1/(1+u_i^2) + F^{-1}(\tau),\\ &\text{Model 2: } Q_2(\tau|u_i) = 2 + \sin(2u_i/3) + 0.5(1+(u_i-3)^2)F^{-1}(\tau).\end{aligned}$$

The number of observations is fixed at $n = 400$ for all setups. We consider three different error distributions: a standard normal distribution, a Student t distribution with two degrees of freedom, and a gamma distribution with shape 4 and scale 1. Fig. 1 displays data examples from both models with all kinds of error term. Note that Model 2 has a heteroscedastic data structure where the quantile curves are no longer parallel shifted as in Model 1; the heavy-tailed t distribution leads to some extreme outliers in both models. Similar examples have been analyzed in Taddy and Kottas (in press), Kottas and Krnjajić (2009) and Fenske et al. (2009).

For each of the generated datasets, we estimated the nonparametric functions for a fixed quantile grid on $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$ by the STAQ model using both MCMC and INLA approaches, and by the boosting algorithm (function gamboost() from the package **mboost**). Following Fenske et al. (2009), we used cubic penalized spline base-learners with second-order difference penalty, 20 inner knots and three degrees of freedom for the boosting models. In addition, an extra test dataset was created to determine the number of boosting iterations for each of the simulated datasets. For both the MCMC and INLA approaches, the CRW2 model is used as the function prior and diffuse gamma priors ($a = 1, c = 0.001$) are assumed for all precision parameters. The value of the tuning parameter γ for INLA was chosen to be 3 for all setups. The results from the MCMC methods are based on 40,000 total iterations, with a 15,000 burn-in period, sampled every fifth draw. The Markov chains converge fairly quickly and mix quite well.

To evaluate the estimation performance, we compute the mean absolute deviation errors (MADEs) as given by

$$\text{MADE} = \frac{1}{n}\sum_{i=1}^{n} |Q_k(\tau|u) - \hat{Q}_k(\tau|u)|, \quad k = 1, 2,$$

for 200 replications, then obtain the quartiles of those 200 MADEs. Tables 1 and 2 show the results of all competing methods for Models 1 and 2, respectively. Consider first the results of the INLA and MCMC approaches. From the tables, the two algorithms yield close performances when estimating three quartiles: the MADE results are mostly located in the same range. We also present the estimated quartile curves for Model 2 of normally and Student t distributed errors in Fig. 2, from which we can hardly tell any differences. Furthermore, the Laplace approximation in INLA provides marginal posterior distributions that are close to the samples of a long MCMC run as shown in Fig. 3 (bottom). When estimating more extreme quantiles, the MCMC approach, however, seems to perform a little worse than the INLA method. Compared to the boosting algorithm, both MCMC and INLA approaches performed equally well in Model 1 and slightly outperformed in Model 2 where the heterogenous error terms were considered. As one referee suggested, we also did some simulations with the sample size smaller than 400, i.e., $n = 100$ and 250. The corresponding MADE results showed quite similar patterns as we observed for $n = 400$, although the values are higher due to the decreased signal-to-noise ratios. Therefore, we conclude that (i) assuming an asymmetric Laplace distribution of the likelihood is appropriate for Bayesian nonparametric quantile inference; (ii) the

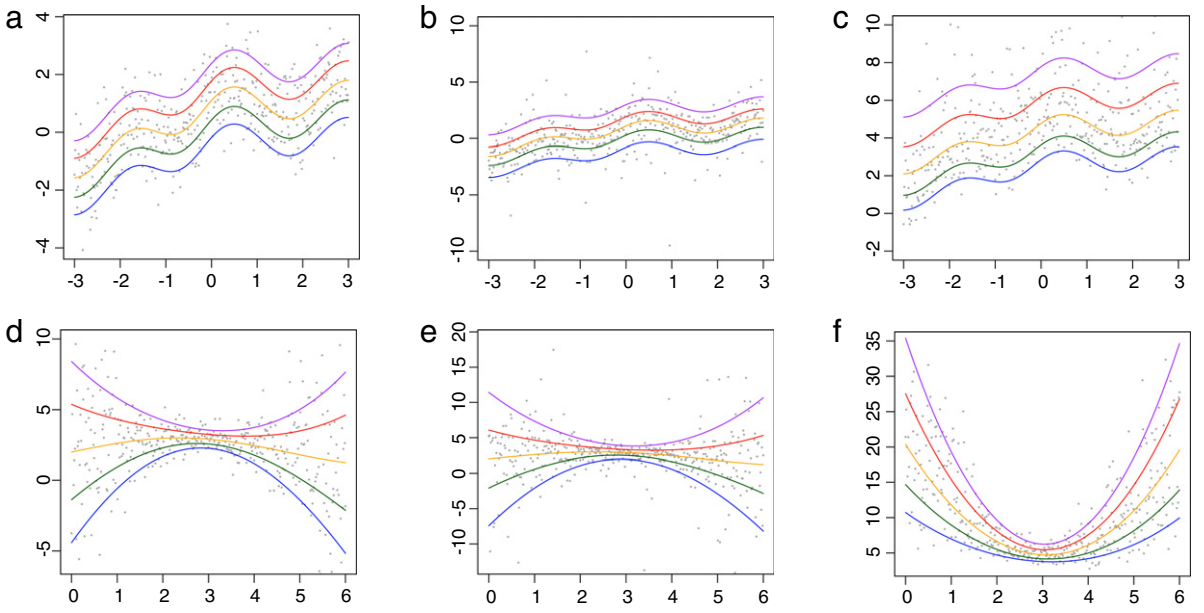

Fig. 1. Simulated examples for Model 1 (top) or Model 2 (bottom) with standard normal (left) or Student t (middle) or gamma (right) distributed error terms. Lines represent true underlying quantile curves for $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$.

Table 1
Median with lower and upper quartiles of estimated MADE from 200 replications of Model 1 simulation setups with three types of error.

τ	MCMC	INLA	Boost
Standard normal errors			
0.10	0.207 (0.169, 0.232)	0.191 (0.155, 0.222)	0.187 (0.169, 0.211)
0.25	0.164 (0.139, 0.192)	0.156 (0.130, 0.187)	0.160 (0.133, 0.184)
0.50	0.145 (0.125, 0.164)	0.139 (0.113, 0.162)	0.143 (0.120, 0.170)
0.75	0.165 (0.140, 0.190)	0.153 (0.128, 0.182)	0.159 (0.131, 0.181)
0.90	0.197 (0.153, 0.236)	0.196 (0.170, 0.233)	0.195 (0.159, 0.221)
Student t distributed errors			
0.10	0.407 (0.347, 0.465)	0.397 (0.359, 0.441)	0.375 (0.321, 0.449)
0.25	0.215 (0.178, 0.256)	0.217 (0.174, 0.250)	0.216 (0.188, 0.252)
0.50	0.166 (0.137, 0.195)	0.162 (0.136, 0.185)	0.166 (0.139, 0.204)
0.75	0.214 (0.178, 0.253)	0.218 (0.173, 0.252)	0.220 (0.185, 0.255)
0.90	0.391 (0.338, 0.452)	0.387 (0.350, 0.443)	0.386 (0.329, 0.461)
Gamma distributed errors			
0.10	0.210 (0.172, 0.260)	0.208 (0.171, 0.245)	0.216 (0.181, 0.251)
0.25	0.229 (0.195, 0.270)	0.225 (0.198, 0.284)	0.233 (0.193, 0.275)
0.50	0.280 (0.225, 0.321)	0.271 (0.222, 0.311)	0.268 (0.226, 0.332)
0.75	0.357 (0.296, 0.426)	0.355 (0.307, 0.420)	0.345 (0.293, 0.417)
0.90	0.457 (0.420, 0.496)	0.444 (0.415, 0.502)	0.461 (0.391, 0.534)

performance of the STAQ model is competitive; (iii) INLA provides quite good approximations to the posterior distributions for valid inference.

To demonstrate the computational efficiency of our method, we experimented with both the MCMC (also 40,000 iterations) and INLA approaches on the simulated datasets of five different sample sizes, $n = 400, 800, 1500, 3000$ and 6000. The MCMC computations were performed on a quad-core Intel Pentium 2.84 GHz CPUs using a Fortran program (code available upon request), while the INLA computation was performed on a single-processor 2.13-GHz laptop using an R interface to the 'inla' program (Martino and Rue, 2008). The computational times are recorded in Table 3. Note that we only post the results for estimating medians of the datasets simulated from Model 1 with normal error terms. Similar performances were found in other situations. As we can see, the MCMC computation increases linearly with sample size n. The INLA approach is obviously much faster: it took less than 20 s for $n = 6000$. (The INLA results were scaled for CPU difference.)

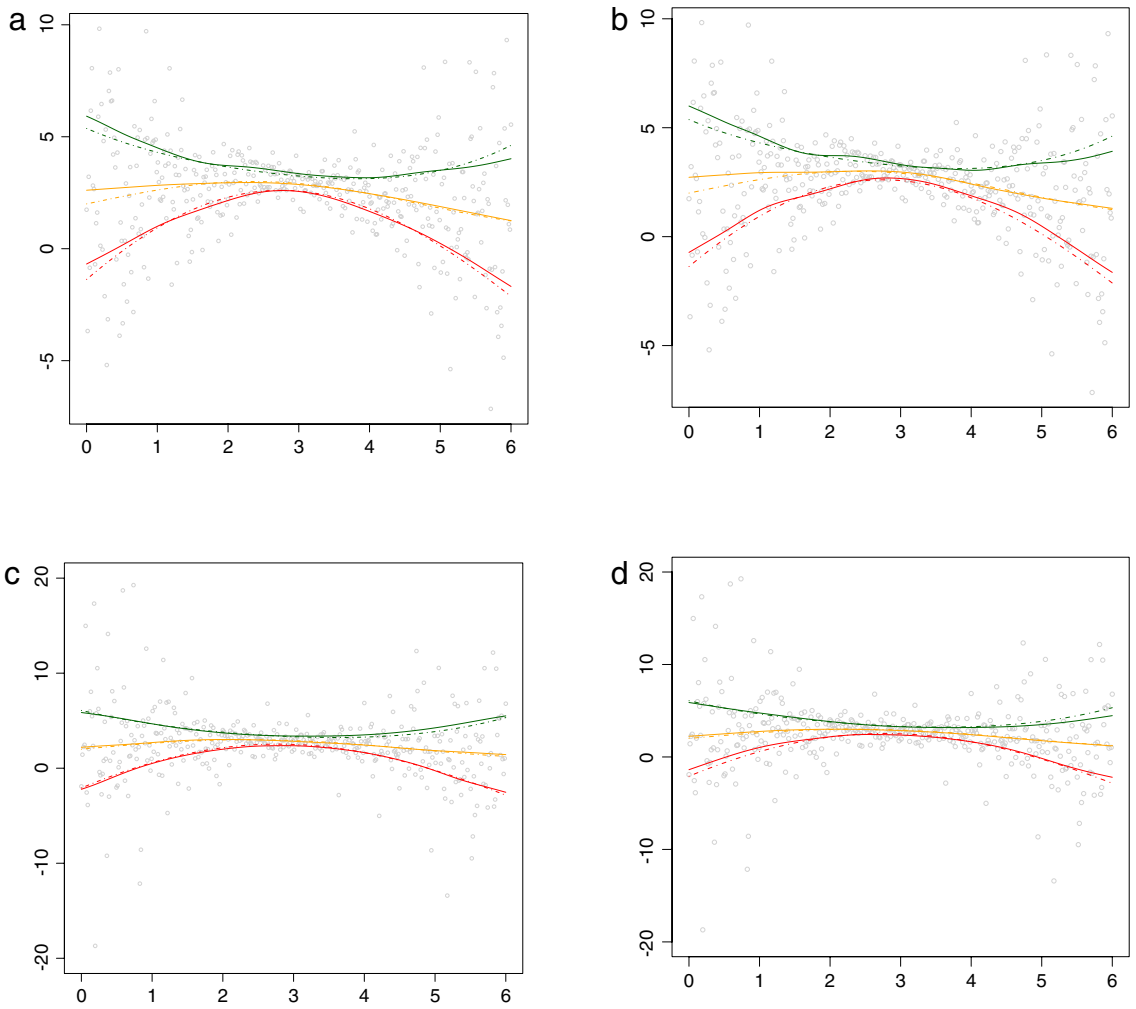

Fig. 2. Example for estimated quartile curves for Model 2 with normal (top) and Student t (bottom) distributed error terms using the INLA (left) and MCMC (right) approaches.

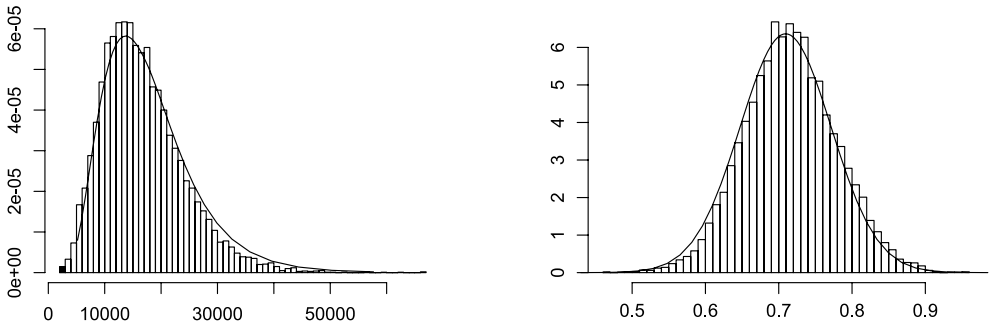

Fig. 3. Comparison between marginal posterior distributions of precision δ_0 (left) and one function evaluation $f(u_i)$ (right) from INLA (line) and MCMC (histogram) samples.

5.2. Munich rental guide

In the following, we analyze the 2003 Munich rental dataset (see Rue and Held, 2005, Section 4.2.1). The response variable Y_i is the rent (per square meter in Euros) for a flat, and the covariates are spatial location ($\boldsymbol{u}_i$), floor space (size_i), year of construction (year_i), and various indicator variables such as an indicator for a flat with no central heating, no warm water, more refined kitchen equipment, and so on (see Table 4 for details). The dataset consists of $n = 2035$ observations. There

Table 2
Median with lower and upper quartiles of estimated MADE from 200 replications of Model 2 simulation setups with three types of error.

τ	MCMC	INLA	Boost
Standard normal errors			
0.10	0.385 (0.304, 0.468)	0.369 (0.292, 0.459)	0.380 (0.303, 0.464)
0.25	0.261 (0.196, 0.335)	0.249 (0.181, 0.312)	0.317 (0.250, 0.368)
0.50	0.206 (0.144, 0.287)	0.212 (0.138, 0.298)	0.294 (0.230, 0.351)
0.75	0.233 (0.159, 0.304)	0.235 (0.183, 0.308)	0.316 (0.265, 0.376)
0.90	0.365 (0.278, 0.489)	0.345 (0.263, 0.473)	0.379 (0.305, 0.469)
Student t distributed errors			
0.10	0.707 (0.536, 0.878)	0.689 (0.510, 0.854)	0.721 (0.541, 0.904)
0.25	0.322 (0.249, 0.401)	0.321 (0.248, 0.415)	0.413 (0.339, 0.521)
0.50	0.217 (0.160, 0.300)	0.221 (0.159, 0.311)	0.328 (0.278, 0.403)
0.75	0.322 (0.209, 0.405)	0.311 (0.236, 0.416)	0.412 (0.340, 0.502)
0.90	0.732 (0.535, 0.937)	0.721 (0.530, 0.924)	0.742 (0.644, 0.985)
Gamma distributed errors			
0.10	0.375 (0.318, 0.433)	0.353 (0.302, 0.411)	0.381 (0.321, 0.488)
0.25	0.387 (0.298, 0.486)	0.388 (0.283, 0.547)	0.451 (0.365, 0.521)
0.50	0.490 (0.398, 0.598)	0.422 (0.323, 0.514)	0.528 (0.412, 0.622)
0.75	0.706 (0.559, 0.839)	0.707 (0.520, 0.842)	0.682 (0.517, 0.826)
0.90	0.946 (0.847, 1.17)	0.933 (0.853, 1.12)	0.969 (0.793, 1.16)

Table 3
Computational time (s) versus sample size.

Sample size	400	800	1500	3000	6000
MCMC	22.64	46.82	87.89	172.29	338.99
INLA	1.01	1.97	3.84	8.25	18.09

Table 4
Variables in the Munich rental dataset.

Variable	Explanation
Rent	Net rent per square meter
Floor.size	Size of the flat in square meters
Year	Year of construction of the building in which the flat is located
Location	Location index (in terms of subquarters)
Gute.Wohnlage	Dummy variable for good locations/good neighborhoods
Beste.Wohnlage	Dummy variable for very good locations/very good neighborhoods
Keine.Wwv	Dummy variable for absence of warm water supply
Keine.Zh	Dummy variable for absence of central heating system
Kein.Badkach	Dummy variable for absence of flagging in the bathroom
Besond.Bad	Dummy variable for special features of the bathroom
Gehobene.Kueche	Dummy variable for more refined kitchen equipment
zim1	Dummy variable for flats with 1 room
zim2	Dummy variable for flats with 2 rooms
zim3	Dummy variable for flats with 3 rooms
zim4	Dummy variable for flats with 4 rooms
zim5	Dummy variable for flats with 5 rooms
zim6	Dummy variable for flats with 6 rooms

are 380 districts in Munich. The floor size varies from 17 to 185 square meters and the year of construction goes from 1918 to 2001, both of which are well known to have nonlinear impact on the rent from previous analyses (Fahrmeir et al., 2004; Rue and Held, 2005). The fact that the response variable has a skewed heterogeneous distribution suggests that quantile regression fitting is quite suitable for the analysis of this dataset.

We build a *geoadditive* model for τth quantile regression,

$$\eta_{\tau i} = \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau + f_{\tau 1}(\text{size}_i) + f_{\tau 2}(\text{year}_i) + f_{\tau 3}(\boldsymbol{u}_i), \tag{15}$$

where $\boldsymbol{x}_i$ represents a 13-dimensional vector of categorical covariates with effects summarized in the vector $\boldsymbol{\beta}_\tau$ that is assigned a diffuse normal prior. The functions $f_{\tau 1}$ and $f_{\tau 2}$ are nonlinear effects of floor size and year of construction. Both are modeled as the CRW2 that we described in Section 3.1.1. The function $f_{\tau 3}$ represents the spatial effect modeled as the intrinsic GMRF in Eq. (11) based on the discrete spatial information of the districts of Munich. Estimation of three quartiles ($\tau = 0.25, 0.50, 0.75$) was implemented by INLA (see a code example in Appendix), with tuning parameter $\gamma = 2$ and hyperparameters $a = 1$, $b = 0.001$. The computation time was less than 8 min. As we shall see, quantile inference can offer some useful insights that could not be detected via mean regression.

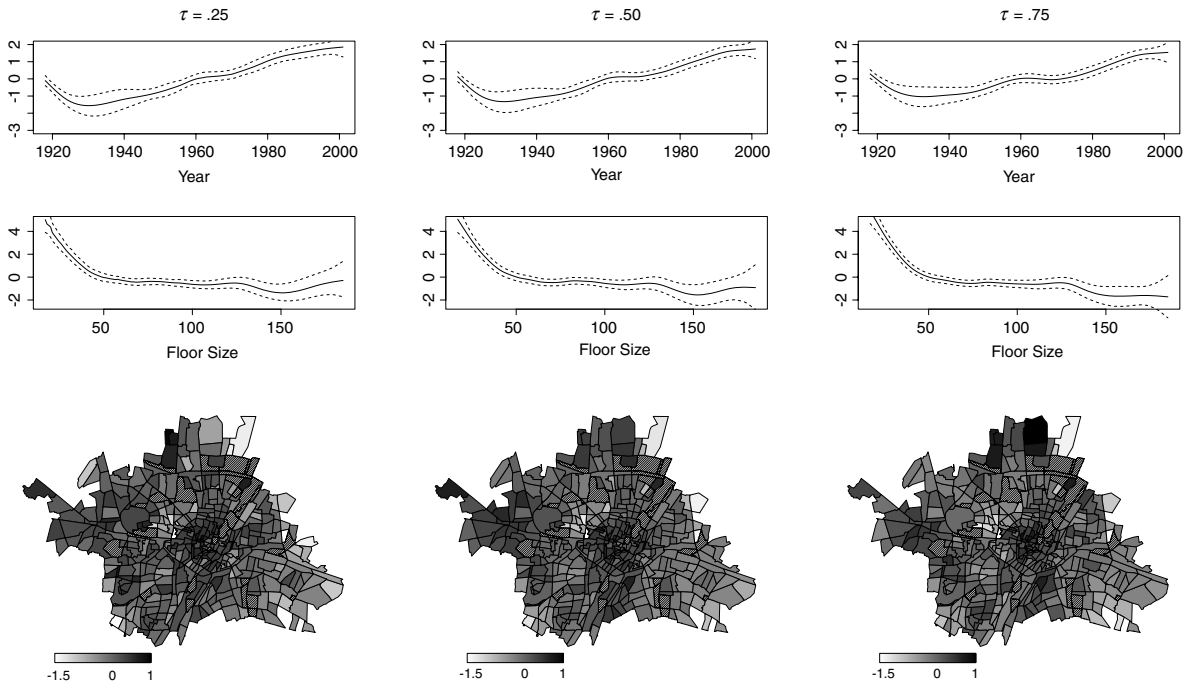

Fig. 4. Munich rental guide. Estimated nonparametric effects with 95% credible bands (top and middle rows) and estimated spatial effects (bottom row) for the quartiles; the shaded areas are districts with no houses, such as parks or fields.

The effects of year of construction and floor size in the model (15) for rents are shown in Fig. 4 (top and middle). Generally speaking, the rents decreased for buildings constructed during the 1920s to 1930s, and then started to increase for newly constructed buildings until the 1960s; after being relatively stable for about 10 years, the rents have kept increasing since 1970s. A closer look shows that the quartiles of rents have slightly different changing patterns over years. Compared to the expensive flats ($\tau = 0.75$), the rents of cheap flats ($\tau = 0.25$) seem to decrease more quickly in the first 10 years but also increase more quickly in the following years. It is the evidence of increasing variance from 1920 to 1930, followed by decreasing variance. Additionally, during the stable period (1960–1970), the rents for the lower quantile increased slightly, were flat for the median, and decreased a little bit for the upper quartile. This can be explained by a boom in construction building in those years, with flats having comparably homogeneous quality. For all quartiles, the effect of floor size is shown as expected with high rents per square meter for smaller flats and only very little variation for flats of size larger than about 50 square meters (except for the size larger than 130 square meters, whose odd behavior can be explained by the lack of data information).

Fig. 4 (bottom) also presents the spatial effect of location in Munich. From expert assessment, we know that rents are high in the center of Munich and in some popular districts along the river Isar and near parks. In contrast, significantly negative effects are found for some districts on the borders of Munich. Furthermore, it appears that the dark colors of the upper quartile are more concentrated in the center of Munich than the other two lower quantiles, while all quartiles have similar patterns in the other districts. This provides interesting evidence of large variation in the central districts of Munich, whereas some of the suburban districts are more homogeneous.

The estimated linear effects as well as their posterior standard deviations are listed in Table 5. In general, all the estimated coefficients are significant, and they explain the response variable appropriately. It is also interesting to see that a particular variable may contribute to different quantiles in different ways. As τ increases, for example, variables Keine.Zh and Kein.Badkach have significantly increasing effects, while Keine.Wwv has a decreasing effect. This means that the absence of a warm water supply would yield a larger rent reduction in the upper quartile than in the lower quartile, while the absences of a central heating system and flagging in the bathroom have reverse effects on quartiles. Presumably, a warm water supply is a basic feature with which only a few flats are not equipped with, and those flats tend to be cheap; however, central heating and flagging are less common, and flats without them could be expensive.

6. Final remarks

In this paper, we have proposed a unified approach for Bayesian quantile inference via MCMC and INLA methods in additive mixed models. Different types of covariate are all treated within the same general framework by assigning appropriate Gaussian-type priors with different forms and degrees of smoothness. Extensive simulation studies demonstrate our approach in terms of both accuracy and computational efficiency. The Munich rental example shows the effectiveness of

Table 5
Munich rental guide: estimated linear effects and their posterior standard deviations.

Variable	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
Gute.Wohnlage	0.578 (0.127)	0.712 (0.110)	0.674 (0.108)
Beste.Wohnlage	1.536 (0.306)	1.582 (0.317)	1.892 (0.323)
Keine.Wwv	−1.526 (0.293)	−1.972 (0.277)	−2.360 (0.298)
Keine.Zh	−1.622 (0.211)	−1.486 (0.202)	−1.123 (0.214)
Kein.Badkach	−0.374 (0.281)	−0.592 (0.108)	−0.597 (0.113)
Besond.Bad	0.479 (0.166)	0.358 (0.153)	0.442 (0.151)
Gehobene.Kueche	1.026 (0.210)	1.152 (0.166)	1.146 (0.172)
zim1	7.062 (0.248)	8.129 (0.242)	9.104 (0.246)
zim2	7.015 (0.211)	8.261 (0.181)	9.633 (0.181)
zim3	6.861 (0.197)	8.118 (0.164)	9.485 (0.169)
zim4	6.497 (0.174)	7.697 (0.182)	9.006 (0.186)
zim5	6.284 (0.349)	7.622 (0.337)	9.473 (0.349)
zim6	6.381 (0.574)	7.513 (0.527)	8.930 (0.566)

our model in accounting for nonlinear and spatial effects from a large dataset. Although we have not presented any example here, the proposed model can be extended to make quantile inference in varying coefficient models and semiparametric mixed models for longitudinal data.

As pointed out in Cai (in press), it is difficult to use the proposed model to estimate extreme quantiles. This is because as τ tends to extremes, no data are available. Also, the parameter τ determines both the skewness and τth quantile for the asymmetric Laplace density, hence limiting its flexibility in modeling tails of an error distribution. To allow more flexible tail behavior of the STAQ model, one possible solution is to replace the global scale parameter δ_0 in (6) with local parameters δ_i associated with the response observation y_i, as suggested in Kottas and Krnjajić (2009). Then, it becomes crucial to choose appropriate priors for δ_i. The Dirichlet process priors used by Kottas and Krnjajić (2009) seem to be a nice option. Further investigation on this issue will be conducted in the future.

Another possible future work is to use the STAQ model to make quantile inference for count or binary data. It is well known that the transformed quantile regression is equivalent to the quantile regression of the transformed variable, i.e., $Q(t(y)) = t(Q(y))$, where $t(\cdot)$ denotes a monotonic transformation and $Q(\cdot)$ denotes the quantile. There may exist a transformation that can create an appropriate continuous response variable that has the same quantiles as the original discrete data, so we can apply the model developed here to the transformed data.

Acknowledgements

Yu Yue's research has been supported by PSC-CUNY research award #60147-39 40. We thank Nora Fenske and Thomas Kneib for their assistance in the simulation studies and Munich rental data example. We also thank two referees for their extremely helpful and insightful comments on our manuscript, which has significantly improved its quality.

Appendix

In order to make the results of our data analysis reproducible, we provide the following R code for estimating the median function in the Munich rental example. Other quantiles can be obtained via changing the value of alpha in the function inla().

```
> library(INLA)
> data(Munich)
> g = system.file("demodata/munich.graph", package="INLA")
> formula = rent ~ f(location,model="besag",graph.file=g,param=c(1,0.001)) +
                   f(year,model="crw2",values=seq(1918,2001),param=c(1,0.001)) +
                   f(floor.size,model="crw2",param=c(1,0.001)) +
                   Gute.Wohnlage + Beste.Wohnlage + Keine.Wwv + Keine.Zh +
                   Kein.Badkach  + Besond.Bad + Gehobene.Kueche +
                   zim1 + zim2 + zim3 + zim4 + zim5 + zim6 -1
> mod = inla(formula,data = Munich, verbose = T,  family = "laplace",
                 control.data=list(alpha=0.5, gamma=2, epsilon=0.01),
                 control.predictor = list(initial = 12), control.inla = list(h=1e-3))
```

References

Besag, J., Kooperberg, C., 1995. On conditional and intrinsic autoregressions. Biometrika 82, 733–746.
Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian *P*-splines. Computational Statistics and Data Analysis 50, 967–991.
Cai, Y., 2009. Polynomial power-Pareto quantile function models. Extremes (in press).
Cai, Z., Xu, X., 2008. Nonparametric quantile estimations for dynamic smooth coefficient models. Journal of the American Statistical Association 103, 1595–1608.

Chhikara, R.S., Folks, L., 1989. The Inverse Gaussian Distribution: Theory, Methodology, and Applications. Marcel Dekker, New York.
De Gooijer, J., Zerom, D., 2003. On additive conditional quantiles with high-dimensional covariates. Journal of the American Statistical Association 98, 135–146.
Dunson, D., Taylor, J., 2005. Approximate Bayesian inference for quantiles. Journal of Nonparametric Statistics 17, 385–400.
Eubank, R.L., Huang, C., Maldonado, Y.M., Wang, N., Wang, S., Buchanan, R.J., 2004. Smoothing spline estimation in varying-coefficient models. Journal of the Royal Statistical Society, Series B: Statistical Methodology 66, 653–667.
Fahrmeir, L., Kneib, T., Lang, S., 2004. Penalized structured additive regression for space–time data: a Bayesian perspective. Statistica Sinica 14, 731–761.
Fahrmeir, L., Lang, S., 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. Journal of the Royal Statistical Society, Series C: Applied Statistics 50, 201–220.
Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modeling Based on Generalized Linear Models. Springer, Berlin.
Fenske, N., Kneib, T., Hothorn, T., 2009. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Technical Report 52. Department of Statistics, University of Munich.
Geraci, M., Bottai, M., 2007. Quantile regression for longitudinal data using the asymmetric laplace distribution. Biostatistics 8, 140–154.
Green, P.J., 1990. Bayesian reconstructions from emission tomography data using a modified EM algorithm. IEEE Transactions on Medical Imaging 9, 84–93.
Horowitz, J.L., Lee, S., 2005. Nonparametric estimation of an additive quantile regression model. Journal of the American Statistical Association 100, 1238–1249.
Kneib, T., Konrath, S., Fahrmeir, L., 2010. High-dimensional structured additive regression models: Bayesian regularisation, smoothing and predictive performance. Journal of the Royal Statistical Society, Series C: Applied Statistics (in press).
Koenker, R., 2005. Quantile Regression. In: Econometric Society Monographs, Cambridge University Press.
Koenker, R., Bassett Jr., Gilbert, 1978. Regression quantiles. Econometrica 46, 33–50.
Koenker, R., Machado, J.A.F., Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94, 1296–1310.
Koenker, R., Mizera, I., 2004. Penalized triograms: total variation regularization for bivariate smoothing. Journal of the Royal Statistical Society, Series B: Statistical Methodology 66, 145–163.
Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.
Kottas, A., Krnjajić, M., 2009. Bayesian semiparametric modeling in quantile regression. Scandinavian Journal of Statistics 36, 297–319.
Lang, S., Brezger, A., 2004. Bayesian *P*-splines. Journal of Computational and Graphical Statistics 13, 183–212.
Li, Y., Liu, Y., Zhu, J., 2007. Quantile regression in reproducing kernel Hilbert spaces. Journal of the American Statistical Association 102, 255–268.
Martino, S., Rue, H., 2008. Implementing approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations: a manual for the inla -program. Technical Report No. 2. Department of Mathematical Sciences, Norwegian University of Science and Technology.
Reich, B., Bondell, H., Wang, H., 2010. Flexible Bayesian quantile regression for independent and clustered data. Biostatistics 11, 337–352.
Rue, H., 2001. Fast sampling of Gaussian Markov random fields. Journal of the Royal Statistical Society, Series B 63, 325–338.
Rue, H., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. In: Monographs on Statistics and Applied Probability, vol. 104. Chapman & Hall, London.
Rue, H., Martino, S., 2007. Approximate Bayesian inference for hierarchical Gaussian Markov random fields. Journal of Statistical Planning and Inference 137, 3177–3192.
Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). Journal of the Royal Statistical Society, Series B 71, 319–392.
Taddy, M., Kottas, A., 2009. A Bayesian nonparametric approach to inference for quantile regression. Journal of Business and Economic Statistics (in press).
Takeuchi, I., Le, Q., Sears, T., Smola, A., 2006. Nonparametric quantile estimation. Journal of Machine Learning Research 7, 1231–1264.
Tsionas, E.G., 2003. Bayesian quantile inference. Journal of Statistical Computation and Simulation 73, 659–674.
Wahba, G., 1990. Spline Models for Observational Data. SIAM [Society for Industrial and Applied Mathematics].
Wecker, W.E., Ansley, C.F., 1983. The signal extraction approach to nonlinear regression and spline smoothing. Journal of the American Statistical Association 78, 81–89.
Yu, K., Lu, Z., 2004. Local linear additive quantile regression. Scandinavian Journal of Statistics 31, 333–346.
Yu, K., Moyeed, R.A., 2001. Bayesian quantile regression. Statistics & Probability Letters 54, 437–447.
Yu, K., Zhang, J., 2005. A three-parameter asymmetric laplace distribution and its extension. Communications in Statistics: Theory and Methods 34, 1867–1879.
Yuan, Y., Yin, G., 2010. Bayesian quantile regression for longitudinal studies with nonignorable missing data. Biometrics 66, 105–114.