<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Program quality with pair programming in CS1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Krnjaji c, Milovan</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2005</td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/3805">http://hdl.handle.net/10379/3805</a></td>
</tr>
</tbody>
</table>

Some rights reserved. For more information, please see the item record link above.
Model-based approaches to nonparametric Bayesian quantile regression

Athanasis Kottas\textsuperscript{1}, Milovan Krunjadi\textsuperscript{2}, Matthew Taddy\textsuperscript{3}
University of California, Santa Cruz\textsuperscript{1}
University of California, Lawrence Livermore National Laboratory\textsuperscript{2}
University of California, Santa Cruz\textsuperscript{3}

\textbf{Abstract}

In several regression applications, a different structural relationship might be anticipated for the higher or lower responses than the average responses. In such cases, quantile regression analysis can uncover important features that would likely be overlooked by mean regression. We develop two distinct Bayesian approaches to fully nonparametric model-based quantile regression. The first approach utilizes an additive regression framework with Gaussian process priors for the quantile regression functions and a scale uniform Dirichlet process mixture prior for the error distribution, which yields flexible unimodal error density shapes. Under the second approach, the joint distribution of the response and the covariates is modeled with a Dirichlet process mixture of multivariate normals, with posterior inference for different quantile curves emerging through the conditional distribution of the response given the covariates. The proposed nonparametric prior probability models allow the data to uncover non-linearities in the quantile regression function and non-standard distributional features in the response distribution. Inference is implemented using a combination of posterior simulation methods for Dirichlet process mixtures. We illustrate the performance of the proposed models using simulated and real data sets.

\textbf{Keywords}: Dirichlet process mixture models; Gaussian process priors; Multivariate normal mixtures; Scale uniform mixtures

1. Introduction

Quantile regression can be used to quantify the relationship between a set of quantiles of the response distribution and available covariates. Because, in general, a set of quantiles provides a more complete description of the response distribution than the mean, quantile regression offers a practically important alternative to traditional mean regression. In many regression examples (e.g., in econometrics, educational studies, and environmental applications), we might expect a different structural relationship for the higher (or lower) responses than the average responses. In such applications, mean, or median, regression approaches would likely overlook important features that could be uncovered by a more general quantile regression analysis.

There is a fairly extensive literature on classical estimation for the standard \(p\)-th quantile regression model, \(y_i = \mathbf{x}_i^T \beta + \epsilon_i\), where \(y_i\) denotes the response observations, \(\mathbf{x}_i\) the corresponding covariate vectors, and \(\epsilon_i\) the errors, which are typically assumed independent from a distribution (with density, say, \(f_p(\cdot)\)) that has \(p\)-th quantile equal to 0. (See, e.g., the review paper by Yu, Lu and Stander, 2003, and the book by Koenker, 2005.) This literature is dominated by semiparametric techniques where the error density \(f_p(\cdot)\) is left unspecified (apart from the restriction \(\int_{-\infty}^{0} f_p(\cdot)\, d\mathbb{P} = p\)). Hence, since there is no probability model for the response distribution, point estimation for \(\beta\) proceeds by optimization of some loss function. For instance, under the standard setting with independent and uncensored responses, the point estimates for \(\beta\) minimize \(\sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^T \beta)\), where \(\rho_p(u) = up\, u I_{(-\infty,0)}(u)\); this form yields the least absolute deviations criterion for \(p = 0.5\), i.e., for the special case of median regression. Any inference beyond point estimation is based on asymptotic arguments or resampling methods. The classical literature includes also work that relaxes the parametric (linear) regression form for the quantile regression function (see, e.g., He, Ng and Portnoy, 1998; Horowitz and Lee, 2005).

By comparison with the existing volume of classical work, the Bayesian literature on quantile regression is relatively limited. The special case of median regression has been considered in Walker and Mallick (1999), Kottas and Gelfand (2001), and Hanson and Johnson (2002). This work is based on a parametric form for the median regression function and nonparametric modeling for the error distribution, using either Pólya tree priors or Dirichlet process mixture priors. (See, e.g., Müller and Quintana, 2004, for reviews of these nonparametric prior models.) Regarding quantile regression, based again on parametric regression functions, Yu and Moyeed (2001) and Tsionas (2003) discuss parametric inference based on the asymmetric Laplace distribution for the errors; Kottas and Krunjadi (2005) develop Bayesian semiparametric models using Dirichlet process mixtures for the error distribution; Hjort and Petrone (2005) study nonparametric inference for the quantile function based on Dirichlet processes, including brief discussion of the semiparametric extension to quantile regression. Moreover, Dunson and Taylor (2005) propose an approximate semi-Bayesian inference method for quantile regression, which, in contrast to the work discussed above, does not involve probabilistic modeling for the response distribution.
The error terms \( \varepsilon_i \) are assumed independent from a distribution with \( p \)-th quantile equal to 0, i.e., \( \int_{-\infty}^{0} f_p(\varepsilon) d\varepsilon = p \), with \( f_p(\cdot) \) denoting the error density.

Our objective is to combine flexible nonparametric prior models for the quantile regression function and the random error density.

In particular, we work with independent Gaussian process (GP) priors for \( h_1(\cdot) \) and \( h_2(\cdot) \). (See, e.g., Neal, 1997, 1998, on GP regression under parametric error distributions.) To avoid identifiability issues, we set the GP mean functions to zero, \( E(h_\ell(x)) = 0 \), for all \( x \). We have also observed empirically that an intercept in (1) is weakly identifiable, especially, for data sets that support skewness in the error distributions. Rather than working with a random intercept with a very informative prior, we opt to fix it to zero centering the response observations before fitting the model. For the GP covariance functions, we use the simple isotropic specification,

\[
\text{Cov}(h_\ell(x), h_{\ell'}(x')) = \tau^2 \exp(-\phi|x-x'|^a),
\]

with random variances \( \tau^2 \) and range parameters \( \phi > 0 \), and fixed power parameters \( a_\ell \in [1, 2], \ell = 1, 2 \).

For the random error density \( f_p(\cdot) \) we use the scale uniform Dirichlet process (DP) mixture prior from Kottas and Krnjajić (2005). The key result for its construction is a representation for non-increasing densities on the positive real line. Specifically, for any non-increasing density \( u(\cdot) \) on \( \mathbb{R}^+ \) there exists a distribution function \( G(\cdot) \) with support on \( \mathbb{R}^+ \), such that \( u(t) \equiv u(t; G) = \int_0^t R \, dG(\theta)^{1-1} \, dG(\theta)^1 \), \( \theta \). For any general mixing distribution \( G(\cdot) \) and thus, for Bayesian modeling, invokes the use of a nonparametric prior for \( G(\cdot) \); see, e.g., Brunner and Lo (1989), Brunner (1995), Lavine and Mack (1995), and Kottas Gelfand (2001) for DP-based modeling involving variations of this representation.

In the context of the quantile regression setting in (1), this result can be used to provide a mixture representation for any unimodal density on the real line with \( p \)-th quantile (and mode) equal to zero, \( \int k_p(\varepsilon; \sigma_1, \sigma_2) dG_1(\sigma_1) dG_2(\sigma_2) \). Here \( G_1 \) and \( G_2 \) are general mixing distributions, supported on \( \mathbb{R}^+ \), and

\[
k_p(\varepsilon; \sigma_1, \sigma_2) = \frac{p}{\sigma_2} 1_{(-\sigma_0, \sigma_0)}(\varepsilon) + \frac{(1-p)}{\sigma_2} 1_{[0, \sigma_2]}(\varepsilon),
\]

with \( 0 < p < 1 \), and \( \sigma_0 > 0, \sigma > 1, 2 \). Assuming independent DP priors, \( D\varphi_{\sigma_1}, G_0 \), for \( G_r, r = 1, 2 \), we obtain the DP mixture model

\[
f_p(\varepsilon; G_1, G_2) = \int \int k_p(\varepsilon; \sigma_1, \sigma_2) dG_1(\sigma_1) dG_2(\sigma_2),
\]

for the error density in (1).

Recall that the DP was developed by Ferguson (1973) as a prior probability model for random distributions.
(equivalently, distribution functions) $G$. A DP $(\alpha, G_0)$ prior for $G$ is defined in terms of two parameters, a parametric base distribution $G_0$ (the mean of the process) and a positive scalar parameter $\alpha$, which can be interpreted as a precision parameter; larger values of $\alpha$ result in realizations $G$ that are closer to $G_0$. We will write $G \sim \text{DP}(\alpha, G_0)$ to indicate that a DP prior is used for the random distribution $G$. In fact, DP-based modeling typically utilizes mixtures of DPs (Antoniak 1974), i.e., a more general version of the DP prior that involves hyperparameters for the random distribution $G$, and the priors for the shape parameters $\delta_1$ and random scale parameters $d_r$.

Specifically, a random distribution $G$ generated from DP$(\alpha, G_0)$ is (almost surely) of the form

$$G(\cdot) = \sum_{i=1}^{\infty} w_i \delta_{\theta_i}(\cdot)$$

where $\delta_y(\cdot)$ denotes a point mass at $y$. The locations of the point masses, $\theta_i$, are i.i.d. realizations from $G_0$; the corresponding weights, $w_i$, arise from a stick-breaking mechanism based on i.i.d. draws $\{\zeta_k : k = 1, 2, \ldots\}$ from a Beta$(1, \alpha)$ distribution. In particular, $w_1 = \zeta_1$, and, for each $\ell = 2, 3, \ldots$, $w_\ell = \zeta_1 \prod_{k=1}^{\ell-1} (1 - \zeta_k)$. Moreover, the sequences $\{\theta_\ell : \ell = 1, 2, \ldots\}$ and $\{\zeta_k : k = 1, 2, \ldots\}$ are independent.

To complete the DP mixture model specification in (3), we take inverse gamma distributions for $G_{r0}$ with fixed shape parameters $c_r$ and random scale parameters $d_r$, $r = 1, 2$, which are assigned gamma priors. (We use the parameterization of the inverse gamma distribution with mean $d_r/(c_r - 1)$, provided $c_r > 1$.) Gamma priors are also placed on the DP precision parameters $\alpha_r$, $r = 1, 2$.

To break the mixture in (3), we introduce latent mixing parameters $\sigma_{\delta_1}$ and $\sigma_{\delta_2}$, corresponding to $y_1$, and express the full Bayesian model in the following hierarchical form:

$$
\begin{align*}
\delta_{i1} &\sim \text{h}_1(x_{i1}), \\
\delta_{i2} &\sim \text{h}_2(x_{i2}), \quad \text{for } i = 1, \ldots, n, \quad \ell = 1, 2, \quad r = 1, 2,
\end{align*}
$$

for $i = 1, \ldots, n$, $\ell = 1, 2$, and $r = 1, 2$, where $\text{h}_1(x, x') = \text{Cov}(h_1(x), h_1(x'))$ and $\text{h}_2(x, x') = \text{Cov}(h_2(x), h_2(x'))$ denote the covariance functions in (2). Moreover, priors are placed on the GP and DP hyperparameters $\psi = (\tau_1^2, \phi_1, \tau_2^2, \phi_2, \alpha_1, d_1, \alpha_2, d_2)$ (see Section 2.2 regarding the priors for the $\tau_j^2, \phi_j$).

To obtain posterior predictive inference, we work with the version of (4) that involves the normal finite dimensional distributions for the $h(x_{i\ell})$ induced by the GP priors, and the priors for the $\sigma_{\delta\ell}$ induced by marginalizing the random distributions $G_{r}$ over their DP priors (Blackwell and MacQueen, 1973). Let $\delta_{i1} = h_1(x_{i1})$ and $\delta_{i2} = h_2(x_{i2})$, for $i = 1, \ldots, n$. Then, the nonparametric model (4) induces the following hierarchical model that involves only finite dimensional parameters:

$$
\begin{align*}
\delta_{i1} \sim &\text{h}_1(x_{i1}), \\
\delta_{i2} \sim &\text{h}_2(x_{i2}), \\
\sigma_{\delta1} &\sim \text{GP}(0, C_1(\tau_1^2, \phi_1)), \\
\sigma_{\delta2} &\sim \text{GP}(0, C_2(\tau_2^2, \phi_2)), \\
G_1 &\sim \text{DP}(\alpha_1, G_{r0}(d_1)), \\
G_2 &\sim \text{DP}(\alpha_2, G_{r0}(d_2)).
\end{align*}
$$

for $i = 1, \ldots, n$, $\ell = 1, 2$, and $r = 1, 2$, again, with priors for the components of $\psi$. Here, the covariance matrix $C_1(\tau_1^2, \phi_1)$ is induced by the covariance function $C_1(\tau_1^2, \phi_1)$, i.e., its $(i, j)$-th element is given by $\text{Cov}(h_1(x_{i\ell}), h_1(x_{j\ell})) \mid \tau_1^2, \phi_1) = \tau_1^2 \exp(-\phi_1|x_{i\ell} - x_{j\ell}|^2)$. Moreover, the priors $p(\sigma_{\delta1}, \sigma_{\delta2} | \alpha_r, d_r)$ are constructed based on the standard DP Polya urn scheme. Specifically, for $r = 1, 2$, $p(\sigma_{\delta1}, \sigma_{\delta2} | \alpha_r, d_r)$ can be represented in terms of successive complete conditionals, with $\sigma_{\delta1} \sim G_{r0}$, and for each $i = 1, \ldots, n$, $p(\sigma_{\delta1} | \sigma_{\delta1}, \sigma_{\delta2}, \alpha_r, d_r)$ given a mixed distribution with point masses $(\alpha_r + i - 1)^{-1}$ at the $\sigma_{\delta1}$, for $j = 1, \ldots, i - 1$, and continuous mass $\alpha_r(\alpha_r + i - 1)^{-1}$ on $G_{r0}(d_r)$.

### 2.2 MCMC posterior simulation

We describe here an MCMC algorithm to obtain posterior samples from $p(\delta_{11}, \delta_{12}, \sigma_{\delta1}, \sigma_{\delta2} | \psi, \text{data})$, where $\delta_{11} = (\delta_{11}, \delta_{12})$, $\ell = 1, 2$, $\sigma_{\delta1} = (\sigma_{\delta1}, \sigma_{\delta2})$, and $\text{data} = \{y_i, x_{1i}, x_{2i} : i = 1, \ldots, n\}$. These posterior samples can then be used to estimate posterior predictive distributions as is discussed in this section.

Updating the $\sigma_{\delta\ell}$, conditionally on the $\delta_{\ell1}$ and all other relevant parameters, proceeds exactly as in the semiparametric model developed in Kottas and Krajalic (2005), the only difference being that $x_{\ell1} \beta$ from the semiparametric model is replaced by $h_{\ell}(x_{\ell1}) + h_{\ell}(x_{\ell2})$. The approach is also the same for updating the $\alpha_r$ and $d_r$.

For each $i = 1, \ldots, n$, the posterior full conditional for $\delta_{i1}$ is given by

$$p(\delta_{i1} | \ldots, \text{data}) \propto N(\delta_{1i}; m_1, V_1)k_r(u_i - \delta_{i1}; \sigma_{\delta1}, \sigma_{\delta2})$$

where $u_i = y_i - \delta_{i2}$, and $N(\delta_{1i}; m_1, V_1)$ is the prior full conditional corresponding to the $n$-variate normal prior for $\delta_{1i}$. Hence, letting $\delta_{1i} = (\delta_{1i} : i' \neq i)$, we have $m_1 = bB^{-1}\delta_{1i} - V_1 = \tau_1^2 - bB^{-1}b^T$. Here, $b$ is the $1 \times (n - 1)$ vector with elements $\tau_1^2 \exp(-\phi_1|x_{i1} - x_{i1}|^2)$ for $i' \neq i$, and $B$ is the $(n - 1) \times (n - 1)$ matrix with elements $\tau_1^2 \exp(-\phi_1|x_{i1} - x_{i1}|^2)$ for $(j, k)$ with $j \neq i$ and $k \neq i$. Now, $p(\delta_{11} | \ldots, \text{data}) \propto N(\delta_{11}; m_1, V_1) \{1 - (1 - p_2)^{-1}1_{(u_i - \sigma_{\delta1}, u_i)}(\delta_{1i}) + p_21_{(u_i - u_i)}(\delta_{1i})\}$, and therefore letting $E_1 = \int_{u_i - \sigma_{\delta1}}^{u_i} N(\delta_{1i}; m_1, V_1)d\delta$, and $E_2 = \int_{u_i}^{\infty} N(\delta_{1i}; m_1, V_1)d\delta$, we obtain that $p(\delta_{11} | \ldots, \text{data})$ is a two-component mixture of truncated normal distributions, $TN(m_1, V_1; \delta_{11} \in (u_i - \sigma_{\delta1}, u_i))$ and $TN(m_1, V_1; \delta_{11} \in (u_i, u_i + \sigma_{\delta1}))$ with weight, associated with the first component, given by $q = \frac{E_1}{E_1 + E_2}$.
Metropolis-Hastings steps can be used to update the sampled directly as a discrete distribution (in which case, to be uniform on $(0, \infty)$ similar to the one above for the prior for $\gamma_1$, $\gamma_2$). Then, the posterior full conditional for $\gamma_i$ can be used to specify a plausible upper bound $b_{\gamma}$ for its uniform prior. For instance, under exponential covariance functions (i.e., $a_{\ell} = 1$ in (2)), $3/\phi_{\ell}$ is the range of dependence (the value of the distance between covariate values that yields correlation about 0.05). If $z_{\ell}$ is the range of the observed $x_{\ell}$, we can use, say, $z_{\ell}/2$ as a guess at the range of dependence, and thus set accordingly the prior upper bound for the $\phi_{\ell}$.

To obtain inference for the quantile regression functions, of direct interest is the posterior predictive distribution for $h_{\ell}(x_{0\ell})$, $\ell = 1, 2$, for any new covariate values $x_{01}$ and $x_{02}$. Of interest is also the posterior predictive error density and the posterior predictive distribution for a new response observation $y_0$. Denote by $(\sigma_{10}, \sigma_{20})$ the new latent mixing parameters corresponding to $y_0$, and also let $\delta_{\ell} = h_{\ell}(x_{0\ell})$, for $\ell = 1, 2$.

Adding $(\sigma_{10}, \sigma_{20})$, $\delta_{01}$, $\delta_{02}$ and $y_0$ to model (4), and marginalizing the $G_{\ell}$ over their DP priors, the corresponding joint posterior $p(y_0, \sigma_{10}, \sigma_{20}, \delta_{01}, \delta_{02}, \delta_{1}, \delta_{2}, \gamma_1, \gamma_2, \psi \mid x_{01}, x_{02}, \text{data})$ is

\[(1-p)\sigma_{2i}^{-1}E_1)/(1-p)\sigma_2^{-1}E_1 + p\sigma_{ii}^{-1}E_2).\] Note that although this is not a standard mixture, in that the components have different support, it can be easily sampled, since it only requires a truncated normal random variate.

For each $i = 1, \ldots, n$, the posterior full conditional for $\delta_{i2}$, $p(\delta_{i2} \mid \ldots, \text{data}) \propto N(\delta_{i2}; m_2, V_2) h_{\ell}(v_i - \delta_{i2}; \gamma_1, \gamma_2)$, where $v_i = y_i - \delta_{i1}$, and $N(\delta_{i2}; m_2, V_2)$ is the prior full conditional for $\delta_{i2}$ resulting from the $N(0, S_{2}(\gamma_2, \phi_2))$ prior for $\gamma_2$. Hence, the approach to update the $\delta_{i2}$ is similar to the one above for the $\delta_{i1}$.

For $\ell = 1, 2$, assume an inverse gamma prior for $\gamma^2$ with shape parameter $a_{\gamma}$ and scale parameter $b_{\gamma}$. Moreover, write $S_{\ell}(\gamma^2, \phi_{\ell}) = \gamma^2 \cdot R(\phi_{\ell})$, where $R(\phi_{\ell})$ is the correlation matrix with $(i,j)$-th element $\exp(-\phi_{\ell} | x_{i\ell} - x_{j\ell}|^{\tau_{i\ell}})$. Then, the posterior full conditional for $\gamma^2$ is an inverse gamma distribution with shape parameter $a_{\gamma} + 0.5n$ and scale parameter $b_{\gamma} + 0.5\delta_{\ell} R^{-1}(\phi_{\ell}) \delta_{\ell}$.

Finally, for $\ell = 1, 2$, the posterior full conditional for $\phi_{\ell}$ is proportional to $p(\phi_{\ell}) | R(\phi_{\ell}) |^{-1/2} \exp\{- (\delta_{\ell}^2 R^{-1}(\phi_{\ell}) \delta_{\ell})/(2 \gamma^2)\}$ where $p(\phi_{\ell})$ is the prior for $\phi_{\ell}$, $\ell = 1, 2$, which is taken to be uniform on $(0, b_{\phi})$. Evidently, this form does not lead to an expression that can be sampled directly.

Metropolis-Hastings steps can be used to update the $\phi_{\ell}$. Alternatively, the full conditional can be discretized and sampled directly as a discrete distribution (in which case, the prior for $\phi_{\ell}$ is discrete uniform on $(0, b_{\phi})$). This latter approach is, in general, more efficient computationally, since the computations involving matrix $R(\phi_{\ell})$ can be performed before the beginning of the MCMC steps. Regardless, because the data typically cannot inform strongly about this GP hyperparameter, we need to specify the priors for $\phi_{\ell}$ with some care. The interpretation of the range parameter $\phi_{\ell}$ can be used to specify a plausible upper bound $b_{\phi}$ for its uniform prior. For instance, under exponential covariance functions (i.e., $a_{\ell} = 1$ in (2)), $3/\phi_{\ell}$ is the range of dependence (the value of the distance between covariate values that yields correlation about 0.05). If $z_{\ell}$ is the range of the observed $x_{\ell}$, we can use, say, $z_{\ell}/2$ as a guess at the range of dependence, and thus set accordingly the prior upper bound for the $\phi_{\ell}$.

Figure 1: Median regression simulation example with normal errors. The blue lines denote point and interval estimates for the regression function (top panel) and the posterior predictive error density (bottom panel).
Here, \( p(\delta_{0\ell} \mid \delta_0, \tau_0, \phi_0) \) are the conditional normal distributions that result from the \((n+1)\)-variate normal distributions for \((\delta_{0\ell}, \delta_0)\) induced by the GP priors, for \( \ell = 1, 2 \). Hence, the mean of \( p(\delta_{0\ell} \mid \delta_0, \tau_0, \phi_0) \) is given by \( \mathbf{b}_{0\ell}S_{\ell}^{-1}(\tau_0^2, \phi_0)\mathbf{d}_\ell \) and the variance by \( \tau_0^2 - \mathbf{b}_{0\ell}S_{\ell}^{-1}(\tau_0^2, \phi_0)\mathbf{b}_{0\ell}^T \), where \( \mathbf{b}_{0\ell} \) is the \( 1 \times n \) vector with elements \( \tau_0^2 \exp(-\phi_0|x_{0\ell} - x_i|^2) \), for \( i = 1, \ldots, n \). Moreover, \( p(\sigma_{r0} \mid \sigma_r, \alpha_r, d_r) \) arises from the DP Polya urn structure, i.e., for \( r = 1, 2 \), \( p(\sigma_{r0} \mid \sigma_r, \alpha_r, d_r) \) can be expressed as

\[
\frac{\alpha_r}{\alpha_r + n}G_{r0}(\sigma_{r0} \mid d_r) + \frac{1}{\alpha_r + n} \sum_{j=1}^{n_r^*} \delta_{\sigma_{rj}}(\sigma_{r0}),
\]

where \( \sigma_{rj}^* \), \( j = 1, \ldots, n_r^* \), are the distinct values in the vector \((\sigma_{r1}, \ldots, \sigma_{rn})\) and \( n_{rj} \) is the size of the \( j \)-th cluster.

Now, the posterior predictive density for \( y_0 \), \( p(y_0 \mid \mathbf{x}_{01}, \mathbf{x}_{02}, \text{data}) \), is given by

\[
\int \int \int k_p(y_0 - \delta_{01} - \delta_{02}; \sigma_{10}, \sigma_{20})p(\delta_{01} \mid \delta_1, \tau_1^2, \phi_1)p(\delta_{02} \mid \delta_2, \tau_2^2, \phi_2)p(\sigma_{10} \mid \sigma_1, \alpha_1, d_1)p(\sigma_{20} \mid \sigma_2, \alpha_2, d_2)
\]

\[
p(\delta_1, \delta_2, \sigma_1, \sigma_2, \psi \mid \text{data})
\]

\[
d\delta_0d\delta_{01}d\sigma_{01}d\sigma_{20}d\sigma_1d\sigma_2d\psi
\]

and the posterior predictive error density,

\[
p(\epsilon_0 \mid \text{data}) = \int \int \int k_p(y_0 - \delta_{01} - \delta_{02}; \sigma_{10}, \sigma_{20})p(\delta_{01} \mid \delta_1, \tau_1^2, \phi_1)p(\delta_{02} \mid \delta_2, \tau_2^2, \phi_2)p(\sigma_{10} \mid \sigma_1, \alpha_1, d_1)p(\sigma_{20} \mid \sigma_2, \alpha_2, d_2)
\]

\[
p(\sigma_{10} \mid \sigma_1, \alpha_1, d_1)p(\sigma_{20} \mid \sigma_2, \alpha_2, d_2)p(\sigma_1, \sigma_2, \psi \mid \text{data})
\]

\[
d\sigma_{10}d\sigma_{20}d\sigma_1d\sigma_2d\psi.
\]

Hence, both can be easily sampled using the expressions for \( p(\delta_{0\ell} \mid \delta_0, \tau_0, \phi_0) \) and \( p(\sigma_{r0} \mid \sigma_r, \alpha_r, d_r) \) and the posterior samples from \( p(\delta_1, \delta_2, \sigma_1, \sigma_2, \psi \mid \text{data}) \).

Finally, for \( \ell = 1, 2 \), the posterior predictive distribution for \( \delta_{0\ell} \),

\[
p(\delta_{0\ell} \mid \mathbf{x}_{0\ell}, \text{data}) = \int \int \int \int k_p(y_0 - \delta_{01} - \delta_{02}; \sigma_{10}, \sigma_{20})p(\delta_{01} \mid \delta_1, \tau_1^2, \phi_1)p(\delta_{02} \mid \delta_2, \tau_2^2, \phi_2)p(\sigma_{10} \mid \sigma_1, \alpha_1, d_1)p(\sigma_{20} \mid \sigma_2, \alpha_2, d_2)
\]

\[
p(\sigma_{10} \mid \sigma_1, \alpha_1, d_1)p(\sigma_{20} \mid \sigma_2, \alpha_2, d_2)p(\sigma_1, \sigma_2, \psi \mid \text{data})
\]

\[
d\delta_{01}d\delta_{02}d\sigma_{01}d\sigma_{20}d\sigma_1d\sigma_2d\psi.
\]

As illustrated in Section 2.3, samples from \( p(\delta_{0\ell} \mid \mathbf{x}_{0\ell}, \text{data}) \) over a grid of points \( x_0 \) yield (pointwise) point and interval estimates for the quantile regression functions \( h_\ell(\cdot), \ell = 1, 2 \).

### 2.3 Simulation study

Here, we consider simulated data examples to study the performance of the model developed in Section 2.1, fo-
Figure 3: Simulation example for 25-th quantile regression. The blue lines denote posterior point and interval estimates for the regression function (top panel) and the posterior predictive error density (bottom panel).

cusing on the case with a single covariate, $x$. In all examples, we take $n = 150$, and work with the non-linear regression function used by Neal (1997) in the context of mean regression with GP priors and either normal or $t$ error distributions. Specifically, the true quantile regression function for all simulated data sets is given by

$$h(x) = 0.4x + 0.5\sin(2.7x) + 1.1(1 + x^2)^{-1}.$$ 

The covariate values are generated from a standard normal distribution, while response values $y$ are drawn by adding errors to $h(x)$ based on four different choices for the error distribution, resulting in four simulated data sets. In each case, the simulated response and covariate values are plotted in the top panel of Figures 1 – 4, where also plotted is the true regression function (denoted by the red line). The bottom panels of Figures 1 – 4 include histograms of the simulated errors as well as the corresponding true error density (again, denoted by the red line).

The first data set (Figure 1) involves normal errors in the median regression setting. The other three data sets are based on errors generated from mixtures of normal distributions chosen to yield skewed error densities with a specific quantile (and mode) equal to 0. In particular, the error distribution in the second data example (Figure 2) is right skewed; here, again, we focus on the median regression case. In the remaining two data examples, we consider the cases of 25-th quantile regression (Figure 3) and 75-th quantile regression (Figure 4), with right skewed and left skewed error distributions, respectively.

We apply the single-covariate version of model (4) to the simulated data. For all four data examples, we take an exponential covariance function for the GP prior, i.e., the version of (2) given by $\tau^2 \exp(-\phi|x - x'|)$, with a discrete uniform prior on $(0, 5)$ for $\phi$, and an inverse gamma prior for $\tau^2$ with mean 1 (and shape parameter 2). Regarding the DP prior specification, we take $c_r = 2$ for the shape parameters of the inverse gamma base distributions $G_{r0}$, $r = 1, 2$, and specify the mean of the gamma priors for the scale parameters $d_r$ using a rough range for the response values. Sensitivity analysis with regard to the dispersion of the prior distributions for the $d_r$ revealed robustness for the resulting posterior predictive inference. Finally, a gamma$(4, 0.5)$ prior (with mean 8) was used for $\alpha_1$ and $\alpha_2$.

For each of the data examples, the top panels of Figures 1 – 4 plot pointwise posterior mean estimates and 95% posterior interval estimates for the respective quantile regression functions. The bottom panels of the figures show the corresponding posterior predictive error densities. In all cases, the model captures quite well the shape of both the non-linear regression function and the error density. Particularly encouraging for the potential of the methodology is the fact that the posterior predictive esti-
mates identify successfully the two random functions (the quantile regression function and the error density function) under different types and amounts of skewness in the true error distributions.

3. A fully nonparametric approach to inference for quantile regression

The starting point for the approach of Section 2 (and for most existing approaches to quantile regression) is the standard additive regression framework \( y = h(x) + \epsilon \), where again the errors \( \epsilon \) are assumed independent from a distribution with \( p \)-th percentile equal to 0. Note that, under this framework (and regardless of the formulation for the regression function), if inference is sought for more than one quantile regression, the particular model needs to be fitted separately for each corresponding \( p \). In particular, note that estimated quantile regression functions for nearby values of \( p \) might not satisfy the explicit ordering of the corresponding percentiles, especially with small sample sizes and/or for extreme percentiles. And this attribute of the additive formulation is shared by any approach that utilizes a (proper) probability model for the error distribution, regardless of the estimation method (likelihood or Bayesian). There are certain approaches that allow, in the context of this framework, simultaneous estimation for more than one quantile regression (e.g., Dunson & Taylor, 2005); however, this is only possible because they do not involve modeling for the errors, but are rather based on approximate methods (e.g., certain pseudo-likelihoods). Hence, the additive quantile regression framework is more suitable for applications where interest lies in explaining one percentile (or a few well-separated percentiles) of the response distribution in terms of available covariates. For such settings, by separating the quantile regression function from the errors, the model formulation in (1) allows readily interpretable inference, incorporation of different types of covariates, and extensions of the inference methods to handle censored observations.

In this section, we develop an alternative approach to inference for quantile regression. This approach does not build on a structured regression model formulation as the one in Section 2, but it yields flexible, fully nonparametric inference for quantile regression.

3.1 The modeling approach

The starting point for this approach is to consider a model for the joint distribution of the response, \( y \), and the set of covariates, \( x = (x_1, ..., x_L) \) (again, we focus here on covariate information on continuous variables). Then, inference for any set of quantile curves can be obtained based on the posterior of the implied conditional response.
distribution given the covariates. Clearly, the richness of the resulting inference relies on the flexibility of the prior probability model for the distribution of \( z = (y, x) \). We employ a DP mixture of multivariate normals, which models the joint density for \( z \) through

\[
 f(z; G) = \int N_{L+1}(z; \mu, \Sigma) dG(\mu, \Sigma), \quad G \sim \text{DP}(\alpha, G_0)
\]

with \( G_0 \) built from independent \( N_{L+1}(m, V) \) and \( \text{IWish}(\nu, S) \) components for \( \mu \) and \( \Sigma \), respectively. We work with random \( m \), \( V \) and \( S \) and fixed \( \nu \). Here, \( \text{IWish}(\nu, S) \) denotes the inverse Wishart distribution for the \((L+1) \times (L+1)\) (positive definite) matrix \( \Sigma \) with density proportional to

\[
 |\Sigma|^{-(\nu+L+2)/2} \exp\{-0.5 \text{tr}(S \Sigma^{-1})\}.
\]

This model has been studied in the context of multivariate density estimation as well as curve fitting (see, e.g., Müller, Erkanli and West, 1996). However, the scope of inference has been limited to posterior point estimates, obtained through posterior predictive distributions, i.e., \( p(z_0|\text{data}) = E(f(z_0; G)|\text{data}) \), where, again, data comprises \( \{z_i = (y_i, x_i) : i = 1, ..., n\} \). Our application to quantile regression requires the entire posterior of \( f(z_0; G) \) at any \( z_0 \), and we thus employ a more general approach to MCMC inference (discussed in Section 3.2) that includes sampling from the posterior of \( G \).

The hierarchical model formulation involves, again, latent mixture parameters associated with each vector of response/covariate observations,

\[
 z_i | \mu_i, \Sigma_i \stackrel{\text{iid}}{\sim} N_{L+1}(z_i; \mu_i, \Sigma_i) \\
 \mu_i, \Sigma_i | G \stackrel{\text{iid}}{\sim} G \\
 G | \alpha, \psi \sim \text{DP}(\alpha, G_0(\psi))
\]

for \( i = 1, ..., n \), with hyperpriors for the DP precision parameter \( \alpha \) and for the parameters, \( \psi = (m, V, S) \), of \( G_0 \). In particular, we use a gamma prior for \( \alpha \), a \( N_{L+1}(m_0, B_m) \) prior for \( m \), an \( \text{IWish}(a_S, B_S) \) prior for \( V \), and a \( \text{Wish}(a_S, B_S) \) prior for the \((L+1) \times (L+1)\) positive definite matrix \( S \) with density proportional to \( |S|^{(a_S - L - 2)/2} \exp\{-0.5 \text{tr}(SB_S^{-1})\} \) (provided \( a_S \geq L + 1 \)).

### 3.2 Posterior inference for quantile regression

We describe here the approach to estimate quantile curves based on the posterior for the conditional density \( f(y|x; G) \) implied by DP mixture model (6).

The first step involves MCMC sampling from the posterior of model (7) with \( G \) marginalized over its DP prior. We update each \( (\mu, \Sigma) \) using algorithm 5 from Neal (2000), which is based on Metropolis-Hastings steps with the prior full conditional of \( (\mu, \Sigma) \) as the proposal. Updating all the \( (\mu_i, \Sigma_i) \), \( i = 1, ..., n \), generates a partition with \( n^* \leq n \) distinct components \( (\mu^*_j, \Sigma^*_j) \), \( j = 1, ..., n^* \). The vector of the \( (\mu, \Sigma) \) is determined through configuration indicators \( w = (w_1, ..., w_n) \) such that \( w_i = j \) if and only if \( (\mu_i, \Sigma_i) = (\mu^*_j, \Sigma^*_j) \). Evidently, the Metropolis-Hastings approach to update the \( (\mu, \Sigma) \) can lead to poor mixing. However, it is straightforward to implement and, combined with the standard trick from Bush and MacEachern (1996) to resample the \( (\mu^*_j, \Sigma^*_j) \), yields an efficient MCMC method. For each \( j = 1, ..., n^* \), the posterior full conditional for \( (\mu^*_j, \Sigma^*_j) \) is proportional to \( g_0(\mu^*_j, \Sigma^*_j; \psi) \prod_{i:w_i=j} N_{L+1}(z_i; \mu^*_j, \Sigma^*_j) \), where \( g_0 \) denotes the density of \( G_0 \). We sample from this full conditional by drawing from the full conditionals for \( \mu^*_j \) and \( \Sigma^*_j \). The former is \((L+1)\)-variate normal with mean vector \( (V^{-1} + n_j \Sigma_j^{-1})^{-1}(V^{-1}m + n_j \Sigma_j^{-1} \bar{z}_j) \) and covariance matrix \( (V^{-1} + n_j \Sigma_j^{-1})^{-1} \), where \( n_j = |\{i : w_i = j\}| \) and \( \bar{z}_j = n_j^{-1} \sum_{i:w_i=j} z_i \). The latter is inverse Wishart with scalar parameter \( \nu + n_j \) and matrix parameter \( S + \sum_{i:w_i=j}(z_i - \mu^*_j)(z_i - \mu^*_j)^T \).

Regarding the DP hyperparameters, we update \( \alpha \) using the auxiliary variable method from Escobar and West (1995). The posterior full conditional for \( m \) is \((L+1)\)-variate normal with mean vector \( (B_m^{-1} + n^*V^{-1})^{-1}(B_m^{-1}a_m + n^*V^{-1} \bar{m}^*) \), with \( \bar{m}^* = n^*\sum_{j=1}^{n^*} \mu^*_j \), and covariance matrix \( (B_m^{-1} + n^*V^{-1})^{-1} \). The full conditional for \( V \) is inverse Wishart with scalar parameter \( \nu + n^* \) and matrix parameter \( B_V + \sum_{j=1}^{n^*}(\mu^*_j - m)(\mu^*_j - m)^T \). Finally, the full conditional for \( S \) is given by a Wishart distribution with scalar parameter \( \alpha + n^* \) and matrix parameter \( B_S^{-1} + \sum_{j=1}^{n^*} \Sigma_j^{-1} \).

Next, note that the full posterior corresponding to model (7) is given by

\[
 p(G, \theta, \alpha, \psi|\text{data}) = p(G|\theta, \alpha, \psi)p(\theta, \alpha, \psi|\text{data})
\]

where \( \theta = \{ (\mu, \Sigma) : i = 1, ..., n \} \), or equivalently, \( \theta = \{ (\mu^*_j, \Sigma^*_j) : j = 1, ..., n^* \}, w \) (Antoniak, 1974). Here, the distribution for \( G|\theta, \alpha, \psi \) corresponds to a DP with updated precision parameter \( \alpha + n \) and mean \( G_0(\cdot; \theta, \alpha, \psi) \), which is a mixed distribution with point masses \( n_1(\alpha + n)^{-1} \) at the \( (\mu^*_j, \Sigma^*_j) \), \( j = 1, ..., n^* \), and continuous mass \( \alpha(\alpha + n)^{-1} \) on \( G_0(\psi) \).

Hence, after we obtain posterior samples from \( p(\theta, \alpha, \psi|\text{data}) \), using the MCMC algorithm described above, we can draw from (8) by augmenting each posterior sample with a draw from \( p(G|\theta, \alpha, \psi) \). The latter requires simulation from the DP with parameters given above, which we implement using the DP constructive definition (discussed in Section 2.1) with a truncation approximation (Gelfand and Kottas, 2002; Kottas, 2006). Therefore, this approach yields samples \( \{G_b, \theta_b, \alpha_b, \psi_b : b = 1, ..., B\} \) from the full posterior (8). Each posterior realization \( G_b \) is a discrete distribution with point masses at \( \vartheta_{rb} = (\mu_{rb}, \Sigma_{rb}) \), \( r = 1, ..., R_b \), drawn i.i.d. from \( G_0(\cdot; \theta, \alpha, \psi) \), and corresponding weights \( \omega_{rb}, r = 1, ..., R_b - 1 \), generated using the stick-breaking construction based on i.i.d. Beta(1, \( \alpha_b \)) draws (we set \( \omega_{R_b, b} = 1 - \sum_{r=1}^{R_b-1} \omega_{rb} \)). Here, \( R_b \) is the number of terms used in the truncation series approximation to the countable series representation for the DP. In general, \( R_b \) depends on the particular posterior realization, and the approximation can be specified up to any desired accuracy (see Kottas, 2006, for a specific rule to choose \( R_b \)).
Now, for any specific combination of response and covariate values, say, \((y_0, x_0)\),
\[
   f(y_0, x_0; G_b) = \int \sum_{r=1}^{R_b} \omega_{rb} N_{L+1}(y_0, x_0; \mu_{rb}, \Sigma_{rb})
\]
is a realization from the posterior of the random mixture density \(f(y, x; G)\), corresponding to the DP mixture model (6), at point \((y, x) = (y_0, x_0)\). Analogously, we can compute \(f(x_0; G_b)\) a draw from the posterior of the marginal \(f(x; G)\) at point \(x = x_0\), and thus, also, \(f(y_0 \mid x_0; G_b) = \int f(y_0, x_0; G_b) / f(x_0; G_b)\), i.e., a realization from the posterior of the conditional density \(f(y \mid x; G)\), at point \((y, x) = (y_0, x_0)\). Repeating over a grid in \(y\), that covers the range of response values of interest, we obtain a posterior realization from the random conditional density function \(f(\cdot \mid x_0; G)\) for the specific covariate values \(x_0\). Note that this is a posterior realization for the entire function, obtained, of course, up to the grid approximation. Now, for any \(0 < p < 1\), the conditional quantile \(q_p(x_0) \equiv q_p(x_0; G)\) satisfies \(\int q_p(x_0) f(y \mid x_0; G) dy = p\). Hence, using numerical integration (with interpolation) of the posterior realizations from the conditional density \(f(\cdot \mid x_0; G)\), yields draws from the posterior of \(q_p(x_0)\) for any set of percentiles that might be of interest (i.e., for any set of \(p \in (0, 1)\)).

Therefore, for any specified \(x_0\), and for any \(0 < p < 1\), we obtain samples from \(p(q_p(x_0) \mid \text{data})\) that can be used to summarize the information from these conditional quantiles in any desired form. In particular, for any set of \(p\) values, working with a grid in \(x\), i.e., over the covariate space, we can compute point and interval estimates for the corresponding quantile curves \(q_p(\cdot; G)\). Evidently, graphical depiction of these estimates for the entire curve is not feasible for problems with more than two covariates. For such applications, one can focus on illustrations involving the quantile regression function given subsets of the covariate vector including specific choices of one or two covariates.

Note that this approach to inference for quantile regression allows both non-linear quantile curves as well as non-standard shapes for the conditional distribution of the response given the covariates. As illustrated in Section 3.3, the latter can be explored through the posterior of \(f(\cdot \mid x_0; G)\) for any set of covariate values \(x_0\) of interest. Moreover, the model does not rely on the additive nonparametric regression formulation and therefore can uncover interactions between covariates that might influence certain quantile regression curves. Finally, an important feature of the approach is that it enables simultaneous inference for any set of quantile regressions that might be of interest in a particular application.
3.3 Data example

We illustrate the methodology of Section 3.2 with the motorcycle data (see, e.g., Silverman, 1985), which consist of 133 measurements of velocity in time for the helmet of a motorcycle crash victim after impact. (The data set is available from the MASS package for R.) Note that model (6) incorporates uncertainty in the covariates, which is arguably not required in this application. However, this example is still useful for illustrative purposes, since the motorcycle data is used as one of the standard tests for nonparametric regression approaches, including quantile regression estimation methods (see, e.g., Koenker, 2005, chapter 7).

We used a rough center and range for the response and covariate values to specify the hyperparameters for $G_0$. In particular, we choose the prior mean for $m$ using the center from the data. Setting $\nu = a_V = a_S = 6$, and working with diagonal matrices for $B_m$, $B_V$, $B_S$, we specify their diagonal elements using a rough range for acceleration and time values. Posterior inference was robust to less (as well as more) informative choices, e.g., involving range values larger (and smaller) than the data ranges. Finally, an exponential prior with mean 5 was placed on $\alpha$.

Figure 5 plots point (posterior mean) and 90% interval estimates for three quantile curves, specifically, for the median regression, 25-th quantile regression, and 75-th quantile regression functions. All posterior estimates capture very well the shape suggested by the data, in particular, adapting to its heteroscedasticity, the key feature of the motorcycle data. This is also reflected in posterior inference for $f(\cdot | x; G)$ for different values $x$ of time. For instance, Figure 6 shows posterior mean estimates and 90% posterior interval estimates for this conditional density at three values of $x$.

4. Discussion

We have developed two different model-based, fully inferential approaches for quantile regression. First, we have discussed a practically important extension of the semiparametric model from Kottas and Krnjajić (2005), combining Gaussian process priors for the regression functions and a Dirichlet process mixture prior for the errors in an additive quantile regression framework. Under the second approach, the joint distribution of the response and the (continuous) covariates is assigned a general Dirichlet process mixture prior model with inference for quantile regressions emerging from the induced conditional posterior distribution of the response given the covariates. We have presented MCMC posterior simulation methods under both approaches. Finally, we have provided illustrations for both models using either simulated or real data examples.

The MCMC algorithms for both models can be readily extended to handle censored response observations that arise, for instance, in survival analysis applications. For such applications, more structured versions for the kernel of the nonparametric mixture model of Section 3 can...
be utilized. For example, the kernel can be built from a Weibull or gamma distribution for the response and, independently, a multivariate normal distribution for the covariates.

An alternative method for posterior simulation from the model of Section 3.1 involves direct approximation of $G$ in model (7), using the constructive definition of its DP$(\alpha, G_0)$ prior, and then application of an MCMC technique for the induced discrete mixture model (see, e.g., Ishwaran and James, 2001). Results from comparison of this method with the approach of Section 3.2 will be reported elsewhere.

Regarding the approach of Section 2, we note that there has been relatively limited work in the Bayesian nonparametrics literature on combination of nonparametric priors for different sets of functions within the same model. Further study, involving either analytical work or empirical work through simulation, is needed for this approach regarding the extent of identifiability issues.

Finally, implementation of inference under both models becomes computationally intensive for moderate number of covariates, and seems prohibitive for large number of covariates with the currently used MCMC algorithms. It will therefore be of interest to study alternative to MCMC inferential methods (e.g., variational approximation techniques for Dirichlet process mixtures) and/or more efficient MCMC posterior simulation methods.

References


