Bayesian Nonparametric Modeling in Quantile Regression

Athanasios Kottas and Milovan Krnjajić *

Abstract

We propose Bayesian nonparametric methodology for quantile regression modeling. In particular, we develop Dirichlet process mixture models for the error distribution in an additive quantile regression formulation. The proposed nonparametric prior probability models allow the data to drive the shape of the error density and thus provide more reliable predictive inference than models based on parametric error distributions. We consider extensions to quantile regression for data sets that include censored observations. Moreover, we employ dependent Dirichlet processes to develop quantile regression models which allow the error distribution to change nonparametrically with the covariates. Posterior inference is implemented using Markov chain Monte Carlo methods. We assess and compare the performance of our models using both simulated and real data sets.

KEY WORDS: Censoring; Dependent Dirichlet processes; Dirichlet process mixture models; Markov chain Monte Carlo; Median regression; Skewness.

1 Introduction

A set of quantiles provides a more complete description of a distribution than the mean, which typically yields an inadequate summary. In the regression context, this observation motivates quantile regression, which can be used to quantify the relationship between a set of quantiles of the response distribution and available covariates. In many regression examples, e.g., in econometrics, educational and social studies, and medicine, we might expect a different structural relationship for the higher (or lower) responses than the *average* responses. In such cases, mean, or median, regression approaches would likely overlook important features that could be uncovered by a more general quantile regression analysis.

*A. Kottas is Assistant Professor and M. Krnjajić is Ph.D. candidate, Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95064, USA. The authors are grateful to David Dunson and Jack Taylor for providing the comet assay data discussed in Section 4.2.2. They also thank Keming Yu for providing the data set analyzed in Section 3.2.

Employing the standard additive regression formulation, the p-th quantile regression model for response observations y_i, with associated covariate vectors $\boldsymbol{x}_i$, $i = 1, ..., n$, can be written as

$$y_i = h(\boldsymbol{x}_i) + \epsilon_i, \tag{1}$$

where the ϵ_i are assumed independent from an error distribution with p-th quantile equal to 0, i.e., $\int_{-\infty}^{0} f_p(\epsilon)\mathrm{d}\epsilon = p$, with $f_p(\cdot)$ denoting the error density. Our objective is to develop flexible nonparametric prior models for the random error density $f_p(\cdot)$. We model $h(\cdot)$ parametrically and, for clarity of exposition, write $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of regression coefficients. A non-linear quantile regression function can also be accommodated modifying appropriately the methods for posterior simulation. Moreover, nonparametric modeling for $h(\cdot)$ could be proposed in addition to our modeling for $f_p(\cdot)$. We briefly discuss this latter extension in Section 5.

There is a fairly extensive literature on classical estimation for model (1); see, e.g., the review papers by Buchinsky (1998) and Yu, Lu and Stander (2003). This literature is dominated by *semiparametric* techniques where $h(\boldsymbol{x})$ is typically expressed as $\boldsymbol{x}^T\boldsymbol{\beta}$, and the error density $f_p(\cdot)$ is left unspecified (apart from the restriction $\int_{-\infty}^{0} f_p(\epsilon)\mathrm{d}\epsilon = p$). Hence, with no likelihood specification for the response distribution, point estimation for $\boldsymbol{\beta}$ proceeds by optimization of some *loss* function. For instance, under the standard setting with independent and uncensored responses, the point estimates for $\boldsymbol{\beta}$ minimize $\sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, where $\rho_p(u) = up - u1_{(-\infty,0)}(u)$, and, in fact, this reduces to the least absolute deviations criterion for $p = 0.5$, i.e., for the median regression case. Any inference beyond point estimation is based on asymptotic arguments or resampling methods and thus relies on the availability of large samples.

A Bayesian modeling approach to this problem enables exact and full inference, given the data, not only for the quantile regression coefficients but also for any functional of the response distribution that may be of interest. As such, it may be an appealing alternative to classical fitting techniques. Although the special case of median regression has been considered in the Bayesian nonparametric literature (see, e.g., Walker and Mallick, 1999; Kottas and Gelfand, 2001; Hanson and Johnson, 2002), little work exists for general quantile regression modeling. See, e.g., Yu and Moyeed (2001) for a parametric approach based on the asymmetric Laplace distribution for the errors, and Dunson, Watson and Taylor (2003) and Dunson and Taylor (2004) for an approximate method based on the substitution likelihood for quantiles. Moreover, the recent work of Hjort and Petrone (2005) studies nonparametric inference for the quantile function, including discussion of the extension to quantile regression.

Here we develop three families of nonparametric error distributions based on Dirichlet process mixture models (Ferguson, 1973; Antoniak, 1974). The first, a scale mixture of asymmetric Laplace densities, extends the parametric work of Yu and Moyeed (2001). Motivated by lim-

itations of this model, we propose two flexible scale mixtures of uniform densities, which can capture the shape (e.g., skewness, tail behavior) of *any* unimodal error density $f_p(\cdot)$. We discuss approaches for prior specification and posterior simulation based on Markov chain Monte Carlo (MCMC) methods. We show how the models can be fitted when some of the observations are censored. Moreover, building on recent work on dependent Dirichlet processes (MacEachern, 2000; De Iorio et al., 2004; Gelfand, Kottas and MacEachern, 2004), we develop quantile regression models which allow the error distribution to change nonparametrically with the covariates.

The plan of the paper is as follows. Section 2 presents the nonparametric mixture models for the quantile regression error density, including methods for prior choice, posterior inference for uncensored and censored data, and model comparison. Section 3 provides examples based on simulated and real data. Section 4 develops modeling for quantile regression with error densities that depend on the covariates, including data illustrations. Section 5 offers a summary and discussion of possible extensions. The Appendix includes details on the MCMC methods for posterior inference.

2 The Models

Mixture models for the error distribution are developed in Section 2.1. Sections 2.2 and 2.3 discuss posterior inference (with more details given in the Appendix) and prior specification, respectively. We consider the extension to censored quantile regression in Section 2.4. Model comparison is addressed in Section 2.5.

2.1 Mixture modeling for the error distribution

2.1.1 Nonparametric scale mixture of asymmetric Laplace densities

A natural starting point in constructing a nonparametric model for the random error density in (1) is to extend a parametric class of distributions through appropriate mixing. To our knowledge, the only parametric family that has been used in this context is the family of asymmetric Laplace distributions with densities

$$k_p^{AL}(\epsilon;\sigma) = \frac{p(1-p)}{\sigma}\exp\left\{-\frac{|\epsilon| + (2p-1)\epsilon}{2\sigma}\right\}, \tag{2}$$

where $0 < p < 1$, $\sigma > 0$ is a scale parameter and $\int_{-\infty}^{0} k_p^{AL}(\epsilon;\sigma)\mathrm{d}\epsilon = p$. Yu and Moyeed (2001) used (2) for the errors in (1), working with $h(\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ and, in fact, with $\sigma = 1$. Note that parameter p determines both skewness and p-th quantile for the density in (2), hence limiting its flexibility in modeling skewness and tail behavior. In particular, $k_p^{AL}(\cdot;\sigma)$ is skewed

for $p \neq 0.5$ and symmetric for $p = 0.5$, i.e., for the median regression case. This is a rather restrictive feature as median regression is typically motivated by the need to capture skewness in the response distribution. We refer to (1) with error density $f_p(\cdot) = k_p^{AL}(\cdot;\sigma)$ as model $\mathcal{M}_0$.

In order to construct a model with more flexible tail behavior, a general scale mixture of asymmetric Laplace densities can be used. We consider a nonparametric such mixture with a Dirichlet process (DP) prior for the mixing distribution, which is supported on R^+. Specifically, denoting by $\text{DP}(\alpha G_0)$ the DP with precision parameter α and base distribution G_0, we define

$$f_p^1(\epsilon; G) = \int k_p^{AL}(\epsilon;\sigma)\text{d}G(\sigma), \qquad G \sim \text{DP}(\alpha G_0). \tag{3}$$

Note that mixing in this fashion preserves the quantiles, i.e., $\int_{-\infty}^0 f_p^1(\epsilon; G)\text{d}\epsilon = p$. We place a gamma prior on α and take an inverse gamma distribution for G_0 with mean $d/(c-1)$, provided $c > 1$. We set $c = 2$, which yields an infinite variance for G_0, and work with a gamma prior for d. Introducing a latent mixing parameter σ_i associated with response observation y_i, the model can be expressed in the hierarchical form

$$\begin{array}{rcl} Y_i \mid \sigma_i & \overset{ind.}{\sim} & k_p^{AL}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta};\sigma_i), \;\; i = 1, ..., n \\ \sigma_i \mid G & \overset{iid}{\sim} & G, \;\; i = 1, ..., n \\ G \mid \alpha, d & \sim & \text{DP}(\alpha G_0) \end{array} \tag{4}$$

with independent normal priors for the components of $\boldsymbol{\beta}$. We refer to (3), or (4), as model $\mathcal{M}_1$.

Mixture model $\mathcal{M}_1$ extends model $\mathcal{M}_0$ with regard to tail behavior in the error distribution. However, scale mixing does not affect the skewness of the kernel of the mixture; $f_p^1(\cdot; G)$ has the same limitation as $k_p^{AL}(\cdot;\sigma)$ regarding skewness.

2.1.2 Nonparametric scale mixtures of uniform densities

The key result for constructing more flexible models than $\mathcal{M}_0$ and $\mathcal{M}_1$ is a *representation* for non-increasing densities on the positive real line. Specifically, for any non-increasing density $f(\cdot)$ on R^+ there exists a distribution function G, with support on R^+, such that $f(t) \equiv f(t; G) = \int \theta^{-1} 1_{[0,\theta)}(t)\text{d}G(\theta)$, i.e., $f(\cdot)$ can be expressed as a scale mixture of uniform densities. The result requires a general mixing distribution G and thus, for Bayesian modeling, invites the use of a nonparametric prior for G; see, e.g., Brunner and Lo (1989), Brunner (1992; 1995), and Kottas and Gelfand (2001), for applications of this representation, which utilize DP priors.

In our context, the result can be employed to provide a mixture representation for any unimodal density on the real line with p-th quantile (and mode) equal to zero, $\iint k_p(\epsilon;\sigma_1,\sigma_2)\text{d}G_1(\sigma_1)\text{d}G_2(\sigma_2)$. Here G_1 and G_2 are general mixing distributions, supported

on R^+, and

$$k_p(\epsilon;\sigma_1,\sigma_2) = \frac{p}{\sigma_1}1_{(-\sigma_1,0)}(\epsilon) + \frac{(1-p)}{\sigma_2}1_{[0,\sigma_2)}(\epsilon), \tag{5}$$

with $0 < p < 1$, and $\sigma_r > 0$, $r = 1, 2$. Assuming independent DP priors for G_1 and G_2, we obtain the model

$$f_p^2(\epsilon; G_1, G_2) = \iint k_p(\epsilon;\sigma_1,\sigma_2)\mathrm{d}G_1(\sigma_1)\mathrm{d}G_2(\sigma_2), \quad G_r \sim \mathrm{DP}(\alpha_r G_{r0}), r = 1, 2 \tag{6}$$

for the error density in (1). In the special case of median regression (i.e., $p = 0.5$), (6) reduces to the nonparametric error model studied in Kottas and Gelfand (2001). In the context of quantile regression, $f_p^2(\cdot; G_1, G_2)$ is sufficiently flexible to capture general forms of skewness and tail behavior. We use gamma priors for α_r, $r = 1, 2$, and inverse gamma distributions for G_{r0} with random means d_r, $r = 1, 2$, which are assigned gamma priors (again, we set the shape parameters c_r for G_{r0} equal to 2). With latent mixing parameters σ_{1i} and σ_{2i} for each response observation y_i, we now obtain the hierarchical model

$$\begin{array}{rcl} Y_i \mid \boldsymbol{\beta}, \sigma_{1i}, \sigma_{2i} & \overset{ind}{\sim} & k_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}; \sigma_{1i}, \sigma_{2i}), \;\; i = 1, ..., n \\ \sigma_{ri} \mid G_r & \overset{iid}{\sim} & G_r, \;\; r = 1, 2, \;\; i = 1, ..., n \\ G_r \mid \alpha_r, d_r & \sim & \mathrm{DP}(\alpha_r G_{r0}), \;\; r = 1, 2, \end{array} \tag{7}$$

again, with independent normal priors for the regression coefficients. Model (6), or (7), will be referred to as model $\mathcal{M}_2$.

The formulation in (6) indicates an alternative nonparametric family of error densities based on a single mixing distribution G, supported on $R^+ \times R^+$ and assigned a DP prior $\mathrm{DP}(\alpha G_0^*)$, where now G_0^* is a parametric distribution on $R^+ \times R^+$. Hence the new model (model $\mathcal{M}_3$) for the random error density is given by

$$f_p^3(\epsilon; G) = \iint k_p(\epsilon;\sigma_1,\sigma_2)\mathrm{d}G(\sigma_1,\sigma_2), \quad G \sim \mathrm{DP}(\alpha G_0^*). \tag{8}$$

Straightforwardly, $\int_{-\infty}^{0} f_p^3(\epsilon; G)\mathrm{d}\epsilon = p$. The hierarchical formulation for model $\mathcal{M}_3$ is analogous to (7), the difference being that now the pair of latent mixing parameters $(\sigma_{1i}, \sigma_{2i})$, given G, are i.i.d. from G, $i = 1, ..., n$. The precision parameter α is again given a gamma prior. We work with a bivariate lognormal distribution for G_0^* with density

$$g_0^*(\sigma_1,\sigma_2) = (2\pi\tau_1\tau_2\sqrt{1-\psi^2})^{-1}\sigma_1^{-1}\sigma_2^{-1}\exp\{-0.5(1-\psi^2)^{-1}(u_1^2 - 2\psi u_1 u_2 + u_2^2)\}, \tag{9}$$

where $u_r = (\log\sigma_r - \mu_r)/\tau_r$, $r = 1, 2$. We fix the location parameters μ_r and take inverse gamma priors for the scale paramaters τ_r^2, and a uniform prior for the dependence parameter

$\psi \in (-1, 1)$. Comparing mixture formulations (6) and (8), model $\mathcal{M}_3$ is expected to be, at least, as flexible as model $\mathcal{M}_2$; for instance, (6) can be viewed as a special case of (8) arising when G has independent marginals G_1 and G_2. However, allowing distinct mixing distributions as in (6), might be preferable for error densities with substantially different structure in their left and right tails. Models $\mathcal{M}_2$ and $\mathcal{M}_3$ are compared in Sections 3.1 and 3.2 on the basis of their predictive performance.

2.2 Posterior inference

We obtain inference under the models discussed in Section 2.1 utilizing well-established posterior simulation methods for DP mixture models. In particular, we use a combination of MCMC methods from Escobar and West (1995), Bush and MacEachern (1996), and Neal (2000). (Some of the details are given in the Appendix.) These methods are based on a marginalization of the random mixing distributions over their DP priors (Blackwell and MacQueen, 1973). Draws from the resulting marginalized posteriors yield the posterior predictive distribution for a *new* response Y_{new} with corresponding covariate vector $\boldsymbol{x}_{\text{new}}$.

We illustrate with model $\mathcal{M}_2$. Denote data $= \{y_i, \boldsymbol{x}_i : i = 1, ..., n\}$, and let $\boldsymbol{\psi}$ collect all model parameters, $\boldsymbol{\psi} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}_r = \{\sigma_{ri} : i = 1, ..., n\}, \alpha_r, d_r : r = 1, 2\}$. The discreteness of the DP priors (Blackwell, 1973; Sethuraman, 1994) induces a clustering in the $\boldsymbol{\sigma}_r$, $r = 1, 2$. For $r = 1, 2$, let n_r^* be the number of distinct elements of the vector $\boldsymbol{\sigma}_r$, and let σ_{rj}^*, $j = 1, ..., n_r^*$, be the distinct σ_{ri}, i.e., the cluster locations. Because G_{r0} is continuous, the clusters are determined by a vector of configuration indicators $\boldsymbol{s}_r = (s_{r1}, ..., s_{rn})$ such that $s_{ri} = j$ if and only if $\sigma_{ri} = \sigma_{rj}^*$ for $i = 1, ..., n$. Moreover, denote by n_{rj} the size of the j-th cluster for $j = 1, ..., n_r^*$. Evidently, $(n_r^*, \boldsymbol{s}_r, \{\sigma_{rj}^* : j = 1, ..., n_r^*\})$ yields an equivalent representation for $\boldsymbol{\sigma}_r$. Now the posterior predictive density for Y_{new} can be expressed as

$$p(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}, \text{data}) = \int k_p(\epsilon_{\text{new}}; \sigma_{1,\text{new}}, \sigma_{2,\text{new}}) p(\sigma_{1,\text{new}} \mid \boldsymbol{\psi}) p(\sigma_{2,\text{new}} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid \text{data}), \tag{10}$$

where $\epsilon_{\text{new}} = y_{\text{new}} - \boldsymbol{x}_{\text{new}}^T \boldsymbol{\beta}$, and for $r = 1, 2$,

$$p(\sigma_{r,\text{new}} \mid \boldsymbol{\psi}) = \frac{\alpha_r}{\alpha_r + n} \mathrm{G}_{r0}(\sigma_{r,\text{new}} \mid d_r) + \frac{1}{\alpha_r + n} \sum_{j=1}^{n_r^*} n_{rj} \delta_{\sigma_{rj}^*}(\sigma_{r,\text{new}}), \tag{11}$$

with δ_a denoting a point mass at a. Note that, combining (10) and (11), we can also write $p(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}, \text{data}) = \int p(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}, \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid \text{data})$, where $p(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}, \boldsymbol{\psi})$ is given by

$$\begin{aligned} &\tfrac{\alpha_1}{\alpha_1+n} \textstyle\int \tfrac{p}{\sigma_1} 1_{(-\sigma_1, 0)}(\epsilon_{\text{new}}) \mathrm{d}G_{10}(\sigma_1 \mid d_1) + \tfrac{\alpha_2}{\alpha_2+n} \int \tfrac{(1-p)}{\sigma_2} 1_{[0, \sigma_2)}(\epsilon_{\text{new}}) \mathrm{d}G_{20}(\sigma_2 \mid d_2) \\ &+ \tfrac{1}{\alpha_1+n} \textstyle\sum_{j=1}^{n_1^*} n_{1j} \tfrac{p}{\sigma_{1j}^*} 1_{(-\sigma_{1j}^*, 0)}(\epsilon_{\text{new}}) \quad + \quad \tfrac{1}{\alpha_2+n} \sum_{j=1}^{n_2^*} n_{2j} \tfrac{(1-p)}{\sigma_{2j}^*} 1_{[0, \sigma_{2j}^*)}(\epsilon_{\text{new}}). \end{aligned} \tag{12}$$

The first two components in (12) allow for new structure with decreasing weights $\alpha_r/(\alpha_r + n)$, $r = 1, 2$, for increasing sample size. In fact, when the sample size is moderate to large (say $n > 50$), the posteriors for α_r, $r = 1, 2$, are typically supported by values small relative to n, whence, with $\alpha_r/(\alpha_r + n) \approx 0$, $p(y_{\text{new}} \mid \boldsymbol{x}_{\text{new}}, \boldsymbol{\psi})$ can be approximated by the last two terms in (12). This is a discrete mixture with number of components n_r^*, $r = 1, 2$, which are driven by the data through the posterior for $\boldsymbol{\psi}$, and, in particular, with different induced clustering structure for the left and right tails of the posterior predictive density.

2.3 Prior specification

In the absence of strong prior information about the quantile regression function, we take independent normal priors, with zero mean and large variance, for the regression coefficients. We use dispersed gamma priors for the precision parameters of the DPs, which control the number of clusters (e.g., the n_r^* under model $\mathcal{M}_2$); larger values increase the probabilities for larger number of clusters (Antoniak, 1974; Escobar and West, 1995). In general, posterior predictive inference is robust to the prior choice for the precision parameters. We work with prior predictive densities to specify the hyperparameters of the base distributions for the DP priors. In addition to revealing the effect of certain choices for the prior hyperparameter values, prior predictive densities facilitate empirical model comparison, because they provide the means of checking that roughly the same amount of prior information is used for all models under consideration. They are easily estimated for all models discussed in Section 2.1; e.g., under model $\mathcal{M}_2$, the prior predictive density at y_0, with corresponding covariate vector $\boldsymbol{x}_0$, is given by

$$p(y_0) = \int k_p(y_0 - \boldsymbol{x}_0^T\boldsymbol{\beta}; \sigma_{10}, \sigma_{20}) p(\boldsymbol{\beta}) \mathrm{d}G_{10}(\sigma_{10} \mid d_1) \mathrm{d}G_{20}(\sigma_{20} \mid d_2) p(d_1) p(d_2),$$

where $p(\boldsymbol{\beta})$ is the prior density for $\boldsymbol{\beta}$, and $p(d_r)$ are the prior densities for d_r, $r = 1, 2$.

2.4 Bayesian nonparametric censored quantile regression

Median, as well as general quantile, regression models for censored survival data have received attention in the classical literature (see, e.g., Yang, 1999, Koenker and Geling, 2001, and the review by Buchinsky, 1998, for earlier references). More recently, there has been some Bayesian work on censored median regression (e.g., Walker and Mallick, 1999; Kottas and Gelfand, 2001; Hanson and Johnson, 2002), and median residual life regression (Gelfand and Kottas, 2003).

All the quantile regression models of Section 2.1 can be extended to handle right, left, or interval censored observations. The extension requires modifications of the posterior simulation techniques. For instance, in the presence of right censoring, let $n = n_o + n_c$, where n_o of the

survival times t_{i_o}, $i_o = 1, ..., n_o$, are observed, whereas for the remaining n_c survival times t_{i_c}, $i_c = 1, ..., n_c$, we have $t_{i_c} > z_{i_c}$ for known censorship times z_{i_c}. Then, the only change required for models $\mathcal{M}_2$ and $\mathcal{M}_3$ is in the first stage of the corresponding hierarchical models to incorporate the right censored observations. For instance, with y_{i_o} and y_{i_c} denoting, on a logarithmic scale, the observed survival times t_{i_o} and the right censorship times z_{i_c}, respectively, and with $\boldsymbol{x}_{i_o}$ and $\boldsymbol{x}_{i_c}$ denoting the corresponding covariate vectors, the first stage in (7) for model $\mathcal{M}_2$ becomes

$$\prod_{i_o=1}^{n_o} k_p(y_{i_o} - \boldsymbol{x}_{i_o}^T\boldsymbol{\beta}; \sigma_{1,i_o}, \sigma_{2,i_o}) \prod_{i_c=1}^{n_c} (1 - K_p(y_{i_c} - \boldsymbol{x}_{i_c}^T\boldsymbol{\beta}; \sigma_{1,i_c}, \sigma_{2,i_c}))$$

where $K_p(\cdot; \sigma_1, \sigma_2)$ denotes the distribution function of $k_p(\cdot; \sigma_1, \sigma_2)$. The Gibbs samplers that we use to fit models $\mathcal{M}_2$ and $\mathcal{M}_3$ under censoring have the same structure with the case of fully observed data (detailed in the Appendix). The difference is that now the full conditionals for the latent mixing parameters σ_{1,i_c}, σ_{2,i_c}, associated with right censored observations, are derived using the survival function $1 - K_p(\cdot; \sigma_{1,i_c}, \sigma_{2,i_c})$ instead of the density function. We note that Kottas and Gelfand (2001) attempted to fit model $\mathcal{M}_2$, with $p = 0.5$, to the right censored data of Section 3.3 using data augmentation, but were not able to report reliable posterior results. Under the data augmentation sampling scheme, the addition of latent variables to impute the censored observations, resulted in a Gibbs sampler with poor mixing. The algorithm we use here overcomes the difficulties with data augmentation. Section 3.3 illustrates inference for censored quantile regression under model $\mathcal{M}_2$, demonstrating its superiority over parametric alternatives.

2.5 Model comparison

Given the different semiparametric specifications discussed in section 2.1 and additional parametric models that might be considered, the need for model comparison arises. Here, we explore model choice in posterior predictive space working with both empirical graphical comparisons and formal posterior predictive criteria.

In particular, in the examples of section 3 we compare posterior predictive densities, posterior predictive survival functions, and posteriors for specific quantiles. For the survival data of section 3.3, which include right censored observations, we also illustrate with conditional predictive ordinate (CPO) plots (see, e.g., Ibrahim, Chen and Sinha, 2001). For model $\mathcal{M}$, and specified covariate vector $\boldsymbol{x}$, denote by $p^{\mathcal{M}}(\cdot|\boldsymbol{x}, \text{data})$ and $S^{\mathcal{M}}(\cdot|\boldsymbol{x}, \text{data})$ the posterior predictive density and survival function, respectively, on the original scale. Then the CPO for an observed survival time t_{i_o} is given by $p^{\mathcal{M}}(t_{i_o}|\boldsymbol{x}_{i_o}, \text{data}(-i_o))$, whereas the CPO for a right censored survival time t_{i_c} is defined as $S^{\mathcal{M}}(z_{i_c}|\boldsymbol{x}_{i_c}, \text{data}(-i_c))$, where data$(-i_o)$ and data$(-i_c)$ denote the data vector excluding t_{i_o} and z_{i_c}, respectively. A large CPO value indicates agreement between the associated

observation and the model. Models can be compared using a plot of all CPO values. In addition, the CPOs can be summarized yielding the cross-validation posterior predictive criterion

$$\mathrm{Q}(\mathcal{M}) = n^{-1} \sum_{i_o=1}^{n_o} \log p^{\mathcal{M}}(t_{i_o}|\boldsymbol{x}_{i_o}, \mathrm{data}(-i_o)) + n^{-1} \sum_{i_c=1}^{n_c} \log S^{\mathcal{M}}(z_{i_c}|\boldsymbol{x}_{i_c}, \mathrm{data}(-i_c)) \qquad (13)$$

(see, e.g., Bernardo and Smith, 2000). For the data set of section 3.2, we work with a criterion based on a posterior predictive loss approach suggested in Gelfand and Ghosh (1998). The criterion favors the model $\mathcal{M}$ which minimizes

$$\mathrm{D}_m(\mathcal{M}) = \sum_{i=1}^{n} V^{\mathcal{M}}(i) + \frac{m}{m+1} \sum_{i=1}^{n} (y_i - E^{\mathcal{M}}(i))^2, \qquad (14)$$

where $m \geq 0$, and $E^{\mathcal{M}}(i)$ and $V^{\mathcal{M}}(i)$ is the mean and variance, respectively, under model $\mathcal{M}$, of the posterior predictive distribution for $Y_{\mathrm{new},i}$ with associated covariate vector $\boldsymbol{x}_i$. The first component in (14) is a penalty term for model complexity whereas the second component is a goodness-of-fit term, with weight determined by the value of m.

3 Data Illustrations

We illustrate the methodology using real and synthetic data sets. For all examples, we followed the approach in Section 2.3 for prior specification working with prior predictive densities. We have also conducted prior sensitivity analysis. The posteriors for the quantile regression coefficients and the DP hyperparameters were robust for a wide range of prior choices.

3.1 Simulation study

To assess the performance of the error models discussed in section 2.1, we ignore covariates and generate data from distributions with a specific quantile fixed at zero, and with varying shapes. In particular, we work with standard Laplace distributions ($\sigma = 1$ in (2)) for three values of p ($p = 0.5$, 0.9, and 0.1), a standard normal distribution, and two mixtures of normals, one with 0.6-th quantile at zero and another with median zero. The components for both normal mixtures are chosen so that the resulting mixture densities are right skewed with non-standard tail behavior. The true densities under the six cases of this simulation experiment are included in Figure 1. All the samples were of size $n = 250$.

Both models $\mathcal{M}_2$ and $\mathcal{M}_3$ capture very successfully the different distributional shapes as illustrated for $\mathcal{M}_2$ in Figure 1. (Posterior predictive densities under models $\mathcal{M}_2$ and $\mathcal{M}_3$ were almost indistinguishable.) Model $\mathcal{M}_1$ fits well the data generated from the asymmetric Laplace

distributions but fails for all other data sets as depicted in Figure 2 for the data drawn from the normal mixture distribution with 0.6-th quantile at zero.

3.2 Immunoglobulin-G data set

Here we work with data discussed in Royston and Altman (1994), and used also by Yu and Moyeed (2001) to illustrate quantile regression analysis based on an asymmetric Laplace error distribution with fixed scale parameter equal to 1 (thus a special case of model $\mathcal{M}_0$). The data set consists of values of serum concentrations (gram/liter) of immunoglobulin-G (IgG) for 298 children, with ages from 6 months to 6 years. As in Yu and Moyeed (2001), we use a quadratic quantile regression model $\beta_0 + \beta_1 x + \beta_2 x^2$, where x denotes age in years.

We have used the predictive loss criterion $\mathrm{D}_m(\mathcal{M})$ in (14) to compare models $\mathcal{M}_0$ through $\mathcal{M}_3$. Table 1 provides results for $m = 1$ and $m \to \infty$, and for 5 quantiles, $p = 0.05$, 0.25, 0.5, 0.75, and 0.95. Based on this criterion, model $\mathcal{M}_2$ outperforms models $\mathcal{M}_0$ and $\mathcal{M}_1$ for all 5 quantiles. Results are similar for model $\mathcal{M}_3$, the difference being that model $\mathcal{M}_1$ is favored in the median regression case over model $\mathcal{M}_3$. We note that model $\mathcal{M}_0$ and, to a lesser extent, model $\mathcal{M}_1$, performs substantially worse than models $\mathcal{M}_2$ and $\mathcal{M}_3$ at the low and high quantile values. This could be attributed to the restrictive feature of models $\mathcal{M}_0$ and $\mathcal{M}_1$ discussed in section 2.1.1, i.e., the fact that the skewness of the error density is determined once p is specified.

The posterior predictive error densities for $p = 0.5$ (i.e., for the median regression case), under all four models, are given in Figure 3. By their definition, models $\mathcal{M}_0$ and $\mathcal{M}_1$ have symmetric error densities in the median regression case. However, the results based on models $\mathcal{M}_2$ and $\mathcal{M}_3$ indicate that the error density is skewed. To illustrate how the quantiles of the IgG serum concentration distribution change with age, Figure 4 shows the posteriors of $\beta_0 + \beta_1 x + \beta_2 x^2$, under model $\mathcal{M}_2$, at six values for age and for four quantiles.

3.3 Small cell lung cancer data

To illustrate the methodology for censored quantile regression, we consider a data set analyzed using median regression models in Ying, Jung and Wei (1995), Yang (1999), Walker and Mallick (1999), and Kottas and Gelfand (2001). It consists of survival times in days for 121 patients with small cell lung cancer; 23 survival times are right censored. Each patient was randomly assigned to one of two treatments A and B, achieving 62 and 59 patients, respectively. To facilitate graphical comparisons between the two treatments, we work with the treatment indicator as the single covariate. (Also available is the patient's age at entry in the clinical study.)

We fit model $\mathcal{M}_2$ to this data set using a $\log_{10}$ transformation of the survival times. Figure 5

provides posterior predictive densities and survival functions under both treatments. It also compares the posteriors for 0.25-th quantile, median, 0.75-th quantile, and 0.90-th quantile survival times for the two treatments. All the results indicate that treatment A is better. Noteworthy are the non-standard shapes for the predictive densities and the bimodalities in the posteriors for some of the quantile survival times. These are features that standard parametric models are unable to uncover; see, e.g., the top panel of Figure 6.

Results under model $\mathcal{M}_3$ (not shown here) were similar with model $\mathcal{M}_2$. In the interest of comparison of model $\mathcal{M}_2$ with simpler parametric alternatives, we fit model $\mathcal{M}_0$ and a Weibull proportional hazards model to the data. Under the latter model, the survival function is $\exp(-t^\gamma \exp(\beta_0+\beta_1 x))$, where $\gamma > 0$ is the Weibull shape parameter and x is the treatment indicator, and the p-th quantile survival time has the simple form $\{-\log(1-p)\exp(-(\beta_0+\beta_1 x))\}^{1/\gamma}$. The CPO plots (Figure 6) indicate a superior predictive performance of model $\mathcal{M}_2$ compared with the two parametric models, as does the the cross-validation criterion $Q(\mathcal{M})$ in (13), taking values -8.01, -6.91, and -11.56 for models $\mathcal{M}_0$, $\mathcal{M}_2$, and the Weibull model, respectively.

4 Dependent Nonparametric Error Distributions for Quantile Regression

4.1 The modeling approach

Here we propose an extension of the standard modeling framework in (1) to a class of quantile regression models where the error density $f_p(\cdot)$ depends on the covariates. For a simpler exposition, we consider a single continuous covariate x with realized values x_m, $m = 1, ..., M$. For any specified quantile p, the error distribution under (1) is the same for all values of x and hence the response distribution changes with x only through the p-th quantile $\beta_0 + \beta_1 x$. Extension to nonparametric covariate-dependent error distributions requires a nonparametric prior model for the stochastic process of error densities indexed by values x in the covariate space $\mathcal{X}$, i.e., for $f_{p,\mathcal{X}} = \{f_{p,x}(\cdot) : x \in \mathcal{X}\}$, where for each fixed x, $\int_{-\infty}^{0} f_{p,x}(\varepsilon)\mathrm{d}\varepsilon = p$. Hence, in this setting, $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ are dependent for all $x \neq x'$. In fact, we would typically seek a specification that yields *similar* $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ for x close to x'. We employ dependent Dirichlet processes (DDPs) to formulate a prior probability model for $f_{p,\mathcal{X}}$. The DDP was developed by MacEachern (1999; 2000) as a nonparametric prior for a stochastic process of random distributions. These distributions are dependent but such that, at each index value, the distribution is a DP. We also refer to De Iorio et al. (2004) for an illustration in the ANOVA setting, and Gelfand, Kottas and MacEachern (2004) for related work on spatial DPs.

We provide the details building on model $\mathcal{M}_2$. A similar approach could be used for model $\mathcal{M}_3$. First, we re-parameterize the kernel (5) of mixture model (6) so that $\sigma_r = \exp(\theta_r)$, where $\theta_r \in R$, $r = 1, 2$. Hence (6) becomes $f_p^2(\epsilon; G_1, G_2) = \iint k_p(\epsilon; \theta_1, \theta_2) \mathrm{d}G_1(\theta_1) \mathrm{d}G_2(\theta_2)$, $G_r \sim \mathrm{DP}(\alpha_r G_{r0})$, $r = 1, 2$, where now we could take $G_{r0} = \mathrm{N}(\mu_r, \tau_r^2)$, $r = 1, 2$, with μ_r and/or τ_r^2 random.

To allow $f_p^2(\epsilon; G_1, G_2)$ to change with x, we need mixing distributions G_1 and G_2 that change with x and are still assigned nonparametric priors; we need prior probability models for the stochastic processes $\{G_r(x) : x \in \mathcal{X}\}$, where $G_r(x)$, $r = 1, 2$, are the mixing distributions for covariate value x. The definition of the DP given by Sethuraman (1994) provides a constructive approach to defining such priors. Based on this definition, a realization G_r, $r = 1, 2$, from $\mathrm{DP}(\alpha_r G_{r0})$ is almost surely of the form $G_r = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \delta_{\psi_{r,\ell}}$ where $\psi_{r,\ell}$ are i.i.d. from G_{r0} and the weights arise from a *stick-breaking* procedure, $\omega_{r,1} = z_{r,1}$, $\omega_{r,\ell} = z_{r,\ell} \prod_{s=1}^{\ell-1}(1 - z_{r,s})$, $\ell = 2, 3, ...$, with $z_{r,s}$ i.i.d. Beta$(1, \alpha_r)$. Moreover, the sequences of random variables $\{\psi_{r,\ell} : \ell = 1, 2, ...\}$ and $\{z_{r,s} : s = 1, 2, ...\}$ are independent.

Hence, an extension of the DP (a prior model for the distribution function G_r) to a DDP (a prior model for the stochastic process $\{G_r(x) : x \in \mathcal{X}\}$) arises by replacing the univariate R-valued random variable $\psi_{r,\ell}$ with a realization from a stochastic process over $\mathcal{X}$, $\psi_{r,\ell,\mathcal{X}} = \{\psi_{r,\ell}(x) : x \in \mathcal{X}\}$. Therefore we are replacing the base distribution function G_{r0}, with support on R, with a base stochastic process $G_{r0,\mathcal{X}}$ over $\mathcal{X}$ taking values in R. The resulting random distribution $G_{r,\mathcal{X}}$ for $\{G_r(x) : x \in \mathcal{X}\}$ has the representation

$$G_{r,\mathcal{X}} = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \delta_{\psi_{r,\ell,\mathcal{X}}} \tag{15}$$

where $\psi_{r,\ell,\mathcal{X}}$ are i.i.d. realizations from $G_{r0,\mathcal{X}}$, i.e., $G_{r,\mathcal{X}}$ arises as a countable mixture of realizations from the base stochastic process $G_{r0,\mathcal{X}}$. Extending earlier notation, we write $G_{r,\mathcal{X}} \sim \mathrm{DDP}(\alpha_r G_{r0,\mathcal{X}})$ to denote that $G_{r,\mathcal{X}}$ follows the DDP prior, and $\theta_{r,\mathcal{X}} = \{\theta_r(x) : x \in \mathcal{X}\} \mid G_{r,\mathcal{X}} \sim G_{r,\mathcal{X}}$ to indicate that $\theta_{r,\mathcal{X}}$ given $G_{r,\mathcal{X}}$ is a realization from $G_{r,\mathcal{X}}$.

An important consequence of the construction leading to (15) is that for any finite set of covariate values the induced prior is a DP. Specifically, for any collection of x values, $\boldsymbol{u} = (x_1, ..., x_L)$, which can include both observed and new covariate values, we have $G_{r,\boldsymbol{u}} = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \delta_{\psi_{r,\ell}(\boldsymbol{u})}$, where the L-dimensional random vectors $\psi_{r,\ell}(\boldsymbol{u}) = (\psi_{r,\ell}(x_1), ..., \psi_{r,\ell}(x_L))$ are i.i.d. with distribution $G_{r0}(\boldsymbol{u})$ induced by the stochastic process $G_{r0,\mathcal{X}}$ at $\boldsymbol{u}$. Therefore, if $\theta_{r,\mathcal{X}} \mid G_{r,\mathcal{X}} \sim G_{r,\mathcal{X}}$, then $G_{r,\mathcal{X}}$ induces at $\boldsymbol{u}$ a $\mathrm{DP}(\alpha_r G_{r0}(\boldsymbol{u}))$ prior on the space of distribution functions for $(\theta_r(x_1), ..., \theta_r(x_L))$. Note that, although the $\psi_{r,\ell}(\boldsymbol{u})$ are i.i.d. from $G_{r0}(\boldsymbol{u})$, for any ℓ, $\psi_{r,\ell}(x_1)$,...,$\psi_{r,\ell}(x_L)$ are dependent. Hence a critical advantage of the DDP model, besides the fact that it enables different error density shapes for different observed covariate values, is that

it can provide posterior predictive inference for the error distribution at unobserved x values allowing learning from nearby covariate values.

A natural choice for $G_{r0,\mathcal{X}}$, $r = 1, 2$, is a Gaussian process, which is taken to be stationary with constant mean, $\mathrm{E}(\theta_r(x) \mid \mu_r) = \mu_r$, and variance, $\mathrm{Var}(\theta_r(x) \mid \tau_r^2) = \tau_r^2$, and exponential correlation function $\mathrm{Corr}(\theta_r(x), \theta_r(x') \mid \phi_r) = \exp(-\phi_r|x - x'|)$, for x, $x' \in \mathcal{X}$, with random hyperparameters μ_r, τ_r^2, and $\phi_r > 0$, $r = 1, 2$. Therefore, for the observed covariate vector $\boldsymbol{x} = (x_1, ..., x_M)$, $G_{r0}(\boldsymbol{x})$ is an M-dimensional normal with mean vector $\mu_r 1_M$ and covariance matrix V_r with (i, j)-th element $\tau_r^2 \exp(-\phi_r|x_i - x_j|)$, $i, j = 1, ..., M$. Deviations from the Gaussianity and stationarity structure imposed in the center $G_{r0,\mathcal{X}}$ of the DDP prior emerge through the countable mixing in (15). In particular, for any x, $x' \in \mathcal{X}$, $\mathrm{E}(\theta_r(x) \mid G_{r,\mathcal{X}}) = \sum_{\ell=1}^{\infty} \omega_{r,\ell}\psi_{r,\ell}(x)$, and $\mathrm{Cov}(\theta_r(x), \theta_r(x') \mid G_{r,\mathcal{X}}) = \sum_{\ell=1}^{\infty} \omega_{r,\ell}\psi_{r,\ell}(x)\psi_{r,\ell}(x') - (\sum_{\ell=1}^{\infty} \omega_{r,\ell}\psi_{r,\ell}(x))\,(\sum_{\ell=1}^{\infty} \omega_{r,\ell}\psi_{r,\ell}(x'))$.

Introducing mixing through independent DDP priors $G_{1,\mathcal{X}}$ and $G_{2,\mathcal{X}}$ yields a prior for the collection $f_{p,\mathcal{X}}$ of quantile regression error densities. In particular, for any x, we obtain model $\mathcal{M}_2$ as the induced DP mixture model,

$$f^2_{p,x}(\epsilon; G_{1,x}, G_{2,x}) = \iint k_p(\epsilon; \theta_1(x), \theta_2(x))\mathrm{d}G_{1,x}(\theta_1(x))\mathrm{d}G_{2,x}(\theta_2(x)),$$

with $G_{r,x} \sim \mathrm{DP}(\alpha_r G_{r0}(x))$, $r = 1, 2$, and $G_{r0}(x) = \mathrm{N}(\mu_r, \tau_r^2)$. However, now the random error densities are dependent with the extent of dependence driven by $G_{1,\mathcal{X}}$ and $G_{2,\mathcal{X}}$. More generally, for the vector $\boldsymbol{x}$, we can write

$$f^2_{p,\boldsymbol{x}}(\boldsymbol{\epsilon}; G_{1,\boldsymbol{x}}, G_{2,\boldsymbol{x}}) = \iint \prod_{m=1}^{M} k_p(\epsilon_m; \theta_1(x_m), \theta_2(x_m))\mathrm{d}G_{1,\boldsymbol{x}}(\boldsymbol{\theta}_1)\mathrm{d}G_{2,\boldsymbol{x}}(\boldsymbol{\theta}_2),$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_M)$, $\boldsymbol{\theta}_r = (\theta_r(x_1), ..., \theta_r(x_M))$, and $G_{r,\boldsymbol{x}} \sim \mathrm{DP}(\alpha_r G_{r0}(\boldsymbol{x}))$, $r = 1, 2$, with $G_{r0}(\boldsymbol{x})$ the M-variate normal distribution described above. We note that, in practice, learning with DDP priors is facilitated by some form of replication in the response values, i.e, more than one response value for each x_m, $m = 1, ..., M$. Let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iM})$, $i = 1, ..., N$, be the i-th response replicate. In the examples of Section 4.2 we have *complete* replicates, i.e., the same number of response observations at each covariate value. Using customary data augmentation methods, the model can also be fitted when some of the y_{im} are missing.

To express the model in a hierarchical form, let $\boldsymbol{\theta}_{ri} = (\theta_{ri}(x_1), ..., \theta_{ri}(x_M))$, $r = 1, 2$, be the latent mixing vectors associated with $\boldsymbol{y}_i$. Moreover, let $f_p(\boldsymbol{y}_i; \boldsymbol{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}) = \prod_{m=1}^{M} k_p(y_{im} - (\beta_0 + \beta_1 x_m); \theta_{1i}(x_m), \theta_{2i}(x_m))$. Then the quantile regression model is given by

$$\begin{array}{rcl} \boldsymbol{Y}_i \mid (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i} & \overset{ind.}{\sim} & f_p(\boldsymbol{y}_i; \boldsymbol{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}), \;\; i = 1, ..., N \\ \boldsymbol{\theta}_{ri} \mid G_{r,\boldsymbol{x}} & \overset{i.i.d.}{\sim} & G_{r,\boldsymbol{x}}, \;\; r = 1, 2, \;\; i = 1, ..., N \\ G_{r,\boldsymbol{x}} \mid \alpha_r, \mu_r, \tau_r^2, \phi_r & \sim & \mathrm{DP}(\alpha_r G_{r0}(\boldsymbol{x}) = \mathrm{N}_M(\mu_r 1_M, V_r)), \;\; r = 1, 2, \end{array} \tag{16}$$

with independent priors for all the hyperparameters. In particular, we take normal priors for β_0, β_1, and μ_r, gamma priors for α_r, inverse gamma priors for τ_r^2, and a uniform prior on $(0, b_\phi)$ for ϕ_1 and ϕ_2. Prior specification for all parameters, other than ϕ_1 and ϕ_2, proceeds along the lines discussed in Section 2.3. To specify b_ϕ, we recall that, under the exponential correlation function, ϕ_r determines the range, $(3/\phi_r)$, of the base Gaussian process $G_{r0,\mathcal{X}}$. (For an isotropic covariance function that decreases to 0 as distance goes to ∞, the range is the distance at which correlation becomes 0.05.) The range is usually presumed to be around one half of the maximum interpoint distance over the index space. But since $3/b_\phi < 3/\phi_r$, we conservatively specify $3/b_\phi = a(\max x_m - \min x_m)$ for a small value of a.

Model (16) is a DP mixture model with the M-variate DP priors for $G_{r,\boldsymbol{x}}$ induced by the DDP priors for $G_{r,\mathcal{X}}$. Hence, again, posterior sampling proceeds by marginalizing $G_{r,\boldsymbol{x}}$ over their DP priors, and utilizing an MCMC method for DP mixtures. (The Appendix includes some of the details.) Regarding posterior predictive inference, interest lies in the posterior predictive density at observed covariate values in $\boldsymbol{x}$ as well as at new (unobserved) values, say $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_U)$. Hence we seek $p(\boldsymbol{y}_0, \tilde{\boldsymbol{y}}_0 \mid \tilde{\boldsymbol{x}}, \text{data})$, where $\boldsymbol{y}_0 = (y_{01}, ..., y_{0M})$, and $\tilde{\boldsymbol{y}}_0 = (\tilde{y}_{01}, ..., \tilde{y}_{0U})$ are associated with $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$, respectively, and data $= \{x_1, ..., x_M; \boldsymbol{y}_1, ..., \boldsymbol{y}_N\}$. Consider, for $r = 1, 2$, the augmented set of latent mixing vectors $\{(\boldsymbol{\theta}_{ri}, \tilde{\boldsymbol{\theta}}_{ri}) : i = 1, ..., N\}$, where $\tilde{\boldsymbol{\theta}}_{ri} = (\theta_{ri}(\tilde{x}_1), ..., \theta_{ri}(\tilde{x}_U))$, corresponding to the augmented response data vector $\{(\boldsymbol{y}_i, \tilde{\boldsymbol{y}}_i) : i = 1, ..., N\}$, which includes the unobserved portion $\tilde{\boldsymbol{y}}_i = (\tilde{y}_{i1}, ..., \tilde{y}_{iU})$ associated with $\tilde{\boldsymbol{x}}$. For $r = 1, 2$, let N_r^* be the number of clusters in the $\{(\boldsymbol{\theta}_{ri}, \tilde{\boldsymbol{\theta}}_{ri}) : i = 1, ..., N\}$, denote by $(\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*)$, $j = 1, ..., N_r^*$, the cluster locations, and let N_{rj} be the j-th cluster size. Again, the clusters are determined by a vector of configuration indicators $\boldsymbol{s}_r = (s_{r1}, ..., s_{rN})$, with $s_{ri} = j$ if and only if $(\boldsymbol{\theta}_{ri}, \tilde{\boldsymbol{\theta}}_{ri}) = (\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*)$, $i = 1, ..., N$. Let $\boldsymbol{\eta}$ collect all parameters corresponding to the augmented data vector, i.e., $\boldsymbol{\eta} = \{\beta_0, \beta_1, N_r^*, \boldsymbol{s}_r, \{(\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*) : j = 1, ..., N_r^*\}, \alpha_r, \mu_r, \tau_r^2, \phi_r : r = 1, 2\}$. Note that $\boldsymbol{\eta} = \{\boldsymbol{\xi}, \{\tilde{\boldsymbol{\theta}}_{rj}^* : j = 1, ..., N_r^*\} : r = 1, 2\}$, where $\boldsymbol{\xi}$ is the parameter vector for model (16), resulting after marginalizing $G_{r,\boldsymbol{x}}$, $r = 1, 2$. Thus the MCMC algorithm discussed in the Appendix yields the posterior for all parameters in $\boldsymbol{\eta}$ other than the $\tilde{\boldsymbol{\theta}}_{rj}^*$. As described below, to obtain posterior draws from the $\tilde{\boldsymbol{\theta}}_{rj}^*$, the additional required sampling is from U-variate normal distributions.

The key observation is that, for $r = 1, 2$, the $(\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*)$ are i.i.d. from $G_{r0}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$, which is a $(M + U)$-variate normal distribution with mean vector $\mu_r 1_{M+U}$ and covariance matrix T_r with elements that depend on values in both $\boldsymbol{x}$ and $\tilde{\boldsymbol{x}}$. In particular, the $M \times M$ upper diagonal block of T_r is given by V_r, the $U \times U$ lower diagonal block consists of $\tau_r^2 \exp(-\phi_r |\tilde{x}_u - \tilde{x}_{u'}|)$, $u, u' = 1, ..., U$, and the $M \times U$ cross-diagonal block has elements $\tau_r^2 \exp(-\phi_r |x_m - \tilde{x}_u|)$, $m = 1, ..., M$, $u = 1, ..., U$. Based on this DP property, we can express the posterior $p(\boldsymbol{\eta} \mid \tilde{\boldsymbol{x}}, \text{data})$

as $p(\boldsymbol{\xi} \mid \text{data}) \prod_{r=1}^{2} \prod_{j=1}^{N_r^*} p(\tilde{\boldsymbol{\theta}}_{rj}^* \mid \boldsymbol{\theta}_{rj}^*, \mu_r, \tau_r^2, \phi_r)$, where $p(\tilde{\boldsymbol{\theta}}_{rj}^* \mid \boldsymbol{\theta}_{rj}^*, \mu_r, \tau_r^2, \phi_r)$ is the conditional U-variate normal density for $\tilde{\boldsymbol{\theta}}_{rj}^*$ given $\boldsymbol{\theta}_{rj}^*$ under the $G_{r0}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ distribution for $(\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*)$. Now, the posterior predictive density is given by

$$\begin{aligned} p(\boldsymbol{y}_0, \tilde{\boldsymbol{y}}_0 \mid \tilde{\boldsymbol{x}}, \text{data}) &= \int f_p(\boldsymbol{y}_0; \boldsymbol{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}) f_p(\tilde{\boldsymbol{y}}_0; \tilde{\boldsymbol{x}}, (\beta_0, \beta_1), \tilde{\boldsymbol{\theta}}_{10}, \tilde{\boldsymbol{\theta}}_{20}) \\ &\qquad p(\boldsymbol{\theta}_{10}, \tilde{\boldsymbol{\theta}}_{10} \mid \boldsymbol{\eta}) p(\boldsymbol{\theta}_{20}, \tilde{\boldsymbol{\theta}}_{20} \mid \boldsymbol{\eta}) p(\boldsymbol{\eta} \mid \tilde{\boldsymbol{x}}, \text{data}), \end{aligned}$$

where, for $r = 1, 2$,

$$p(\boldsymbol{\theta}_{r0}, \tilde{\boldsymbol{\theta}}_{r0} \mid \boldsymbol{\eta}) = \frac{\alpha_r}{\alpha_r + N} g_{r0}(\boldsymbol{\theta}_{r0}, \tilde{\boldsymbol{\theta}}_{r0} \mid \mu_r, \tau_r^2 \phi_r) + \frac{1}{\alpha_r + N} \sum_{j=1}^{N_r^*} N_{rj} \delta_{(\boldsymbol{\theta}_{rj}^*, \tilde{\boldsymbol{\theta}}_{rj}^*)}(\boldsymbol{\theta}_{r0}, \tilde{\boldsymbol{\theta}}_{r0}),$$

with $g_{r0}(\boldsymbol{\theta}_{r0}, \tilde{\boldsymbol{\theta}}_{r0} \mid \mu_r, \tau_r^2 \phi_r)$ denoting the density of the $G_{r0}(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ distribution.

4.2 Examples

We present two examples to illustrate the DDP quantile regression model. The first (Section 4.2.1) is based on a simulated data set; the second (Section 4.2.2) considers the comet assay data from Dunson and Taylor (2004).

4.2.1 Simulated data

We consider simulated data with $N = 100$ response values generated according to $y_{im} = \beta_0 + \beta_1 x_m + \varepsilon_{im}$, for each of $M = 5$ covariate values. In particular, we use the same covariate values with the real data of section 4.2.2, $x_m = 0, 5, 20, 50, 100$, and set $\beta_0 = 0$ and $\beta_1 = 0.45$. The errors ε_{im} were generated from a median-zero split normal distribution $0.5\text{TN}(\varepsilon \mid 0, \varphi\sigma^2; \varepsilon \in (-\infty, 0))$ + $0.5\text{TN}(\varepsilon \mid 0, \varphi^{-1}\sigma^2; \varepsilon \in [0, \infty))$, where $\text{TN}(\cdot \mid \mu, \tau^2; \cdot \in A)$ denotes a $\text{N}(\cdot \mid \mu, \tau^2)$ distribution truncated over set A. Here $\varphi > 0$ is a skewness parameter and σ is a scale parameter. To allow different error density shapes for different x values, we let φ and σ^2 be functions of x, specifically, $\varphi(x) = \exp(-(x - 50)/80)$ and $\sigma^2(x) = \exp(0.0012(x - 55)^2)$, which yields left and right skewed errors for $x \in [0, 50)$ and $x > 50$, respectively, and symmetric errors at $x = 50$.

We fit model (16), with $p = 0.5$, to the data. (We used a uniform prior on $(0, 1.5)$ for ϕ_1 and ϕ_2.) The DDP model captures successfully the different shapes of the error density at the five observed x values as illustrated in Figure 7. It also enables learning at new x values where responses are not observed (Figure 7 includes prediction at two such values, $x = 10$ and $x = 95$). The amount of learning depends on how close the unobserved x values are to the observed x_m. For instance, posterior predictive densities associated with values for x in $(50, 90)$ (not shown) were similar and had roughly symmetric shapes.

To assess the benefits of employing the extra level of complexity in the DDP model, we have also fitted model $\mathcal{M}_2$ to the data (again, for $p = 0.5$). The resulting posterior predictive densities (not shown) had the same shape (slightly left skewed) for all values of x, with range shifted according to the estimated median regression term. Hence, as anticipated, model (16) yields superior predictive performance. The fact that the DDP mixture model explains the error structure better than DP mixture model $\mathcal{M}_2$ is also reflected in posterior inference for the median regression coefficients. The posteriors for β_0 and β_1 are more dispersed under model $\mathcal{M}_2$ than model (16) and fail to capture the true values. In particular, under model $\mathcal{M}_2$, point estimates (posterior medians) and 95% central interval estimates for β_0, and β_1 are given by -0.425, $(-0.814, -0.107)$, and 0.457, $(0.453, 0.464)$, respectively. The corresponding posterior estimates under the DDP mixture model are -0.091, $(-0.182, -0.006)$, and 0.451, $(0.449, 0.453)$, respectively.

4.2.2 Data from a genotoxicity experiment

Here we work with the data set studied in Dunson and Taylor (2004). A brief description is provided below; see Dunson and Taylor (2004) for more details. The data were drawn from a genotoxicity experiment assessing the effect of oxidative damage on the frequency of DNA strand breaks. Samples of cells exposed to different levels of hydrogen peroxide ($d_m = 0$, 5, 20, 50, 100 uM H_2O_2) were prepared for use in the comet assay. After electrophoresis, the DNA from the nucleus of cells with a high frequency of DNA strand breaks exhibit a comet-style shape, with the nucleus forming the ball-like head and the cut DNA strands the tail. Cells with a low frequency tend to maintain an approximately spherical shape with less tail. The objective of the experiment was to evaluate the sensitivity of the comet assay in detecting genotoxic effects of hydrogen peroxide. Measurements on the % DNA in the comet tail are available for 100 cells in each of the five dose groups. As in Dunson and Taylor (2004), we take the response, y_{im}, for cell i ($i = 1, ..., N = 100$) in dose group m ($m = 1, ..., M = 5$), to be the % DNA in the comet tail divided by 100 (whence $0 \leq y_{im} \leq 1$). (Alternatively, a logit transformation could be used to ensure real-valued responses.) The p-th quantile regression function is assumed to be $\beta_0 + \beta_1 x_m$, where $x_m = \log(d_m + 1)$ and β_0, β_1 are the p-th quantile regression coefficients.

Dunson and Taylor (2004) demonstrated the utility of a quantile regression analysis for these data, obtaining approximate inference based on the substitution likelihood for quantiles. They concluded that the linear quantile regression model provides a good fit to the data, and found evidence of an increasing trend in the direction of larger slopes for higher quantiles.

The histograms of the response values at the five dose groups (included in Figure 8) suggest

that the different shapes for the response distribution will not be captured by the quantile regression term. Thus the methodology of Section 4.1 seems well suited for these data. We obtain results under the DDP model (16) for $p = 0.1$, 0.25, 0.5, 0.75, and 0.9. (Here, the prior for ϕ_1 and ϕ_2 was uniform on $(0, 6)$.) Regarding inference for the regression coefficients, of primary interest are the five slope parameters β_1, which indicate how the 0.1, 0.25, 0.5, 0.75, and 0.9 quantiles of the distribution of % DNA in the comet tail change with dose. Posterior medians and 95% central posterior interval estimates for β_1 are 0.0152 and $(0.0097, 0.0180)$ for $p = 0.1$; 0.0248 and $(0.0208, 0.0292)$ for $p = 0.25$; 0.0487 and $(0.0442, 0.0564)$ for $p = 0.5$; 0.0671 and $(0.0613, 0.0713)$ for $p = 0.75$; 0.0747 and $(0.0716, 0.0765)$ for $p = 0.9$. Hence we too find evidence that the frequency of DNA strand breaks increases with increasing dose of hydrogen peroxide, and that this increasing trend is stronger at the higher quantiles. The posterior densities for the β_1 under the DDP model (not shown) are more concentrated than the corresponding quasi-posterior densities reported in Dunson and Taylor (2004). Again, this can be attributed to the flexibility of the DDP mixture model (16), i.e., its ability to model accurately the error structure with different distributional shapes for different ranges of dose values. This is illustrated in Figure 8, which includes posterior predictive densities at the five observed and at three new dose values.

5 Summary and Future Work

We have developed modeling approaches for the error distribution in quantile regression using a representation for unimodal densities on the real line with a specified quantile equal to zero. We have studied two formulations through nonparametric scale mixtures of uniform densities with DP priors for the mixing distributions. We have demonstrated the superiority of these classes of nonparametric mixture models over parametric models as well as over a class of nonparametric scale mixtures of asymmetric Laplace densities, which extends existing parametric work for the quantile regression error distribution. We have discussed methods for prior specification, posterior inference based on MCMC techniques, and model comparison. We have shown how the MCMC posterior simulation methods can be extended to handle censored observations. We have also proposed methodology for quantile regression error densities that change with values in the covariate space, using DDP mixing for the scale mixture of uniform densities. Finally, we have provided illustrations for all the models with simulated and real data.

An extension of interest involves nonparametric modeling for the quantile regression function $h(\cdot)$ in (1). For instance, Gaussian process priors provide a flexible framework for formulation of such modeling; see, e.g., Neal (1998) on Gaussian process regression under parametric error

distributions. The combination of model-based nonparametric approaches for $h(\cdot)$ and the error distribution $f_p(\cdot)$ in (1) can potentially yield a practically important class of models, which allow the data to uncover non-linearities in the quantile regression function as well as non-standard distributional features in the errors. Fitting of such fully nonparametric quantile regression models will require extensions of the current methodology as well as carefully designed MCMC algorithms. We will report on this work in a future article.

References

Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.

Bernardo, J.M., and Smith, A.F.M. (2000), *Bayesian Theory*, Chichester: Wiley.

Blackwell, D. (1973), "Discreteness of Ferguson Selections," *The Annals of Statistics*, 1, 356-358.

Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.

Brunner, L.J. (1992), "Bayesian nonparametric methods for data from a unimodal density," *Statistics and Probability Letters*, 14, 195-199.

Brunner, L.J. (1995), "Bayesian Linear Regression With Error Terms That Have Symmetric Unimodal Densities," *Journal of Nonparametric Statistics*, 4, 335-348.

Brunner, L.J., and Lo, A.Y. (1989), "Bayes methods for a symmetric unimodal density and its mode," *Annals of Statistics*, 17, 1550-1566.

Buchinsky, M. (1998), "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research," *The Journal of Human Resources*, 33, 88-126.

Bush, C.A., and MacEachern, S.N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275-285.

De Iorio, M., Müller, P., Rosner, G.L., and MacEachern, S.N. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205-215.

Dunson, D.B., and Taylor, J.A. (2004), "Approximate Bayesian Inference for Quantiles," ISDS Discussion Paper 2004-17, Duke University. (To appear in *Journal of Nonparametric Statistics*.)

Dunson, D.B., Watson, M., and Taylor, J.A. (2003), "Bayesian Latent Variable Models for Median Regression on Multiple Outcomes," *Biometrics*, 59, 296-304.

Escobar, M.D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association,* 90, 577-588.

Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics,* 1, 209-230.

Gelfand, A.E., and Ghosh, S.K. (1998), "Model choice: A minimum posterior predictive loss approach," *Biometrika*, 85, 1-11.

Gelfand, A.E., and Kottas, A. (2003), "Bayesian Semiparametric Regression for Median Residual Life," *Scandinavian Journal of Statistics*, 30, 651-665.

Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2004), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," AMS Technical Report 2004-05, University of California, Santa Cruz. (To appear in *Journal of the American Statistical Association*.)

Hanson, T., and Johnson, W.O. (2002), "Modeling Regression Error with a Mixture of Pólya Trees," *Journal of the American Statistical Association*, 97, 1020-1033.

Hjort, N.L., and Petrone, S. (2005), "Nonparametric Quantile Inference Using Dirichlet Processes," unpublished manuscript.

Ibrahim, J.G., Chen, M-H., and Sinha, D. (2001), *Bayesian survival analysis*, New York: Springer.

Koenker, R., and Geling, O. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis," *Journal of the American Statistical Association*, 96, 458-468.

Kottas, A., and Gelfand, A.E. (2001), "Bayesian Semiparametric Median Regression Modeling," *Journal of the American Statistical Association*, 96, 1458-1468.

MacEachern, S.N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50-55.

MacEachern, S.N. (2000), "Dependent Dirichlet Processes," Technical Report, Department of Statistics, The Ohio State University.

Neal, R.M. (1998), "Regression and Classification using Gaussian Process Priors," in *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, eds. J.M. Bernardo, Berger, J.O., A.P. Dawid and A.F.M. Smith, Oxford University Press, pp. 475-501.

Neal, R.M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249-65.

Royston, P., Altman, D.G. (1994), "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling," (with discussion). *Applied Statistics*, 43, 429-467.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639-650.

Walker, S.G., and Mallick, B.K. (1999), "A Bayesian Semiparametric Accelerated Failure Time Model," *Biometrics*, 55, 477-483.

Yang, S. (1999), "Censored median regression using weighted empirical survival and hazard functions," *Journal of the American Statistical Association*, 94, 137-145.

Ying, Z., Jung, S.H., and Wei, L.J. (1995), "Survival analysis with median regression models," *Journal of the American Statistical Association*, 90, 178-184.

Yu, K., Moyeed, R.A. (2001), "Bayesian Quantile Regression," *Statistics and Probability Letters*, 54, 437-447.

Yu, K., Lu, Z., and Stander, J. (2003), "Quantile regression: applications and current research areas," *The Statistician*, 52, 331-350.

Appendix: MCMC Posterior Inference

We present here some of the details for MCMC posterior simulation for models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$, discussed in Section 2.1, and for the DDP quantile regression model developed in Section 4.1. For the clustering structure induced by the DP priors, we work with the notation used in Sections 2.2 and 4.1 for models $\mathcal{M}_2$ and the DDP mixture, respectively. We use the superscript "-" to denote all relevant quantities for the vector of latent mixing parameters with its ith element

removed. Moreover, IG(a, b) denotes the inverse gamma distribution with mean $b/(a-1)$, for $a > 1$. Finally, we use the standard bracket notation for marginal and conditional densities.

Model $\mathcal{M}_1$: Marginalizing G in (4) over its DP prior, the posterior becomes

$$[\sigma_1, ..., \sigma_n, \boldsymbol{\beta}, \alpha, d \mid \text{data}] \propto [\sigma_1, ..., \sigma_n \mid \alpha, d][\boldsymbol{\beta}][\alpha][d] \prod_{i=1}^{n} k_p^{AL}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_i),$$

where $[\sigma_1, ..., \sigma_n \mid \alpha, d]$ is the joint prior for the latent mixing parameters σ_i induced by the Pólya urn representation of the DP (Blackwell and MacQueen, 1973).

The full conditional for each σ_i is a mixed distribution with point masses at σ_j^{*-}, $j = 1, ..., n^{*-}$, and continuous component given by an IG$(c+1, d+\gamma_i)$ distribution, where $\gamma_i = (y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})\{p1_{(0,\infty)}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}) + (p-1)1_{(-\infty,0]}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})\}$. The weights for the point masses are proportional to $n_j^- k_p^{AL}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_j^{*-})$. The weight for the continuous component is proportional to $\alpha \int k_p^{AL}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma) \mathrm{d}G_0(\sigma) = \alpha c d^c p(1-p)(d+\gamma_i)^{-c-1}$.

To resample the cluster locations (Bush and MacEachern, 1996), once all the σ_i are updated, we note that the full conditional for each σ_j^* is proportional to $g_0(\sigma_j^*) \prod_{\{i:s_i=j\}} k_p^{AL}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_j^*)$, resulting in an IG$(c + n_j, d + \sum_{\{i:s_i=j\}} \gamma_i)$ distribution.

We sample $\boldsymbol{\beta}$ using a random-walk Metropolis step. We update α using the method of Escobar and West (1995). Finally, the full conditional for d is a gamma distribution.

Model $\mathcal{M}_2$: Integrating G_1 and G_2 in (7) over their DP priors, we obtain the induced posterior

$$[\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \boldsymbol{\beta}, \alpha_1, \alpha_2, d_1, d_2 \mid \text{data}] \propto [\boldsymbol{\beta}] \prod_{r=1}^{2} [\sigma_{r1}, ..., \sigma_{rn} \mid \alpha_r, d_r][\alpha_r][d_r] \prod_{i=1}^{n} k_p(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_{1i}, \sigma_{2i}),$$

where, again, the joint priors $[\sigma_{r1}, ..., \sigma_{rn} \mid \alpha_r, d_r]$, $r = 1, 2$, arise from the Pólya urn structure of the DP priors for G_r, $r = 1, 2$.

For each $i = 1, ..., n$, the full conditional for σ_{1i} has point masses at σ_{1j}^{*-}, $j = 1, ..., n_1^{*-}$, and a continuous component, which is an IG$(c_1 + 1, d_1)$ distribution truncated below by $-(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})$ if $y_i < \boldsymbol{x}_i{}^T\boldsymbol{\beta}$, and an IG$(c_1, d_1)$ distribution when $y_i \geq \boldsymbol{x}_i{}^T\boldsymbol{\beta}$. The weight for this continuous part is proportional to

$$p\alpha_1 c_1 d_1^{-1}\{1 - F_{IG(c_1+1,d_1)}(-(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}))\}1_{(-\infty,0)}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}) + (1-p)\alpha_1\sigma_{2i}^{-1}1_{[0,\sigma_{2i})}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}),$$

where $F_{IG(a,b)}(u)$ denotes the IG(a, b) distribution function at u. The weights for the point masses are proportional to $n_{1j}^- k_p(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_{1j}^{*-}, \sigma_{2i})$. Analogously, we obtain the full conditionals for σ_{2i}, $i = 1, ..., n$.

We next resample the cluster locations σ_{rj}^*, $j = 1, ..., n_r^*$, $r = 1, 2$. The full conditional for each σ_{1j}^* is proportional to $g_{10}(\sigma_{1j}^*) \prod_{\{i:s_{1i}=j\}} k_p(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}; \sigma_{1j}^*, \sigma_{2i})$, where g_{10} is the density associated

with G_{10}. Letting $\mathcal{A}_1 = \{i : s_{1i} = j,\ \ y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta} < 0\}$, and $L_1 = |\mathcal{A}_1|$, the full conditional is an $\text{IG}(c_1 + L_1, d_1)$ distribution truncated below by $\max_{i\in\mathcal{A}_1}(-(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}))$. Similarly, it can be shown that the full conditional for each σ^*_{2j} is an $\text{IG}(c_2 + L_2, d_2)$ distribution truncated below by $\max_{i\in\mathcal{A}_2}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})$, where $\mathcal{A}_2 = \{i : s_{2i} = j,\ \ y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta} \geq 0\}$, and $L_2 = |\mathcal{A}_2|$.

We update α_1 and α_2 using the approach of Escobar and West (1995). The gamma priors for d_1 and d_2 lead to gamma full conditionals for these parameters. The vector $\boldsymbol{\beta}$ can be sampled using a random-walk Metropolis step. Alternatively, Gibbs sampling over each of the elements of $\boldsymbol{\beta}$ can be used following an approach similar to the one in Kottas and Gelfand (2001).

Model $\mathcal{M}_3$: Here the marginalized posterior is proportional to

$$[\boldsymbol{\beta}][\alpha][\tau_1^2][\tau_2^2][\psi][(\sigma_{11},\sigma_{21}),...,(\sigma_{1n},\sigma_{2n}) \mid \alpha,\tau_1^2,\tau_2^2,\psi]\prod_{i=1}^{n} k_p(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta};\sigma_{1i},\sigma_{2i}).$$

We update the $(\sigma_{1i},\sigma_{2i})$, $i = 1,...,n$, using algorithm 5 from Neal (2000). The full conditional for the cluster $(\sigma^*_{1j},\sigma^*_{2j})$ is proportional to ${\sigma^*_{1j}}^{-L_1}{\sigma^*_{2j}}^{-L_2} g_0^*(\sigma^*_{1j},\sigma^*_{2j})1_{(\sigma^*_{1j}>w_1)}1_{(\sigma^*_{2j}>w_2)}$, $j = 1,...,n^*$. Here $L_1 = |A_1|$, $L_2 = |A_2|$, $A_1 = \{i : s_i = j,\ y_i < \boldsymbol{x}_i{}^T\boldsymbol{\beta}\}$, $A_2 = \{i : s_i = j,\ y_i \geq \boldsymbol{x}_i{}^T\boldsymbol{\beta}\}$, $w_1 = \max_{i\in A_1}(-(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}))$, $w_2 = \max_{i\in A_2}(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})$, and g_0^* is defined in (9). We extend the Gibbs sampler to draw from the two full conditionals for σ^*_{1j} and σ^*_{2j}, which are truncated lognormal distributions.

Again, α is sampled using the Escobar and West (1995) method. The Metropolis step for $\boldsymbol{\beta}$ is the same with model $\mathcal{M}_2$. We discretize the full conditional density for ψ over its support $(-1,1)$ and sample it directly. Finally, we update τ_1^2 and τ_2^2 using slice sampling.

DDP quantile regression model: After integrating out each of $G_{r,\boldsymbol{x}}$ over its $\text{DP}(\alpha_r G_{r0}(\boldsymbol{x}))$ prior, we can utilize Gibbs sampling to draw from the resulting marginalized posterior corresponding to model (16).

We use algorithm 6 from Neal (2000) to update the $\boldsymbol{\theta}_{ri}$, $i = 1,...,N$, $r = 1,2$. The regression coefficients (β_0,β_1) are sampled with a random-walk Metropolis update. As with the other models, the precision parameters α_r are sampled as in Escobar and West (1995). The full conditional distributions for the μ_r and the τ_r^2 are normal and inverse gamma, respectively. Finally, since the full conditionals for the ϕ_r are not of some standard form, we discretize them over their support $(0, b_\phi)$ and sample them as discrete distributions.

Table 1: IgG data. Values for the posterior predictive loss criterion $D_m(\mathcal{M})$, with $m = 1$ (D_1) and $m \to \infty$ (D_∞), for models $\mathcal{M} = \mathcal{M}_0$ through $\mathcal{M}_3$ under five quantiles, $p = 0.05$, 0.25, 0.5, 0.75, and 0.95. Here, P and G denote the penalty term, $\sum_{i=1}^{n} V^{\mathcal{M}}(i)$, and goodness-of-fit term, $\sum_{i=1}^{n}(y_i - E^{\mathcal{M}}(i))^2$, respectively.

	p	P	G	D_1	D_∞
$\mathcal{M}_0$	0.05	113802	74330	150968	188133
$\mathcal{M}_1$	0.05	22926	1515	23684	24442
$\mathcal{M}_2$	0.05	1514	1164	2097	2679
$\mathcal{M}_3$	0.05	3882	1188	4476	5071
$\mathcal{M}_0$	0.25	5185	1541	5956	6726
$\mathcal{M}_1$	0.25	2351	1181	2942	3533
$\mathcal{M}_2$	0.25	1361	1157	1940	2518
$\mathcal{M}_3$	0.25	1551	1169	2136	2721
$\mathcal{M}_0$	0.50	2866	1151	3442	4018
$\mathcal{M}_1$	0.50	1811	1147	2385	2959
$\mathcal{M}_2$	0.50	1555	1152	2132	2708
$\mathcal{M}_3$	0.50	2045	1161	2626	3206
$\mathcal{M}_0$	0.75	5592	1811	6498	7404
$\mathcal{M}_1$	0.75	2338	1261	2969	3599
$\mathcal{M}_2$	0.75	1898	1165	2481	3064
$\mathcal{M}_3$	0.75	1702	1165	2285	2868
$\mathcal{M}_0$	0.95	114240	71518	150000	185759
$\mathcal{M}_1$	0.95	58460	2373	59646	60833
$\mathcal{M}_2$	0.95	1220	1149	1795	2369
$\mathcal{M}_3$	0.95	1298	1158	1877	2456

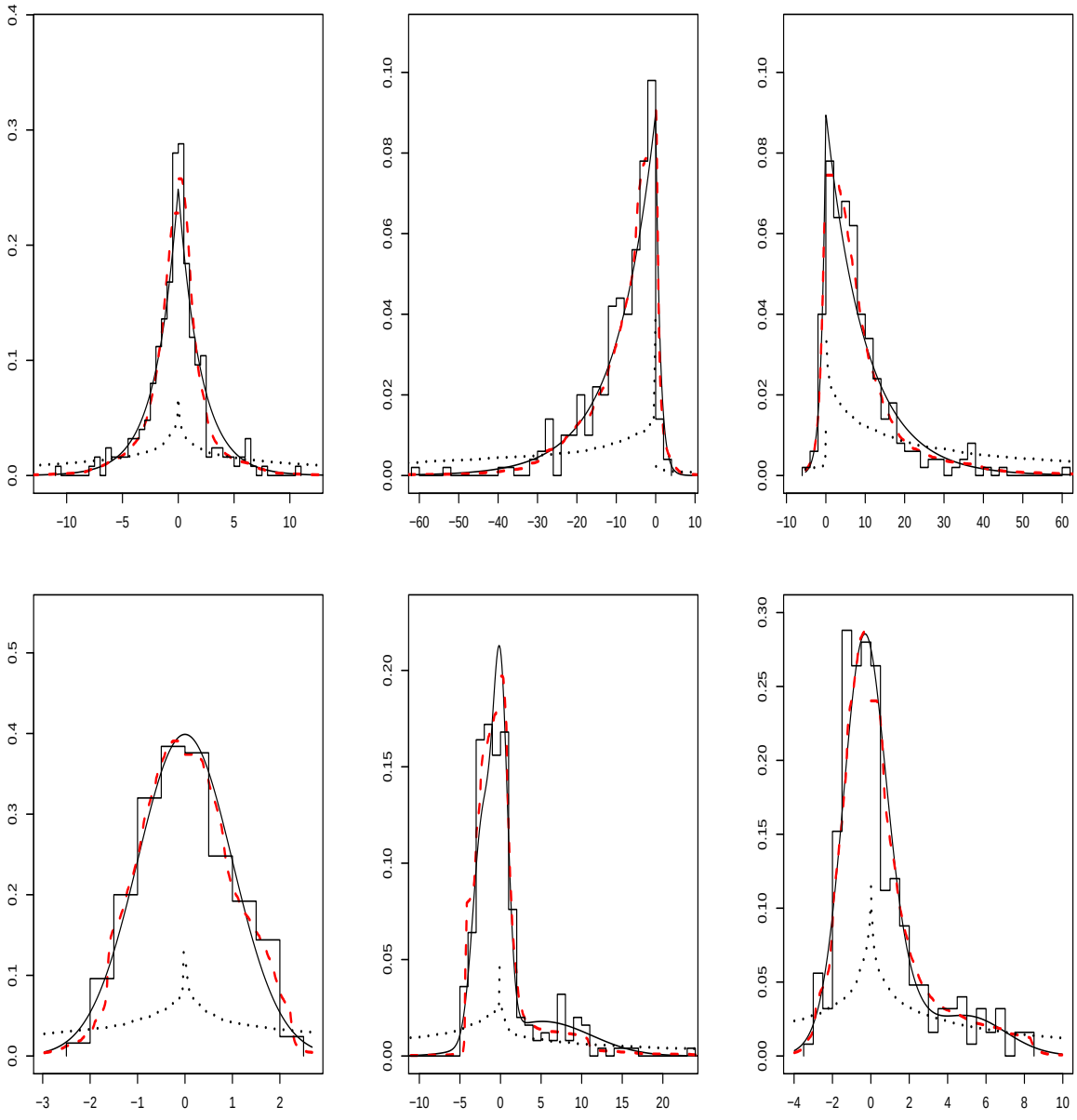

Figure 1: Simulation study. Prior predictive (dotted lines) and posterior predictive (dashed lines) densities under model $\mathcal{M}_2$. The top panels correspond to the standard Laplace densities, with $p = 0.5$, 0.9, and 0.1 for the left, middle, and right panel, respectively. The bottom panels include the standard normal density (left panel), and the two normal mixture densities, the first (middle panel) with 0.6-th quantile at zero, and the second (right panel) with median zero. The solid lines denote the true densities; the histograms of the simulated data are also included.

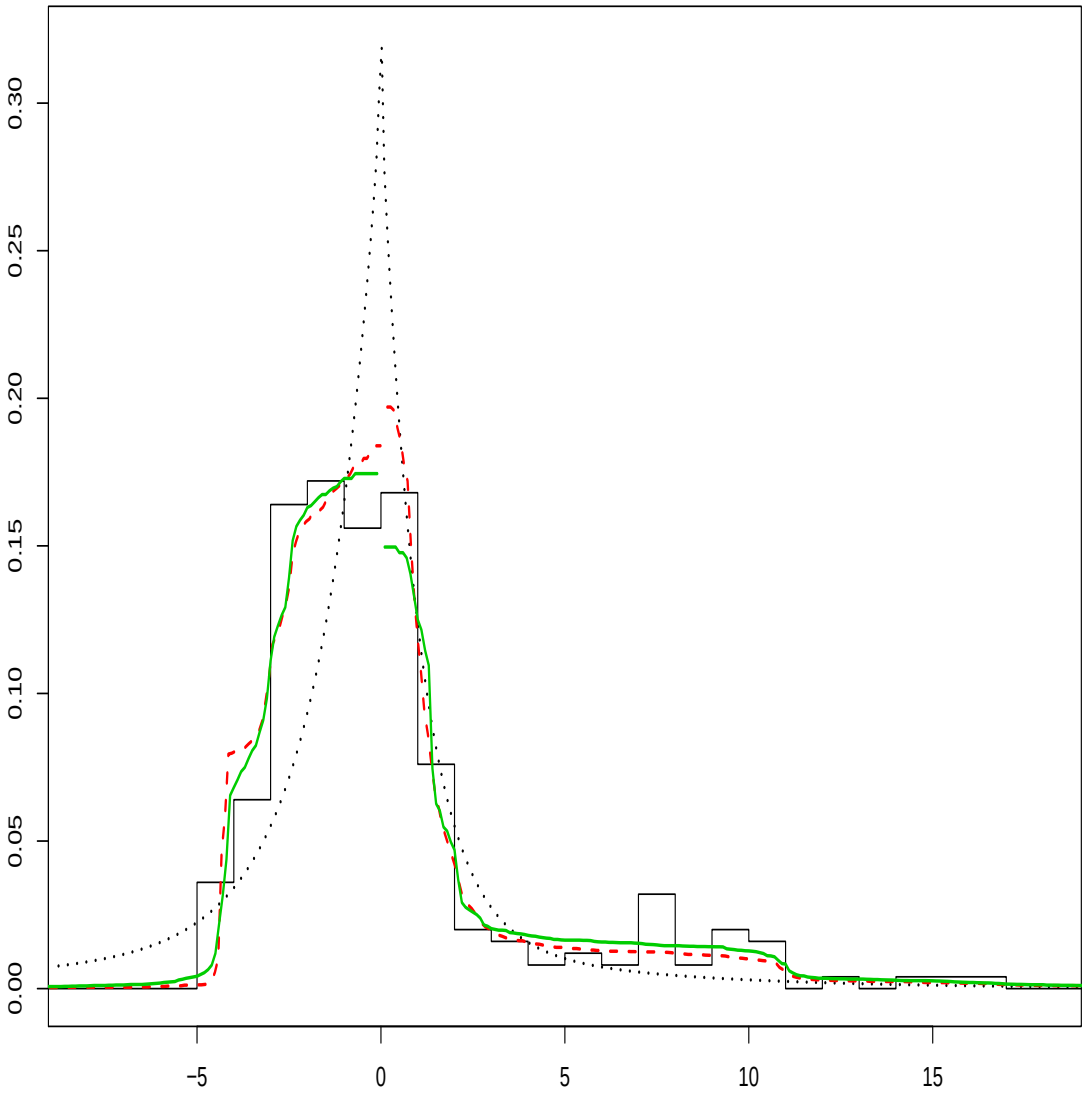

Figure 2: Simulation study. Posterior predictive densities under model $\mathcal{M}_1$ (dotted line), model $\mathcal{M}_2$ (dashed line), and model $\mathcal{M}_3$ (solid line) for the case of the right skewed normal mixture with 0.6-th quantile at zero. The histogram of the simulated data is also included.

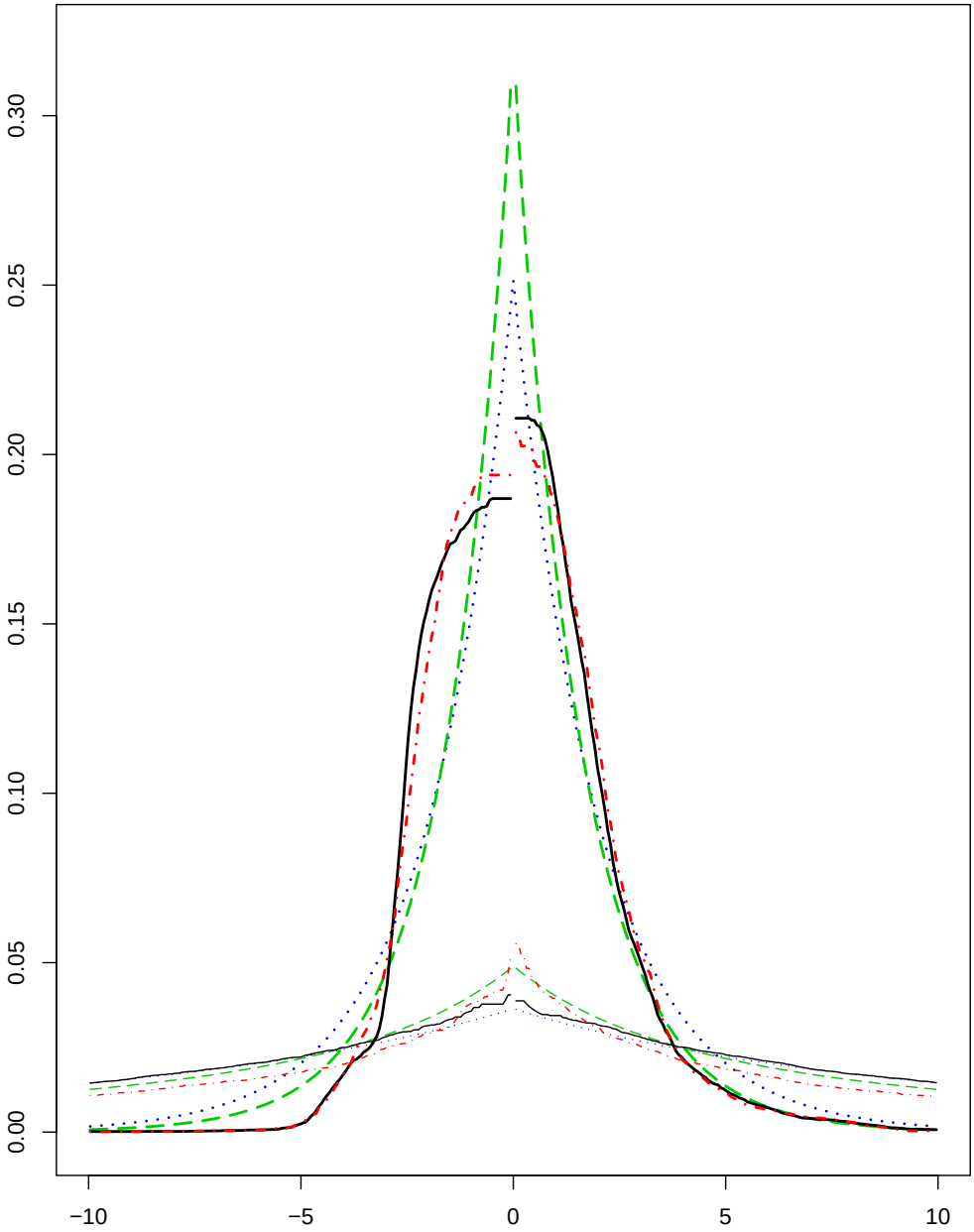

Figure 3: IgG data. Prior and posterior predictive error densities for the median regression case. Both are denoted by the dotted lines for model $\mathcal{M}_0$, by the dashed lines for model $\mathcal{M}_1$, by the solid lines for model $\mathcal{M}_2$, and by the dotted-dashed lines for model $\mathcal{M}_3$.

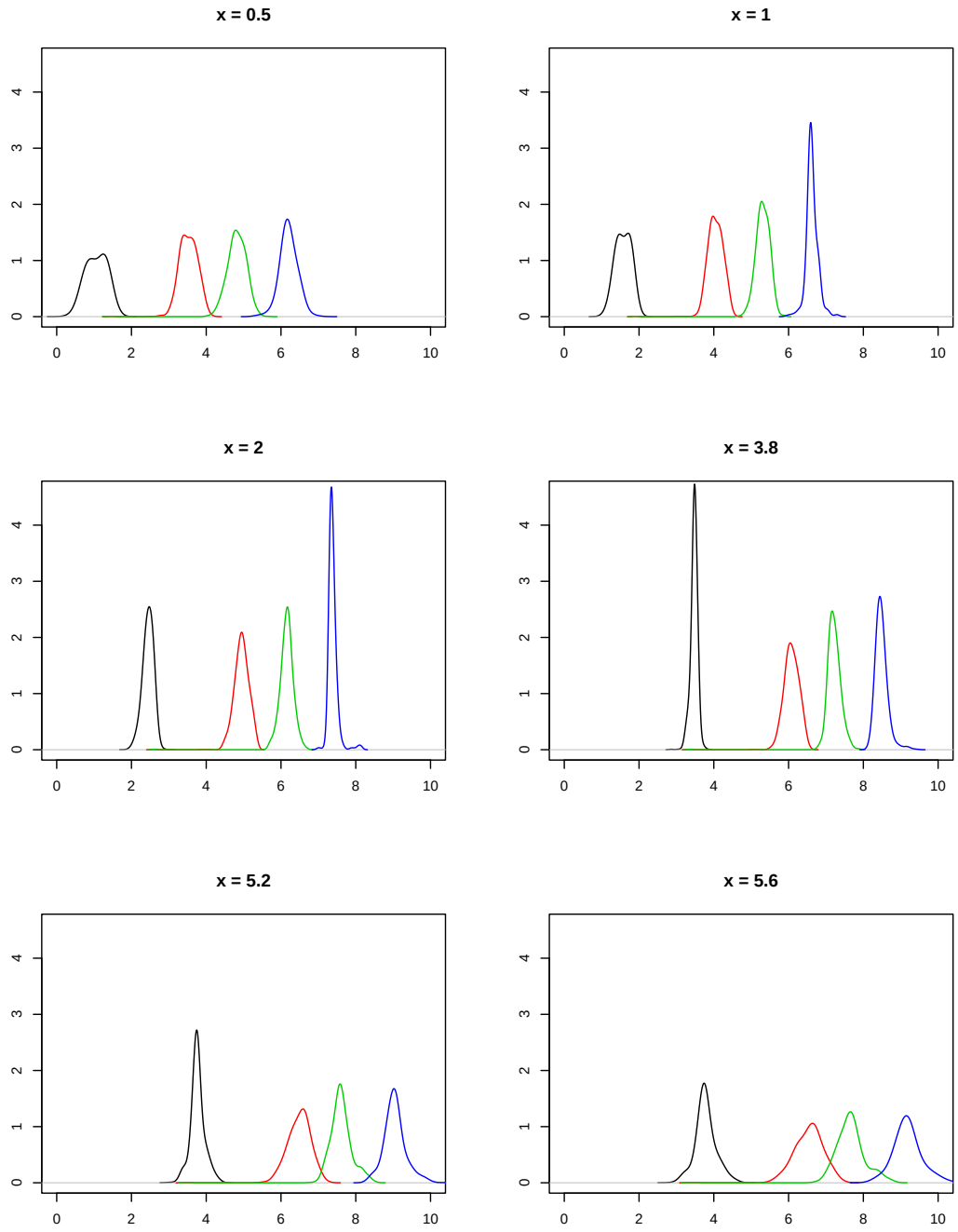

Figure 4: IgG data. Based on model $\mathcal{M}_2$, each panel provides posteriors for four quantiles, specifically, for $p = 0.05$, 0.5, 0.75, and 0.95. Included are results at six values for age, $x = 0.5$, 1, 2, 3.8, 5.2, and 5.6 years.

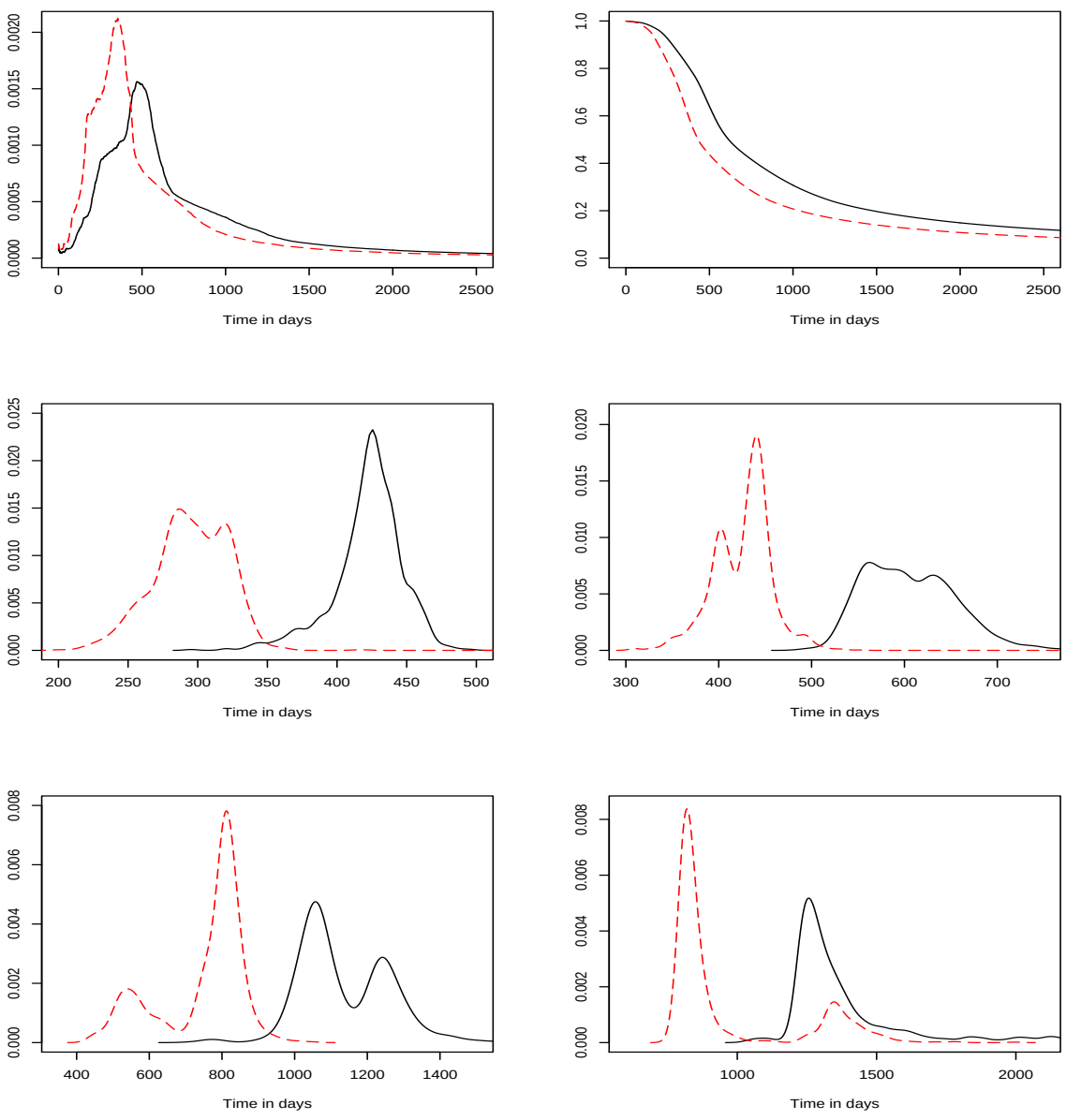

Figure 5: Small cell lung cancer data. The top row includes posterior predictive densities (left panel) and survival functions (right panel) under treatments A and B. The middle row displays the posteriors for the 25th percentile survival time (left panel) and for the median survival time (right panel). The bottom row shows the posteriors for the 75th and the 90th percentile survival times (left and right panels, respectively). All the results are based on model $\mathcal{M}_2$. In each panel the solid and dashed lines represent treatments A and B, respectively.

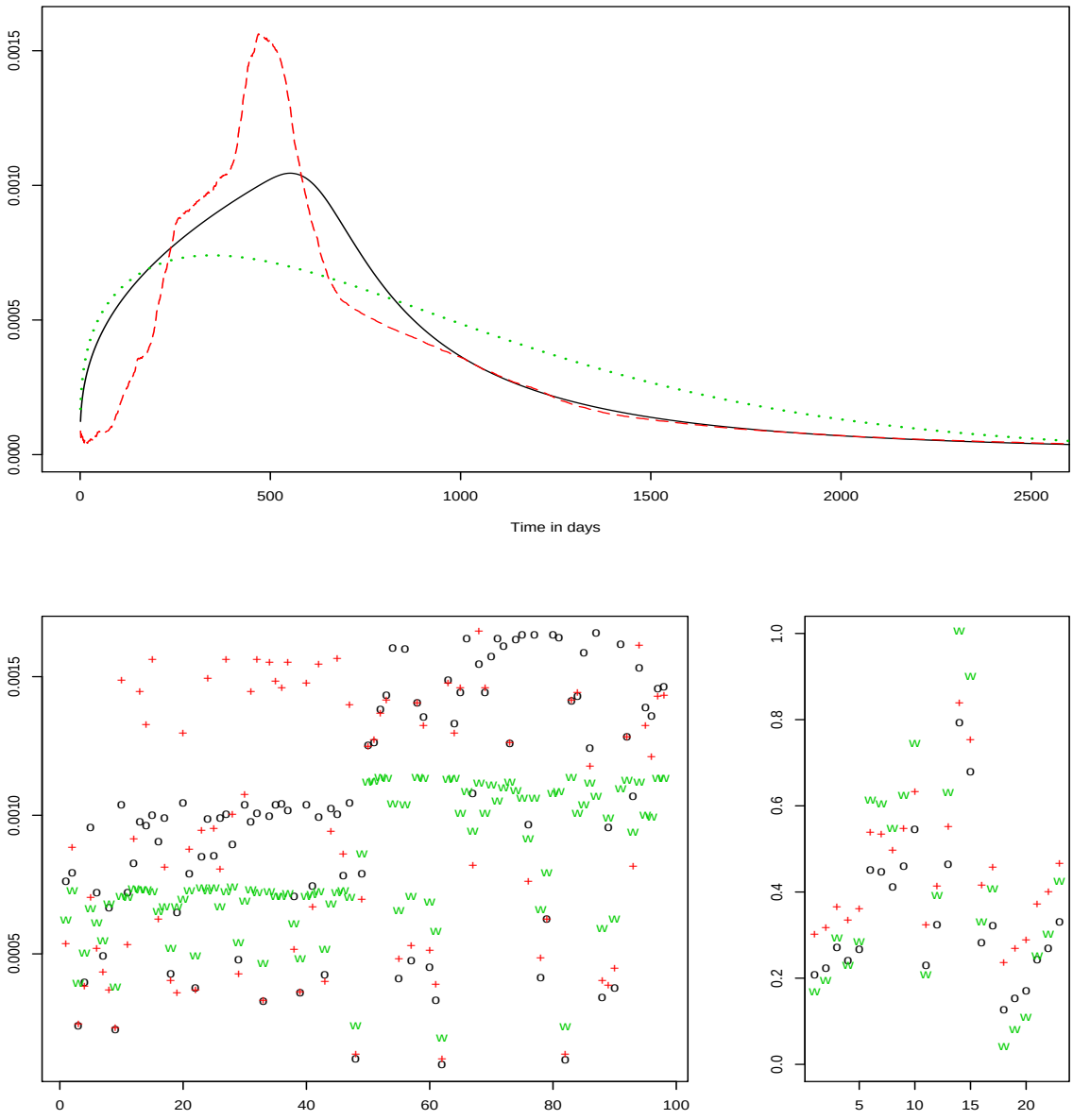

Figure 6: Small cell lung cancer data. The top panel displays posterior predictive densities for treatment A under model $\mathcal{M}_0$ (solid line), model $\mathcal{M}_2$ (dashed line), and a parametric Weibull model (dotted line). The bottom panels include CPO plots for the uncensored and the censored data (left and right panel, respectively). In both cases, "o" denotes points under model $\mathcal{M}_0$, "+" under model $\mathcal{M}_2$, and "w" under the Weibull model.

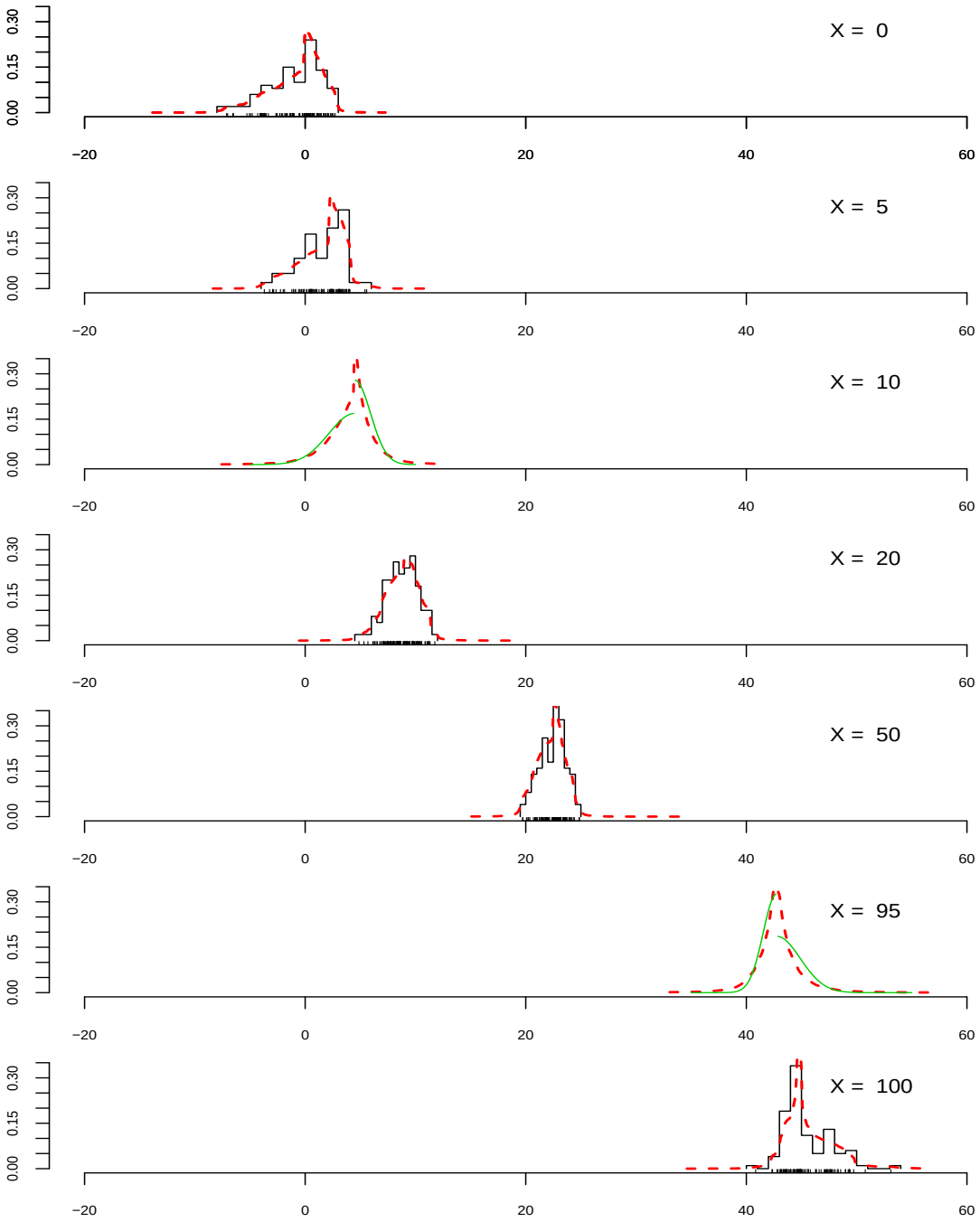

Figure 7: Simulation experiment for the DDP quantile regression model. Posterior predictive densites under the DDP model (dashed lines) at the five observed covariate values (overlaid on histograms of the corresponding response observations), and at two new covariate values, $x = 10$ and $x = 95$ (overlaid on corresponding true densities denoted by solid lines).

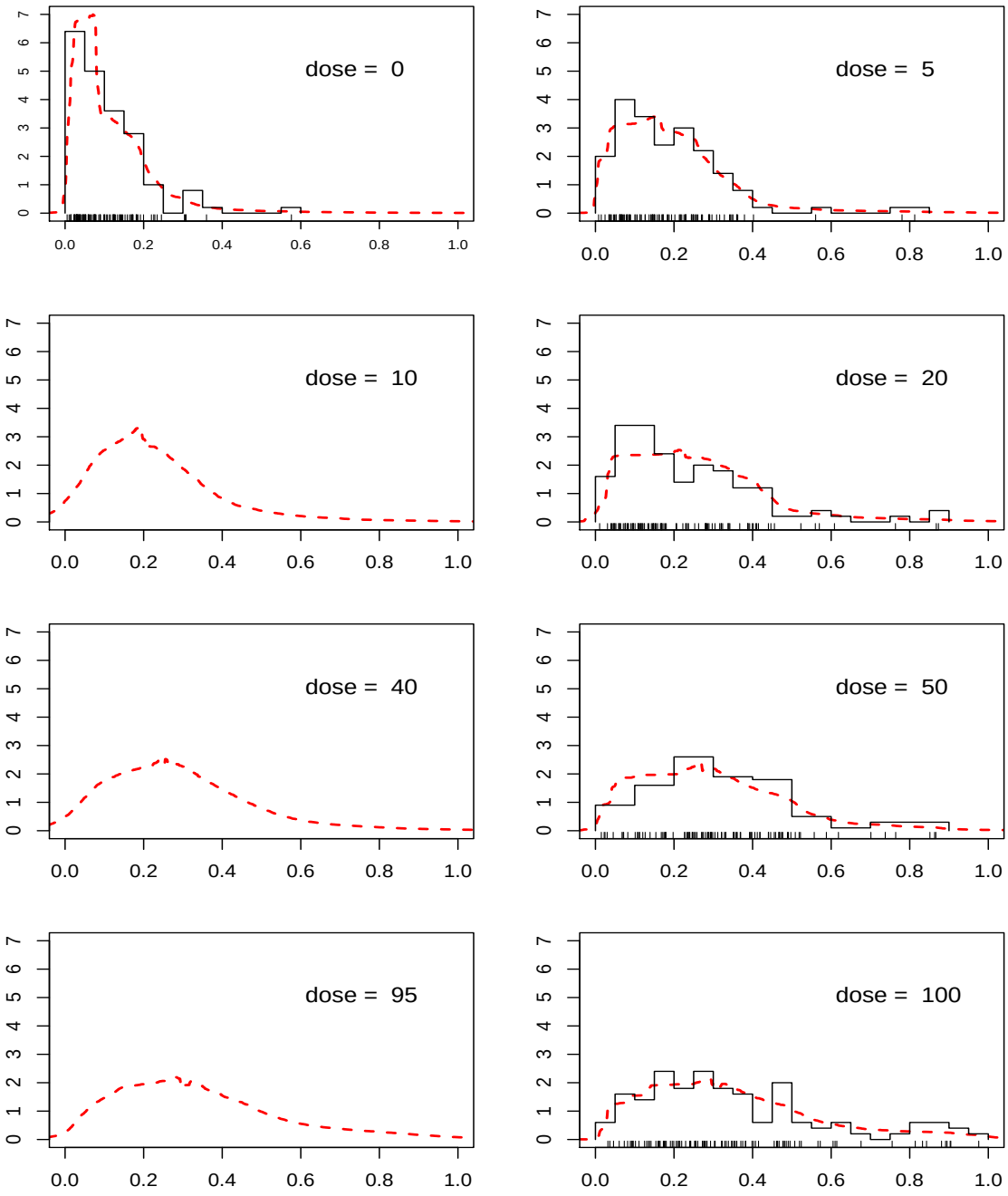

Figure 8: Comet assay data. Posterior predictive densites under the DDP model (dashed lines) at the five observed dose values, overlaid on histograms of the corresponding response observations, and at three new dose values (10, 40, and 95).