Scandinavian Journal of Statistics, Vol. 36: 297–319, 2009
doi: 10.1111/j.1467-9469.2008.00626.x

Bayesian Semiparametric Modelling in Quantile Regression

ATHANASIOS KOTTAS

Department of Applied Mathematics and Statistics, University of California, Santa Cruz

MILOVAN KRNJAJIĆ

University of California, Lawrence Livermore National Laboratory

ABSTRACT. We propose a Bayesian semiparametric methodology for quantile regression modelling. In particular, working with parametric quantile regression functions, we develop Dirichlet process mixture models for the error distribution in an additive quantile regression formulation. The proposed non-parametric prior probability models allow the shape of the error density to adapt to the data 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 that allow the error distribution to change non-parametrically 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, median regression, scale uniform mixtures, 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.

This paper develops a Bayesian semiparametric framework for quantile regression. We employ the standard additive regression formulation where the quantile regression function is separated from the errors in the response distribution. A distinguishing feature of the proposed approaches, relative to the main body of the existing literature, is that they are based on probabilistic modelling for unknown distributions. As importantly, this modelling relaxes, or avoids altogether, parametric assumptions that might not be supported by the data and whose influence on the results is typically difficult to anticipate. As illustrated with our data examples, this feature yields practically important advantages with regard to the range of resulting inferences.

Here, our focus is on flexible non-parametric modelling for the error distribution for applications where a parametric form for the quantile regression function suffices. (As discussed in section 5, the inferential scope of the framework can be extended by incorporating more flexible prior models for the regression function.) We develop non-parametric prior probability models for the error distribution, using Dirichlet process mixture models (Ferguson, 1973; Antoniak, 1974). 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, using dependent Dirichlet processes (MacEachern, 2000), we develop quantile regression models that allow the error distribution to change non-parametrically with the covariates.

The plan of the paper is as follows. Section 2 includes a general discussion of quantile regression to motivate our modelling framework. Moreover, it presents the proposed methodology, including approaches to prior specification, posterior inference for uncensored and censored data and model comparison. Section 3 provides illustrations of the models developed in section 2 through simulated and real data. Section 4 develops modelling for quantile regression with error densities that depend on the covariates, including a data example. Section 5 offers a summary and discussion of possible extensions. Finally, the two appendices include technical details on the MCMC methods for posterior inference, and on posterior propriety under a flat prior for one of the models.

2. Methodology

Section 2.1 discusses the standard additive quantile regression framework. Mixture models for the error distribution under this framework are explained in section 2.2. Sections 2.3 and 2.4 present methods for inference (with more details given in appendix A) and prior specification, respectively. We consider the extension to censored quantile regression in section 2.5. Model comparison is addressed in section 2.6.

2.1. The modelling framework for quantile regression

Quantile regression approaches most typically build on the traditional mean (or median) regression setting where the mean (median) of the response distribution is modelled as a function of covariates. Hence, this setting requires response distributions that are parameterized in terms of their mean (median), which is set equal to 0 in the corresponding error distribution. Extending this standard additive regression formulation, the pth quantile regression model for (continuous) response observations y_i, with associated covariate vectors $\mathbf{x}_i$, $i=1,\ldots,n$, can be written as

$$y_i = h(\mathbf{x}_i) + \varepsilon_i. \tag{1}$$

Here, the ε_i are assumed independent from an error distribution with pth quantile equal to 0, i.e. $\int_{-\infty}^{0} f_p(\varepsilon)\,\mathrm{d}\varepsilon = p$, with $f_p(\cdot)$ denoting the error density.

Our objective is to develop flexible non-parametric prior models for the random error density $f_p(\cdot)$. We model $h(\cdot)$ parametrically and, for clarity of exposition, write $h(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, 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, non-parametric modelling for $h(\cdot)$ could be proposed in addition to our modelling for $f_p(\cdot)$; section 5 includes a brief discussion of this latter extension.

There is an extensive literature on classical estimation for model (1); see, e.g. the review paper by Yu *et al.* (2003) and the book by Koenker (2005). These works are dominated by *semiparametric* techniques where $h(\mathbf{x})$ is typically expressed as $\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, and the error density $f_p(\cdot)$ is left unspecified (apart from the restriction $\int_{-\infty}^{0} f_p(\varepsilon)\,\mathrm{d}\varepsilon = p$). Hence, without probabilistic modelling 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 $\sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, 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 modelling 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 non-parametrics literature (see, e.g. Walker & Mallick, 1999; Kottas & Gelfand, 2001; Hanson & Johnson, 2002), relatively little work exists for general quantile regression modelling. The existing literature includes the parametric approaches in Yu & Moyeed (2001) and Tsionas (2003), which are based on the asymmetric Laplace distribution for the errors. Moreover, the work of Hjort & Petrone (2007) studies non-parametric inference for the quantile function, including discussion of the extension to quantile regression.

Here, we develop two families of non-parametric error distributions based on Dirichlet process (DP) mixture models. The first model, a scale mixture of asymmetric Laplace densities, extends the parametric work of Yu & Moyeed (2001) and Tsionas (2003). Motivated by limitations of this model, we propose a flexible scale mixture of uniform densities, which can capture the shape (e.g. skewness, tail behaviour) of general unimodal error densities $f_p(\cdot)$. Again, our focus is on building a model-based, fully inferential framework for semiparametric quantile regression. In particular, we place emphasis on practically important extensions of model (1) to censored quantile regression (section 2.5) and quantile regression process models that allow the error density to change non-parametrically with the covariates (section 4).

We conclude this section with a critical discussion of the regression formulation in (1). Note that, 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 or for extreme percentiles. And this attribute of the additive formulation (1) 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 (1) with $h(\mathbf{x})=\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}$, *simultaneous* estimation for more than one quantile regression (see, e.g. Dunson & Taylor, 2005); however, this is only possible because they do not involve probabilistic modelling for the errors. 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.

Section 5 discusses an alternative approach to quantile regression that does not follow the structure of model (1) (and thus, to some extent, sacrifices interpretability) and is more computationally intensive, but yields a flexible, fully non-parametric, method of simultaneous inference for quantile curves.

2.2. Mixture modelling for the error distribution

Non-parametric scale mixture of asymmetric Laplace densities. A natural starting point in constructing a non-parametric model for the random error density in (1) is to extend a parametric class of distributions through appropriate mixing. The parametric model that is typically used in this context is the family of asymmetric Laplace distributions with densities

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

where $0<p<1$, $\sigma>0$ is a scale parameter, and $\int_{-\infty}^{0} k_p^{\mathrm{AL}}(\varepsilon;\sigma)\,\mathrm{d}\varepsilon=p$. For instance, Yu & Moyeed (2001) and Tsionas (2003) used (2) for the errors in (1), working with $h(\mathbf{x}_i)=\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. (In fact, Yu & Moyeed, 2001, worked with the special case of this parametric model where $\sigma=1$.) Note that parameter p determines both skewness and pth quantile for the density in (2), hence limiting its flexibility in modelling skewness and tail behaviour. In particular, $k_p^{\mathrm{AL}}(\cdot;\sigma)$ is skewed for $p\neq 0.5$ and symmetric for $p=0.5$, i.e. for the median regression case. This is a 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 $f_p(\cdot)=k_p^{\mathrm{AL}}(\cdot;\sigma)$ as model $\mathcal{M}_0$.

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

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

Note that mixing in this fashion preserves the quantiles, i.e. $\int_{-\infty}^{0} f_p^1(\varepsilon;G)\,\mathrm{d}\varepsilon=p$. 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

$$\begin{aligned}
&Y_i\,|\,\boldsymbol{\beta},\sigma_i \overset{\text{ind.}}{\sim} k_p^{\mathrm{AL}}(y_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta};\sigma_i), \quad i=1,\ldots,n\\
&\sigma_i\,|\,G \overset{\text{i.i.d.}}{\sim} G, \quad i=1,\ldots,n\\
&G\,|\,\alpha,d\sim\mathrm{DP}(\alpha,G_0)
\end{aligned} \tag{4}$$

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 behaviour 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 $k_p^{\mathrm{AL}}(\cdot;\sigma)$ regarding skewness.

Non-parametric scale mixture of uniform densities. The key result for constructing more flexible models than $\mathcal{M}_0$ and $\mathcal{M}_1$ is the representation of non-increasing densities on the positive real line as scale mixtures of uniform densities. Specifically, a density $f(\cdot)$ on R^+ is non-increasing if and only if there exists a distribution function G on R^+ such that $f(t)\equiv f(t;G)=\int \theta^{-1}1_{[0,\theta)}(t)\,\mathrm{d}G(\theta)$ (see, e.g. Feller, 1971, p. 158). The representation requires a general mixing distribution G and thus, for Bayesian modelling, motivates naturally the use of a non-parametric prior for G; see, e.g. Brunner & Lo (1989), Brunner (1992, 1995), Lavine & Mockus (1995), Kottas & Gelfand (2001) and Hansen & Lauritzen (2002) for DP-based modelling involving variations of this representation.

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

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

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

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

for the error density in (1). In the special case of median regression (i.e. $p=0.5$), (6) reduces to the non-parametric error model in Kottas & Gelfand (2001). For quantile regression, $f_p^2(\cdot; G_1, G_2)$ is sufficiently flexible to capture general forms of skewness and tail behaviour. As mentioned above, a consequence of the representation theorem that leads to the formulation of $f_p^2(\cdot; G_1, G_2)$ is that the mode of the error density is equal to 0. In a strict sense, this could be viewed as a limitation of model (6). However, note that, in the context of the framework implied by (1), it does not seem natural to seek unimodal error densities with mode significantly different than the specific pth quantile. We argue that the inferential benefits resulting from the flexibility of (6) outweigh any potential restriction from this particular model attribute.

We use Gamma priors for the DP precision parameters 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

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

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

Note that the formulation in (6) indicates an alternative non-parametric family of error densities, $\iint k_p(\varepsilon; \sigma_1, \sigma_2)\, \mathrm{d}G(\sigma_1, \sigma_2)$, using a single mixing distribution G supported on $R^+ \times R^+$. The corresponding semiparametric quantile regression model based on a DP prior, $\mathrm{DP}(\alpha, G_0^*)$ for G, where now G_0^* is a parametric distribution on $R^+ \times R^+$, is developed in the technical report version of the paper (Kottas & Krnjajić, 2005). The predictive performance of this model for the data sets considered in section 3 was similar to that of model $\mathcal{M}_2$.

2.3. Posterior inference

We obtain inference under the models developed in section 2.2 utilizing well-established posterior simulation methods for DP mixture models. In particular, we use a combination of MCMC methods from Escobar & West (1995), Bush & MacEachern (1996) and Neal (2000). (The details are given in appendix A.) These methods are based on a marginalization of the random mixing distributions over their DP priors (Blackwell & MacQueen, 1973). Draws from the resulting marginalized posteriors yield the posterior predictive distribution for a *new* response Y_{new} with corresponding covariate vector $\mathbf{x}_{\text{new}}$.

We illustrate with model $\mathcal{M}_2$. Denote data $=\{(y_i, \mathbf{x}_i) : i=1,\ldots,n\}$, and let $\boldsymbol{\psi}$ collect all model parameters, $\boldsymbol{\psi} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}_r = (\sigma_{ri} : i=1,\ldots,n), \alpha_r, d_r : r=1,2\}$. The discreteness of the DP priors (Ferguson, 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}^*, $j=1,\ldots,n_r^*$, be the distinct σ_{ri}, i.e. the cluster locations. Because G_{r0} are continuous distributions, the clusters are determined by a vector of configuration indicators $\mathbf{s}_r = (s_{r1},\ldots,s_{rn})$ such that $s_{ri}=j$ if

and only if $\sigma_{ri} = \sigma^*_{rj}$ for $i = 1, \ldots, n$. Moreover, denote by n_{rj} the size of the jth cluster for $j = 1, \ldots, n^*_r$. Evidently, $\{n^*_r, \mathbf{s}_r, (\sigma^*_{rj} : j = 1, \ldots, n^*_r)\}$ yields an equivalent representation for $\boldsymbol{\sigma}_r$. Now the posterior predictive density for Y_{new} can be expressed as

$$p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \text{data}) = \int k_p(\varepsilon_{\text{new}}; \sigma_{1,\text{new}}, \sigma_{2,\text{new}}) p(\sigma_{1,\text{new}} \mid \boldsymbol{\psi}) p(\sigma_{2,\text{new}} \mid \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid \text{data}), \tag{8}$$

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

$$p(\sigma_{r,\text{new}} \mid \boldsymbol{\psi}) = \frac{\alpha_r}{\alpha_r + n} \mathrm{G}_{r0}(\sigma_{r,\text{new}} \mid d_r) + \frac{1}{\alpha_r + n} \sum_{j=1}^{n^*_r} n_{rj} \delta_{\sigma^*_{rj}}(\sigma_{r,\text{new}}), \tag{9}$$

with δ_a denoting a point mass at a. Note that, combining (8) and (9), we can also write $p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \text{data}) = \int p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \boldsymbol{\psi}) p(\boldsymbol{\psi} \mid \text{data})$, where $p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \boldsymbol{\psi})$ is given by

$$\begin{aligned} &\frac{\alpha_1}{\alpha_1 + n} \int \frac{p}{\sigma_1} 1_{(-\sigma_1, 0)}(\varepsilon_{\text{new}}) \, \mathrm{d}G_{10}(\sigma_1 \mid d_1) + \frac{\alpha_2}{\alpha_2 + n} \int \frac{(1-p)}{\sigma_2} 1_{[0, \sigma_2)}(\varepsilon_{\text{new}}) \, \mathrm{d}G_{20}(\sigma_2 \mid d_2) \\ &\quad + \frac{1}{\alpha_1 + n} \sum_{j=1}^{n^*_1} n_{1j} \frac{p}{\sigma^*_{1j}} 1_{(-\sigma^*_{1j}, 0)}(\varepsilon_{\text{new}}) + \frac{1}{\alpha_2 + n} \sum_{j=1}^{n^*_2} n_{2j} \frac{(1-p)}{\sigma^*_{2j}} 1_{[0, \sigma^*_{2j})}(\varepsilon_{\text{new}}). \end{aligned} \tag{10}$$

The first two components in (10) 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, the posteriors for α_r, $r = 1, 2$, are typically supported by values small relative to n, whence, with $\alpha_r/(\alpha_r + n) \approx 0$, $p(y_{\text{new}} \mid \mathbf{x}_{\text{new}}, \boldsymbol{\psi})$ can be approximated by the last two terms in (10). 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.

2.4. 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. A sensitivity analysis for the data sets considered in sections 3.2, 3.3 and 4.2 indicated that the posteriors of the regression coefficients were robust to a range of choices for the corresponding prior variances. Although not used in our data analyses, one might wish to also consider a flat (improper) prior for $\boldsymbol{\beta}$. Establishing propriety of the resulting posterior for $\boldsymbol{\beta}$ under the various models developed in sections 2.2 and 4.1 is beyond the scope of this paper. However, appendix B provides a result in this direction for a simplified version of model $\mathcal{M}_2$.

We use relatively dispersed Gamma priors for the DP precision parameters, 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 & West, 1995). In general, posterior predictive inference was robust to the prior choice for the precision parameters.

We work with prior predictive densities to specify the hyperparameters of the DP base distributions. In addition to revealing the effect of certain choices for the prior hyperparameter values, prior predictive densities facilitate empirical model comparison, because they enable 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 2.2; e.g. under model $\mathcal{M}_2$, the prior predictive density at y_0, with corresponding covariate vector $\mathbf{x}_0$, is given by

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

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

2.5. Bayesian semiparametric censored quantile regression

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

All the quantile regression models of section 2.2 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}, $i_o = 1, \ldots, n_o$, are observed, whereas for the remaining n_c survival times t_{i_c}, $i_c = 1, \ldots, n_c$, we have $t_{i_c} > z_{i_c}$ for known censorship times z_{i_c}. Then, the only change required for model $\mathcal{M}_2$ is in the first stage of the corresponding hierarchical model 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 $\mathbf{x}_{i_o}$ and $\mathbf{x}_{i_c}$ denoting the corresponding covariate vectors, the first stage in (7) becomes

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

where $K_p(\cdot; \sigma_1, \sigma_2)$ denotes the distribution function of $k_p(\cdot; \sigma_1, \sigma_2)$. The Gibbs samplers under censoring have the same structure with the case of fully observed data (detailed in appendix A). 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 & 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 two parametric alternatives.

2.6. Model comparison

Given the different semiparametric specifications discussed in section 2.2 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 *et al.*, 2001). For model $\mathcal{M}$, and specified covariate vector $\mathbf{x}$, denoted by $p^{\mathcal{M}}(\cdot \mid \mathbf{x}, \text{data})$ and $S^{\mathcal{M}}(\cdot \mid \mathbf{x}, \text{data})$ 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 $p^{\mathcal{M}}(t_{i_o} \mid \mathbf{x}_{i_o}, \text{data}(-i_o))$, whereas the CPO for a right censored survival time t_{i_c} is defined as $S^{\mathcal{M}}(z_{i_c} \mid \mathbf{x}_{i_c}, \text{data}(-i_c))$, where $\text{data}(-i_o)$ and $\text{data}(-i_c)$ 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

$$Q(\mathcal{M}) = n^{-1} \sum_{i_o=1}^{n_o} \log p^{\mathcal{M}}(t_{i_o} \mid \mathbf{x}_{i_o}, \text{data}(-i_o)) + n^{-1} \sum_{i_c=1}^{n_c} \log S^{\mathcal{M}}(z_{i_c} \mid \mathbf{x}_{i_c}, \text{data}(-i_c)) \tag{11}$$

(see, e.g. Bernardo & 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 & Ghosh (1998). The criterion favours the model $\mathcal{M}$ which minimizes

$$D_m(\mathcal{M}) = \sum_{i=1}^{n} V^{\mathcal{M}}(i) + \frac{m}{m+1} \sum_{i=1}^{n} \{y_i - E^{\mathcal{M}}(i)\}^2, \tag{12}$$

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 $Y_{\text{new},i}$ with associated covariate vector $\mathbf{x}_i$. The first component in (12) 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 of section 2.4 for prior specification.

3.1. Simulation study

To assess the performance of the error models discussed in section 2.2, 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 0.6th 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 behaviour. The true densities under the six cases of this simulation experiment are included in Fig. 1. All the samples were of size $n = 250$.

Model $\mathcal{M}_2$ captures very successfully the different density shapes as illustrated in Fig. 1. Results under model $\mathcal{M}_1$ (not shown) indicate that this model fits well the data generated from the asymmetric Laplace distributions, but fares worse than model $\mathcal{M}_2$ for the other three data sets. For instance, for the data drawn from the normal mixture distribution with 0.6th quantile at zero, model $\mathcal{M}_1$ yields a heavier left tail and a higher peak than the ones suggested by the data, and captured by model $\mathcal{M}_2$.

3.2. Immunoglobulin-G data set

Here, we work with data discussed in Royston & Altman (1994), and used also by Yu & 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/litre) of immunoglobulin-G (IgG) for 298 children, with ages from 6 months to 6 years. As in Yu & 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 $D_m(\mathcal{M})$ in (12) to compare models $\mathcal{M}_0$, $\mathcal{M}_1$ and $\mathcal{M}_2$. Table 1 provides results for $m = 1$ and $m \to \infty$, and for five 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 five quantiles. We note that model $\mathcal{M}_0$ and, to a lesser extent, model $\mathcal{M}_1$, perform substantially worse than model $\mathcal{M}_2$ at the low and high quantile values. This could be attributed to

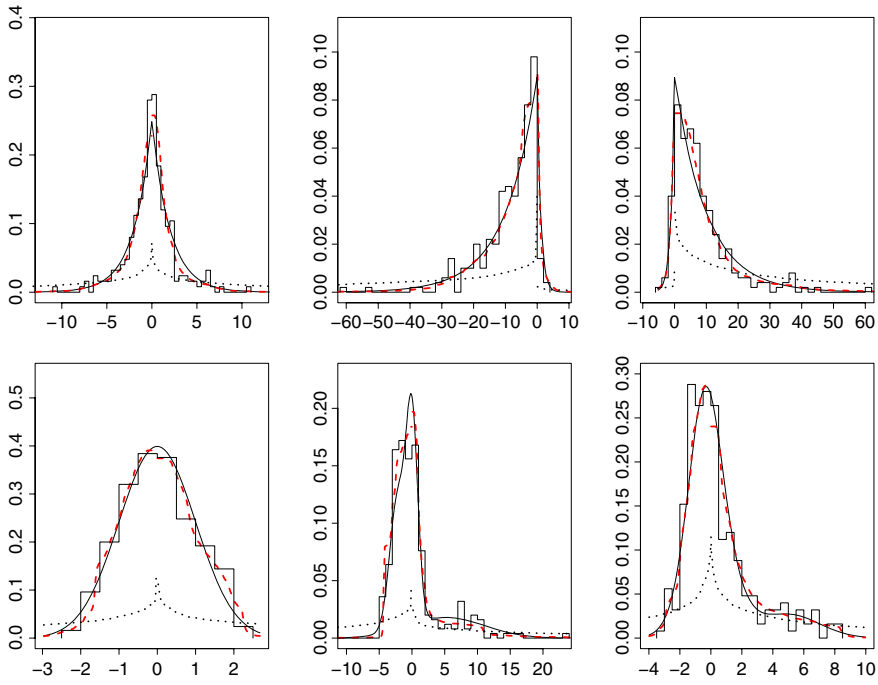

Fig. 1. Simulation study. Prior predictive (dotted lines) and posterior predictive (dashed lines) densities under model $\mathcal{M}_2$. The top panels correspond to the standard Laplace densities, with $p=0.5$, 0.9 and 0.1 for the left, middle and right panel, respectively. The bottom panels include the standard normal density (left panel), and the two normal mixture densities, the first (middle panel) with 0.6th 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.

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* $\mathcal{M}=\mathcal{M}_0$*,* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *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
$\mathcal{M}_0$	0.25	5185	1541	5956	6726
$\mathcal{M}_1$	0.25	2351	1181	2942	3533
$\mathcal{M}_2$	0.25	1361	1157	1940	2518
$\mathcal{M}_0$	0.50	2866	1151	3442	4018
$\mathcal{M}_1$	0.50	1811	1147	2385	2959
$\mathcal{M}_2$	0.50	1555	1152	2132	2708
$\mathcal{M}_0$	0.75	5592	1811	6498	7404
$\mathcal{M}_1$	0.75	2338	1261	2969	3599
$\mathcal{M}_2$	0.75	1898	1165	2481	3064
$\mathcal{M}_0$	0.95	114240	71518	150000	185759
$\mathcal{M}_1$	0.95	58460	2373	59646	60833
$\mathcal{M}_2$	0.95	1220	1149	1795	2369

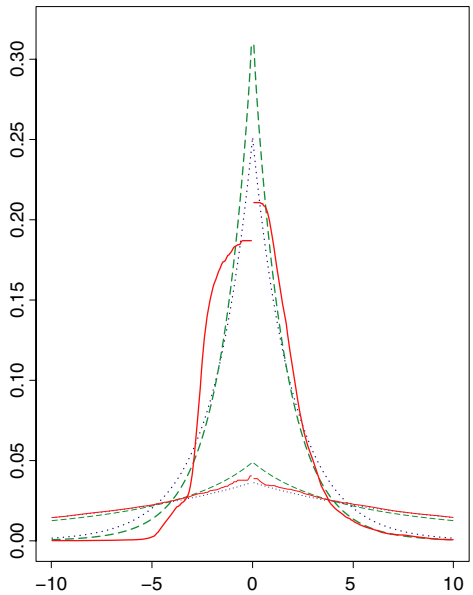

Fig. 2. IgG data. Prior and posterior predictive error densities for the median regression case. Both are denoted by the dotted lines for model $\mathcal{M}_0$, dashed lines for model $\mathcal{M}_1$ and solid lines for model $\mathcal{M}_2$.

the restrictive feature of models $\mathcal{M}_0$ and $\mathcal{M}_1$ discussed in section 2.2, 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 three models, are given in Fig. 2. 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 model $\mathcal{M}_2$ indicate that the error density is skewed. To illustrate how the quantiles of the IgG serum concentration distribution change with age, Fig. 3 shows the posteriors of $\beta_0+\beta_1 x+\beta_2 x^2$, under model $\mathcal{M}_2$, of 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 analysed using median regression models in Ying *et al.* (1995), Walker & Mallick (1999), Yang (1999) and Kottas & 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, a covariate that, based on the earlier analyses, does not appear to be very important in explaining the response.)

We fit model $\mathcal{M}_2$ to this data set using a $\log_{10}$ transformation of the survival times. Figure 4 provides posterior predictive densities and survival functions under both treatments. It also compares the posteriors for 0.25th quantile, median, 0.75th quantile and 0.90th 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 Fig. 5).

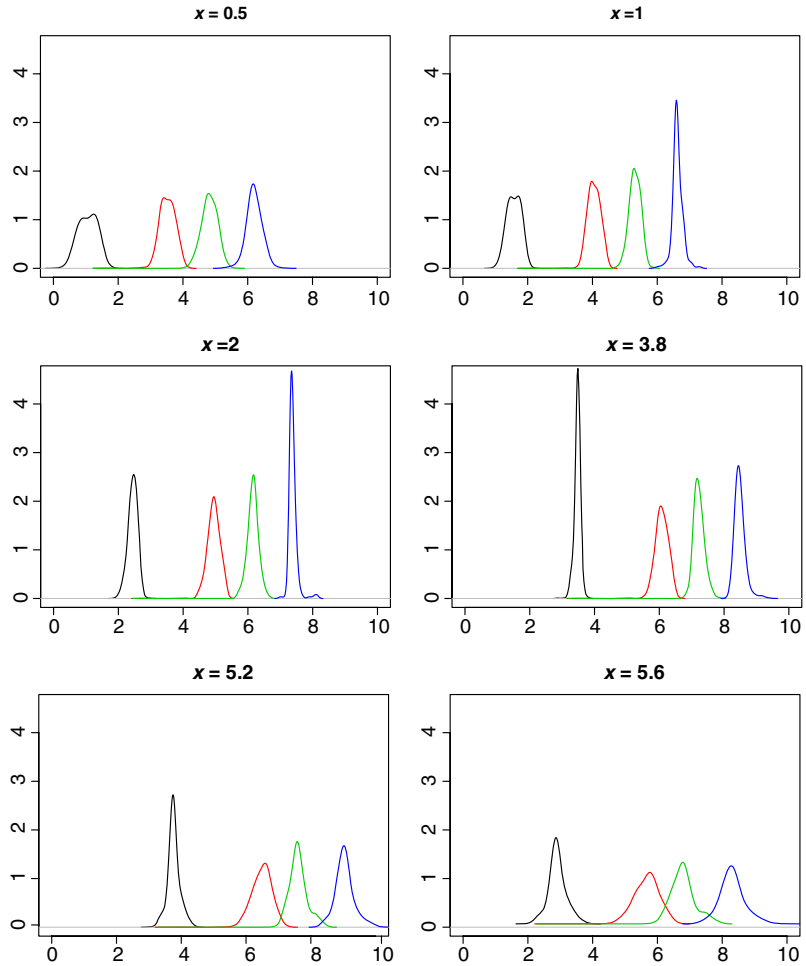

Fig. 3. IgG data. Based on model $\mathcal{M}_2$, each panel provides posteriors for four quantiles, specifically, for $p = 0.05$, 0.5, 0.75 and 0.95. Included are results of six values for age, $x = 0.5$, 1, 2, 3.8, 5.2 and 5.6 years.

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 pth quantile survival time has the simple form $[-\log(1-p)\exp\{-(\beta_0+\beta_1 x)\}]^{1/\gamma}$. The CPO plots (Fig. 5) indicate a superior predictive performance of model $\mathcal{M}_2$ compared with the two parametric models, as does the cross-validation criterion $Q(\mathcal{M})$ in (11), taking values -8.01, -6.91 and -11.56 for models $\mathcal{M}_0$, $\mathcal{M}_2$ and the Weibull model, respectively.

4. Dependent non-parametric error distributions for quantile regression

4.1. The modelling approach

Here we propose an extension of the standard modelling framework in (1) to a class of quantile regression models where the error density $f_p(\cdot)$ depends on the covariates. For a

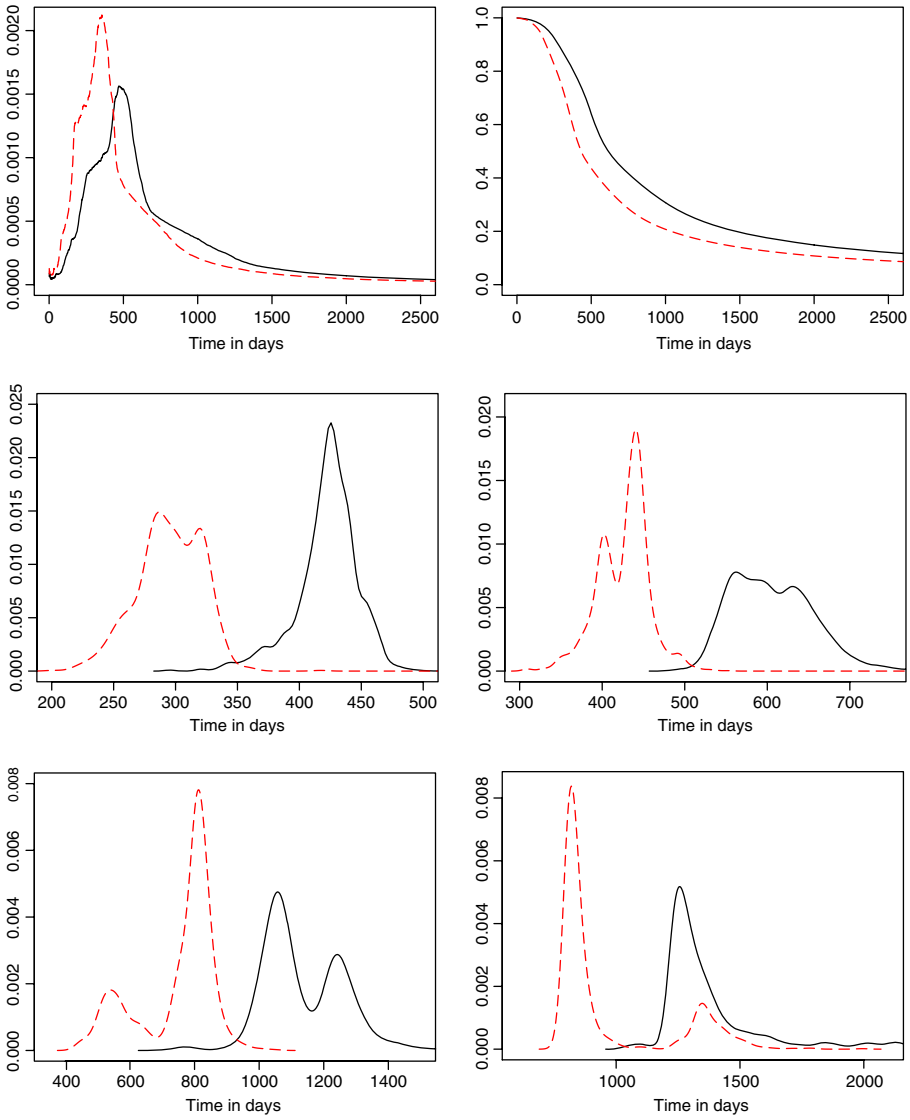

Fig. 4. Small cell lung cancer data. The top row includes posterior predictive densities (left panel) and survival functions (right panel) under treatments A and B. The middle row displays the posteriors for the 25th percentile survival time (left panel) and for the median survival time (right panel). 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.

simpler exposition, we consider a single continuous covariate x with observed values x_m, $m=1,\ldots,M$. 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 pth quantile $\beta_0+\beta_1 x$. Extension to non-parametric covariate-dependent error distributions requires a non-parametric 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, $\int_{-\infty}^{0} f_{p,x}(\varepsilon)\,\mathrm{d}\varepsilon=p$.

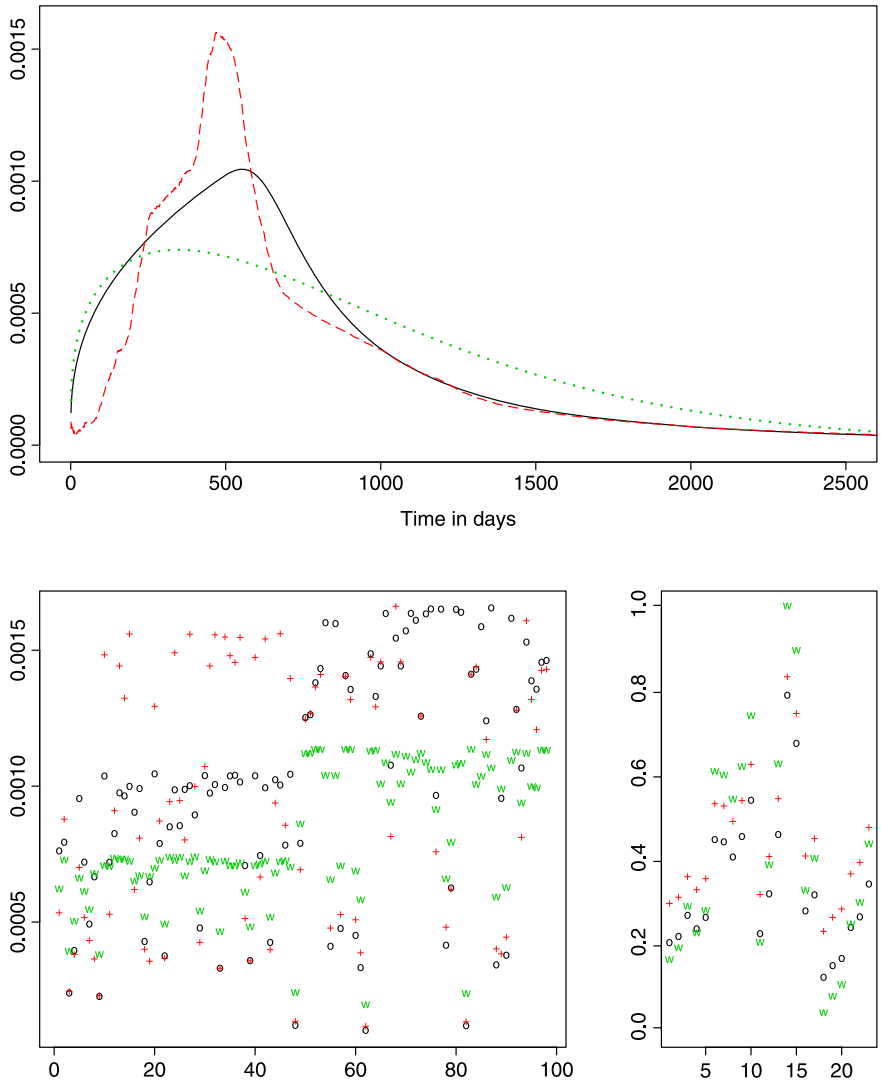

Fig. 5. Small cell lung cancer data. The top panel displays posterior predictive densities of survival times for treatment A under model $\mathcal{M}_0$ (solid line), model $\mathcal{M}_2$ (dashed line) and a parametric Weibull model (dotted line). The bottom panels include CPO plots for the uncensored and the censored data (left and right panel, respectively). In both cases, 'o' denotes CPO values under model $\mathcal{M}_0$, '+' under model $\mathcal{M}_2$ and 'w' under the Weibull model.

Hence, in this setting, $f_{p,x}(\cdot)$ and $f_{p,x'}(\cdot)$ are dependent for $x \neq x'$. 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 probability model for $f_{p,\mathcal{X}}$. The DDP was developed by MacEachern (1999; 2000) as a non-parametric prior for a stochastic process of random distributions; these distributions are dependent but such that, at each index value, the distribution is a DP (see also De Iorio *et al.*, 2004; Gelfand *et al.*, 2005).

We provide the details building on model $\mathcal{M}_2$. 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

$$f_p^2(\varepsilon; G_1, G_2) = \int\int k_p(\varepsilon; \theta_1, \theta_2)\, \mathrm{d}G_1(\theta_1)\, \mathrm{d}G_2(\theta_2),$$
$$G_r \sim \mathrm{DP}(\alpha_r, G_{r0}),$$

$r = 1, 2$, where now we could take $G_{r0} = N(\mu_r, \tau_r^2)$, $r = 1, 2$, with μ_r and/or τ_r^2 random.

To allow $f_p^2(\varepsilon; G_1, G_2)$ to change with x, we need mixing distributions G_1 and G_2 that change with x and are assigned non-parametric priors; we need prior 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 DP definition 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,

$$\omega_{r,1} = z_{r,1}, \omega_{r,\ell} = z_{r,\ell} \prod_{s=1}^{\ell-1} (1 - z_{r,s}), \quad \ell = 2, 3, \ldots,$$

with $z_{r,s}$ i.i.d. Beta$(1, \alpha_r)$. Moreover, the sequences $\{\psi_{r,\ell} : \ell = 1, 2, \ldots\}$ and $\{z_{r,s} : s = 1, 2, \ldots\}$ 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

$$G_{r,\mathcal{X}} = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \delta_{\psi_{r,\ell,\mathcal{X}}} \tag{13}$$

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 (13) is that for any finite set of covariate values the induced prior is a DP. Specifically, for any collection of x values, $\mathbf{u} = (x_1, \ldots, x_L)$, which can include both observed and new covariate values, we have

$$G_{r,\mathbf{u}} = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \delta_{\psi_{r,\ell}(\mathbf{u})},$$

where the L-dimensional random vectors $\psi_{r,\ell}(\mathbf{u}) = (\psi_{r,\ell}(x_1), \ldots, \psi_{r,\ell}(x_L))$ are i.i.d. with distribution $G_{r0}(\mathbf{u})$ induced by the stochastic process $G_{r0,\mathcal{X}}$ at $\mathbf{u}$. Therefore, if $\theta_{r,\mathcal{X}} \mid G_{r,\mathcal{X}} \sim G_{r,\mathcal{X}}$, then $G_{r,\mathcal{X}}$ induces at $\mathbf{u}$ a $\mathrm{DP}(\alpha_r, G_{r0}(\mathbf{u}))$ prior on the space of distribution functions for $(\theta_r(x_1), \ldots, \theta_r(x_L))$. Note that, although the $\psi_{r,\ell}(\mathbf{u})$ are i.i.d. from $G_{r0}(\mathbf{u})$, for any ℓ, the random variables $\psi_{r,\ell}(x_1), \ldots, \psi_{r,\ell}(x_L)$ are dependent. Hence, in addition to modelling different error density shapes for different observed covariate values, the DDP approach provides 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, $E(\theta_r(x) \mid \mu_r) = \mu_r$, and variance, $\mathrm{var}(\theta_r(x) \mid \tau_r^2) = \tau_r^2$, and exponential correlation function $\mathrm{corr}(\theta_r(x), \theta_r(x') \mid \phi_r) = \exp(-\phi_r \mid x - x' \mid)$, 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 $\mathbf{x} = (x_1, \ldots, x_M)$, $G_{r0}(\mathbf{x})$ is an M-dimensional normal with mean vector $\mu_r 1_M$ and covariance matrix V_r with (i,j)th element $\tau_r^2 \exp(-\phi_r |x_i - x_j|)$, $i, j = 1, \ldots, M$. Departures from the Gaussianity and stationarity structure implied by the centre $G_{r0,\mathcal{X}}$ of the DDP prior, emerge through the countable mixing in (13). In particular, for any x, $x' \in \mathcal{X}$,

$$E(\theta_r(x) \mid G_{r,\mathcal{X}}) = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \psi_{r,\ell}(x)$$

and

$$\operatorname{cov}(\theta_r(x), \theta_r(x') \mid G_{r,\mathcal{X}}) = \sum_{\ell=1}^{\infty} \omega_{r,\ell} \psi_{r,\ell}(x) \psi_{r,\ell}(x') - \left\{ \sum_{\ell=1}^{\infty} \omega_{r,\ell} \psi_{r,\ell}(x) \right\} \left\{ \sum_{\ell=1}^{\infty} \omega_{r,\ell} \psi_{r,\ell}(x') \right\}.$$

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

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

with $G_{r,x} \sim \mathrm{DP}(\alpha_r, G_{r0}(x))$, $r = 1, 2$, and $G_{r0}(x) = N(\mu_r, \tau_r^2)$. 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 $\mathbf{x}$, we can write

$$f_{p,\mathbf{x}}^2(\boldsymbol{\varepsilon}; G_{1,\mathbf{x}}, G_{2,\mathbf{x}}) = \iint \prod_{m=1}^{M} k_p(\varepsilon_m; \theta_1(x_m), \theta_2(x_m)) \, \mathrm{d}G_{1,\mathbf{x}}(\boldsymbol{\theta}_1) \, \mathrm{d}G_{2,\mathbf{x}}(\boldsymbol{\theta}_2),$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_M)$, $\boldsymbol{\theta}_r = (\theta_r(x_1), \ldots, \theta_r(x_M))$ and $G_{r,\mathbf{x}} \sim \mathrm{DP}(\alpha_r, G_{r0}(\mathbf{x}))$, $r = 1, 2$, with $G_{r0}(\mathbf{x})$ the M-variate normal distribution given above. We note that, in practice, learning with DDP priors requires some form of replication in the response values, e.g. more than one response value for each x_m, $m = 1, \ldots, M$. Assume that $\mathbf{y}_i = (y_{i1}, \ldots, y_{iM})$, $i = 1, \ldots, N$, is the ith response replicate. Customary data augmentation methods can be used to fit the model when *complete* replicates are not available, i.e. when some of the y_{im} are missing.

Let $\boldsymbol{\theta}_{ri} = (\theta_{ri}(x_1), \ldots, \theta_{ri}(x_M))$, $r = 1, 2$, be the latent mixing vectors associated with $\mathbf{y}_i$, and $f_p(\mathbf{y}_i; \mathbf{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}) = \prod_{m=1}^{M} k_p(y_{im} - (\beta_0 + \beta_1 x_m); \theta_{1i}(x_m), \theta_{2i}(x_m))$. Then, the hierarchical quantile regression model induced by the DDP prior is given by

$$\begin{aligned}
&\mathbf{Y}_i \mid (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i} \overset{\text{ind.}}{\sim} f_p(\mathbf{y}_i; \mathbf{x}, (\beta_0, \beta_1), \boldsymbol{\theta}_{1i}, \boldsymbol{\theta}_{2i}), \quad i = 1, \ldots, N \\
&\boldsymbol{\theta}_{ri} \mid G_{r,\mathbf{x}} \overset{\text{i.i.d.}}{\sim} G_{r,\mathbf{x}}, \quad r = 1, 2, \ \ i = 1, \ldots, N \\
&G_{r,\mathbf{x}} \mid \alpha_r, \mu_r, \tau_r^2, \phi_r \sim \mathrm{DP}(\alpha_r, G_{r0}(\mathbf{x}) = N_M(\mu_r 1_M, V_r)), \quad r = 1, 2,
\end{aligned} \tag{14}$$

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 2.4. 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 (14) is a DP mixture model with the M-variate DP priors for $G_{r,\mathbf{x}}$ induced by the DDP priors for $G_{r,\mathcal{X}}$. Hence, again, posterior sampling proceeds by marginalizing $G_{r,\mathbf{x}}$ over their DP priors, and utilizing an MCMC method for DP mixtures (appendix A discusses

details). Regarding posterior predictive inference, interest lies in the posterior predictive density at observed covariate values in $\mathbf{x}$ as well as at new (unobserved) values, say $\tilde{\mathbf{x}}=(\tilde{x}_1,\ldots,\tilde{x}_U)$. Hence, we seek $p(\mathbf{y}_0,\tilde{\mathbf{y}}_0\,|\,\tilde{\mathbf{x}},\text{data})$, where $\mathbf{y}_0=(y_{01},\ldots,y_{0M})$ and $\tilde{\mathbf{y}}_0=(\tilde{y}_{01},\ldots,\tilde{y}_{0U})$ are associated with $\mathbf{x}$ and $\tilde{\mathbf{x}}$, respectively, and $\text{data}=\{\mathbf{x},\mathbf{y}_1,\ldots,\mathbf{y}_N\}$. Consider, for $r=1,2$, the augmented set of latent mixing vectors $\{(\boldsymbol{\theta}_{ri},\tilde{\boldsymbol{\theta}}_{ri}):i=1,\ldots,N\}$, where $\tilde{\boldsymbol{\theta}}_{ri}=(\theta_{ri}(\tilde{x}_1),\ldots,\theta_{ri}(\tilde{x}_U))$, corresponding to the augmented response data vector $\{(\mathbf{y}_i,\tilde{\mathbf{y}}_i):i=1,\ldots,N\}$, which includes the unobserved portion $\tilde{\mathbf{y}}_i=(\tilde{y}_{i1},\ldots,\tilde{y}_{iU})$ associated with $\tilde{\mathbf{x}}$. For $r=1,2$, let N_r^* be the number of clusters in the $\{(\boldsymbol{\theta}_{ri},\tilde{\boldsymbol{\theta}}_{ri}):i=1,\ldots,N\}$, denote by $(\boldsymbol{\theta}_{rj}^*,\tilde{\boldsymbol{\theta}}_{rj}^*)$, $j=1,\ldots,N_r^*$, the cluster locations, and let N_{rj} be the jth cluster size. Again, the clusters are determined by a vector of configuration indicators $\mathbf{s}_r=(s_{r1},\ldots,s_{rN})$, with $s_{ri}=j$ if and only if $(\boldsymbol{\theta}_{ri},\tilde{\boldsymbol{\theta}}_{ri})=(\boldsymbol{\theta}_{rj}^*,\tilde{\boldsymbol{\theta}}_{rj}^*)$, $i=1,\ldots,N$. Let $\boldsymbol{\eta}$ collect all parameters corresponding to the augmented data vector, i.e. $\boldsymbol{\eta}=\{\beta_0,\beta_1,N_r^*,\mathbf{s}_r,((\boldsymbol{\theta}_{rj}^*,\tilde{\boldsymbol{\theta}}_{rj}^*):j=1,\ldots,N_r^*),\alpha_r,\mu_r,\tau_r^2,\phi_r:r=1,2\}$. Note that $\boldsymbol{\eta}=\{\boldsymbol{\xi},(\tilde{\boldsymbol{\theta}}_{rj}^*:j=1,\ldots,N_r^*):r=1,2\}$, where $\boldsymbol{\xi}$ is the parameter vector for model (14), resulting after marginalizing $G_{r,\mathbf{x}}$, $r=1,2$. Thus, the MCMC algorithm (discussed in appendix A) 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, given the partition implied by $\mathbf{s}_r$, the $(\boldsymbol{\theta}_{rj}^*,\tilde{\boldsymbol{\theta}}_{rj}^*)$ are i.i.d. from $G_{r0}(\mathbf{x},\tilde{\mathbf{x}})$, 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 $\mathbf{x}$ and $\tilde{\mathbf{x}}$. 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 $\tau_r^2\exp(-\phi_r|\tilde{x}_u-\tilde{x}_{u'}|)$, $u,u'=1,\ldots,U$, and the $M\times U$ cross-diagonal block has elements $\tau_r^2\exp(-\phi_r|x_m-\tilde{x}_u|)$, $m=1,\ldots,M,u=1,\ldots,U$. Based on this DP property, we can express the posterior $p(\boldsymbol{\eta}\,|\,\tilde{\mathbf{x}},\text{data})$ as $p(\boldsymbol{\xi}\,|\,\text{data})\prod_{r=1}^2\prod_{j=1}^{N_r^*}p(\tilde{\boldsymbol{\theta}}_{rj}^*\,|\,\boldsymbol{\theta}_{rj}^*,\mu_r,\tau_r^2,\phi_r)$, where $p(\tilde{\boldsymbol{\theta}}_{rj}^*\,|\,\boldsymbol{\theta}_{rj}^*,\mu_r,\tau_r^2,\phi_r)$ is the conditional U-variate normal density for $\tilde{\boldsymbol{\theta}}_{rj}^*$ given $\boldsymbol{\theta}_{rj}^*$ under the $G_{r0}(\mathbf{x},\tilde{\mathbf{x}})$ distribution for $(\boldsymbol{\theta}_{rj}^*,\tilde{\boldsymbol{\theta}}_{rj}^*)$. Now, the posterior predictive density is given by

$$p(\mathbf{y}_0,\tilde{\mathbf{y}}_0\,|\,\tilde{\mathbf{x}},\text{data})=\int f_p(\mathbf{y}_0;\mathbf{x},(\beta_0,\beta_1),\boldsymbol{\theta}_{10},\boldsymbol{\theta}_{20})f_p(\tilde{\mathbf{y}}_0;\tilde{\mathbf{x}},(\beta_0,\beta_1),\tilde{\boldsymbol{\theta}}_{10},\tilde{\boldsymbol{\theta}}_{20})\times p(\boldsymbol{\theta}_{10},\tilde{\boldsymbol{\theta}}_{10}\,|\,\boldsymbol{\eta})p(\boldsymbol{\theta}_{20},\tilde{\boldsymbol{\theta}}_{20}\,|\,\boldsymbol{\eta})p(\boldsymbol{\eta}\,|\,\tilde{\mathbf{x}},\text{data}),$$

where, for $r=1,2$,

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

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

4.2. An example with simulated data

The practical utility of the proposed modelling framework is studied in more detail in the technical report version of the paper (Kottas & Krnjajić, 2005) through the analysis of real data from a genotoxicity experiment. Here, we discuss results from a simulation experiment, which was designed to illustrate the predictive performance of the DDP quantile regression model.

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, $x_m=0,5,20,50,100$, with $\beta_0=0$ and $\beta_1=0.45$. The errors ε_{im} were generated from a median-zero split normal distribution, $0.5\text{TN}(\varepsilon\,|\,0,\varphi\sigma^2;\varepsilon\in(-\infty,0))+0.5\text{TN}(\varepsilon\,|\,0,\varphi^{-1}\sigma^2;\varepsilon\in[0,\infty))$, where $\text{TN}(\cdot\,|\,\mu,\tau^2;\cdot\in A)$ denotes an $N(\cdot\,|\,\mu,\tau^2)$ 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 (14), with $p = 0.5$, to the simulated 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 Fig. 6. It also enables learning at new x values

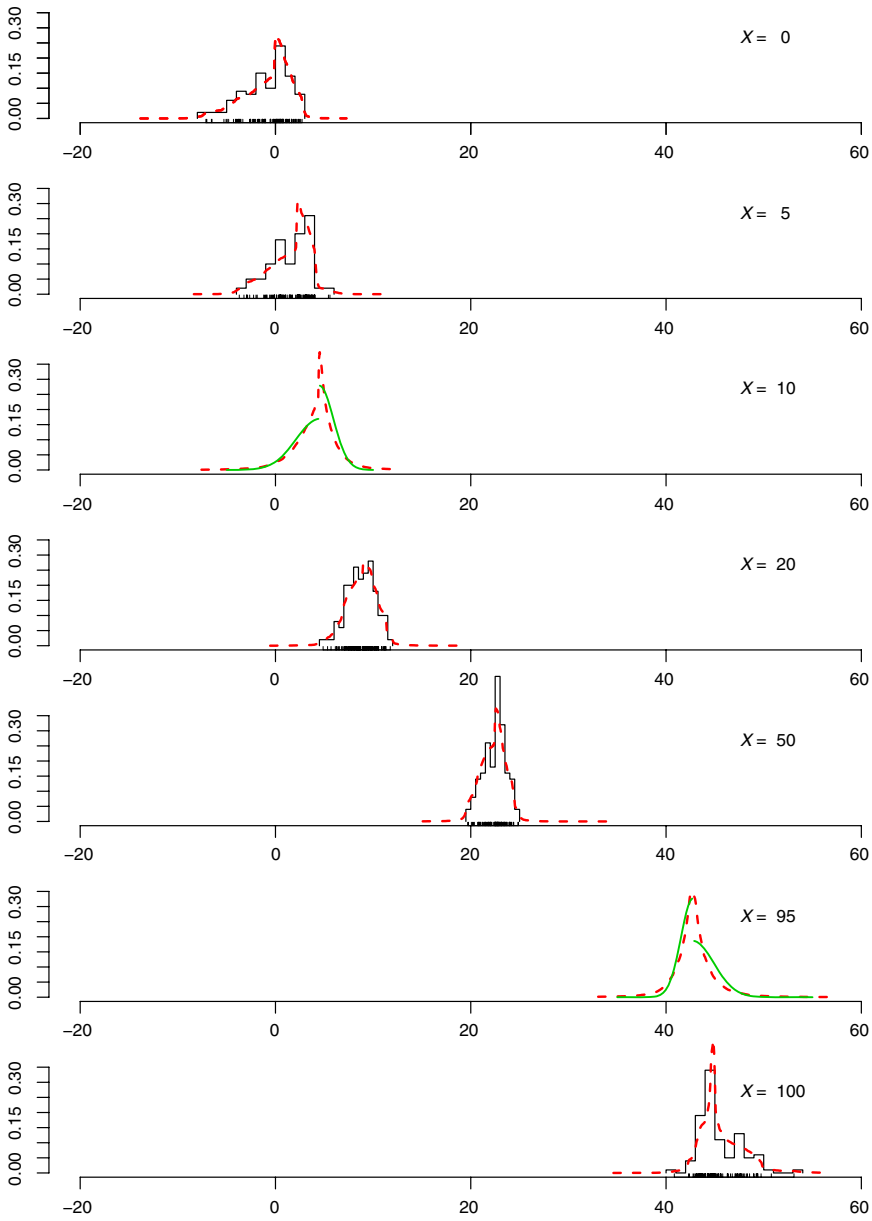

Fig. 6. Simulation example for the DDP quantile regression model. Posterior predictive densites under the DDP model (dashed lines) at the five observed covariate values (overlaid on histograms of the corresponding response observations), and at two new covariate values, $x = 10$ and $x = 95$ (overlaid on corresponding true densities denoted by solid lines).

where responses are not observed (Fig. 6 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 (14) 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 (14) 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.

5. Discussion

We have developed modelling 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. The prior probability model for the errors is defined through scale mixtures of uniform densities with DP priors for the mixing distributions. We have demonstrated the superiority of this class of non-parametric mixture models over certain parametric models as well as over a class of non-parametric 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.

We conclude with a brief discussion of an extension of the proposed modelling approach and an alternative framework for quantile regression. (Some early results on both of these research directions are reported in Kottas *et al.*, 2007.) As mentioned in section 2.1, a practically important extension involves non-parametric modelling for the quantile regression function $h(\cdot)$ in (1). In particular, a natural starting point for such modelling is provided by the additive non-parametric regression setting, under which $h(\mathbf{x})$ is expressed as a sum of covariate-specific regression functions $h_k(x_k)$, where $\mathbf{x}=\{x_k : k=1,\ldots,K\}$. In the presence of a relatively small number of covariates, Gaussian process priors for the $h_k(\cdot)$ provide a flexible means of completing the model specification (see, e.g. Neal, 1998, on Gaussian process regression under parametric error distributions). The combination of model-based non-parametric approaches for $h(\cdot)$ and the error distribution $f_p(\cdot)$ in (1) yields a class of models that allow the data to uncover non-linearities in the quantile regression function as well as non-standard distributional features in the errors. However, it requires careful study of the extent of identifiability issues as well as efficient MCMC algorithms for posterior simulation.

A different approach to fully non-parametric quantile regression can be developed by modelling the joint distribution of the response and a set of (continuous) covariates with a DP mixture (say, in the spirit of Müller *et al.*, 1996) and then obtaining the induced inference

for different quantile curves (using, say, the approach of Gelfand & Kottas, 2002). This framework can be extended to also handle categorical covariates and/or censored observations. (See Scaccia & Green, 2003, for related work where the conditional distribution of the response given a single continuous covariate is modelled with a discrete normal mixture with covariate-dependent weights.) Evidently, this approach does not correspond to the additive quantile regression setting in (1); moreover, it becomes computationally intensive (prohibitive) with a moderate (large) number of covariates. However, it offers a unifying framework for flexible non-linear, non-parametric inference for any number of quantile regression functions. Full details on these lines of research will be reported in a future article.

Acknowledgements

The authors wish to thank two referees for helpful comments. They also thank Keming Yu for providing the data set analysed in section 3.2. The work of the first author was supported in part by NSF grant DMS-0505085.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

Bernardo, J. M. & Smith, A. F. M. (2000). *Bayesian theory*. Wiley, Chichester.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1**, 353–355.

Brunner, L. J. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statist. Probab. Lett.* **14**, 195–199.

Brunner, L. J. (1995). Bayesian linear regression with error terms that have symmetric unimodal densities. *J. Nonparametr. Statist.* **4**, 335–348.

Brunner, L. J. & Lo, A. Y. (1989). Bayes methods for a symmetric unimodal density and its mode. *Ann. Statist.* **17**, 1550–1566.

Bush, C. A. & MacEachern, S. N. (1996). A semiparametric Bayesian model for randomised block designs. *Biometrika* **83**, 275–285.

De Iorio, M., Müller, P., Rosner, G. L. & MacEachern, S. N. (2004). An ANOVA model for dependent random measures. *J. Amer. Statist. Assoc.* **99**, 205–215.

Dunson, D. B. & Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *J. Nonparametr. Statist.* **17**, 385–400.

Escobar, M. D. & West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588.

Feller, W. (1971). *An introduction to probability theory and its applications*, Vol. II, 2nd edn. Wiley, New York.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

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

Gelfand, A. E. & Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *J. Comput. Graph. Statist.* **11**, 289–305.

Gelfand, A. E. & Kottas, A. (2003). Bayesian semiparametric regression for median residual life. *Scand. J. Statist.* **30**, 651–665.

Gelfand, A. E., Kottas, A. & MacEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021–1035.

Hansen, M. B. & Lauritzen, S. L. (2002). Nonparametric Bayes inference for concave distribution functions. *Statist. Neerlandica* **56**, 110–127.

Hanson, T. & Johnson, W. O. (2002). Modeling regression error with a mixture of Pólya trees. *J. Amer. Statist. Assoc.* **97**, 1020–1033.

Hjort, N. L. & Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In *Advances in statistical modeling and inference: essays in Honor of Kjell A. Doksum* (ed. V. Nair), 463–492. World Scientific Publishing Company, Singapore.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). *Bayesian survival analysis*. Springer, New York.
Koenker, R. (2005). *Quantile regression*. Cambridge University Press, New York.
Koenker, R. & Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *J. Amer. Statist. Assoc.* **96**, 458–468.
Kottas, A. & Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *J. Amer. Statist. Assoc.* **96**, 1458–1468.
Kottas, A. & Krnjajić, M. (2005). Bayesian nonparametric modeling in quantile regression. Technical report AMS 2005-06, University of California, Santa Cruz (available at: http://www.ams.ucsc.edu/reports).
Kottas, A., Krnjajić, M. & Taddy, M. (2007). Model-based approaches to nonparametric Bayesian quantile regression. In *Proceedings of the 2007 Joint Statistical Meetings*, American Statistical Association, pp. 1137–1148.
Lavine, M. & Mockus, A. (1995). A nonparametric Bayes method for isotonic regression. *J. Statist. Plann. Inference* **46**, 235–248.
MacEachern, S. N. (1999). Dependent nonparametric processes. In *Proceedings of the 1999 Joint Statistical Meetings*, American Statistical Association, pp. 50–55.
MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical Report, Department of Statistics, The Ohio State University.
Müller, P., Erkanli, A. & West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67–79.
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, J. O. Berger, A. P. Dawid & A. F. M. Smith), 475–501. Oxford University Press.
Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9**, 249–265.
Quintana, F. A. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *J. Roy. Statist. Soc. Ser. B Statist. Methodol.* **65**, 557–574.
Royston, P. & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Appl. Statist.* **43**, 429–467.
Scaccia, L. & Green, P. J. (2003). Bayesian growth curves using normal mixtures with nonparametric weights. *J. Comput. Graph. Statist.* **12**, 308–331.
Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4**, 639–650.
Tsionas, E. G. (2003). Bayesian quantile inference. *J. Statist. Comput. Simul.* **73**, 659–674.
Walker, S. G. & 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. *J. Amer. Statist. Assoc.* **94**, 137–145.
Ying, Z., Jung, S. H. & Wei, L. J. (1995). Survival analysis with median regression models. *J. Amer. Statist. Assoc.* **90**, 178–184.
Yu, K. & Moyeed, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54**, 437–447.
Yu, K., Lu, Z. & Stander, J. (2003). Quantile regression: applications and current research areas. *The Statistician* **52**, 331–350.

Received June 2006, in final form June 2008

Athanasios Kottas, Department of Applied Mathematics and Statistics, Baskin School of Engineering, MS: SOE2, 1156 High Street, University of California, Santa Cruz, CA 95064, USA.
E-mail: thanos@ams.ucsc.edu

Appendix A

We present here some of the details for MCMC posterior simulation for models $\mathcal{M}_1$ and $\mathcal{M}_2$, discussed in section 2.2 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 2.3 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

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

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

The full conditional for each σ_i is a mixed distribution with point masses at $\sigma_j^{*-}, j=1,\ldots,n^{*-}$, and continuous component given by an $\text{IG}(c+1,d+\gamma_i)$ distribution, where

$$\gamma_i=(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta})\{p1_{(0,\infty)}(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta})+(p-1)1_{(-\infty,0]}(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta})\}.$$

The weights for the point masses are proportional to $n_j^- k_p^{\text{AL}}(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta};\sigma_j^{*-})$. The weight for the continuous component is proportional to $\alpha\int k_p^{\text{AL}}(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta};\sigma)g_0(\sigma)\,\mathrm{d}\sigma=\alpha c d^c p(1-p)(d+\gamma_i)^{-c-1}$ (here, g_0 is the density of G_0). To resample the cluster locations once all the σ_i are updated, we use the approach from Bush & MacEachern (1996). Specifically, the full conditional for each σ_j^* is proportional to $g_0(\sigma_j^*)\prod_{\{i:s_i=j\}}k_p^{\text{AL}}(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta};\sigma_j^*)$, resulting in an $\text{IG}(c+n_j,d+\sum_{\{i:s_i=j\}}\gamma_i)$ distribution.

We sample $\boldsymbol{\beta}$ using a random-walk Metropolis step. We update α using the method of Escobar & 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

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

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

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

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

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

We next resample the cluster locations $\sigma_{rj}^*, j=1,\ldots,n_r^*, r=1,2$. The full conditional for each σ_{1j}^* is proportional to $g_{10}(\sigma_{1j}^*)\prod_{\{i:s_{1i}=j\}}k_p(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta};\sigma_{1j}^*,\sigma_{2i})$, where g_{10} is the density associated with G_{10}. Letting $\mathcal{A}_1=\{i:s_{1i}=j,y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta}<0\}$, and $L_1=|\mathcal{A}_1|$, the full conditional is an $\text{IG}(c_1+L_1,d_1)$ distribution truncated below by $\max_{i\in\mathcal{A}_1}\{-(y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta})\}$. 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 $\max_{i\in\mathcal{A}_2}\{y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta}\}$, where $\mathcal{A}_2=\{i:s_{2i}=j,y_i-\mathbf{x}_i^{\text{T}}\boldsymbol{\beta}\geq 0\}$, and $L_2=|\mathcal{A}_2|$.

We update α_1 and α_2 using the approach of Escobar & 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 & Gelfand (2001).

DDP quantile regression model

We use algorithm 6 from Neal (2000) to update the $\boldsymbol{\theta}_{ri}$, $i=1,\ldots,N$, $r=1,2$. The regression coefficients (β_0,β_1) are updated with random-walk Metropolis steps. As with the other models, the precision parameters α_r are sampled as in Escobar & 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.

Appendix B

Here, we provide a result on posterior propriety under flat priors for the quantile regression coefficients. In particular, we consider the special case of model $\mathcal{M}_2$ that includes only an intercept, i.e. with the first stage of (7) reducing to $k_p(y_i-\beta_0;\sigma_{1i},\sigma_{2i})$, $i=1,\ldots,n$. Moreover, for a simpler exposition, we work with fixed DP prior hyperparameters α_r and d_r, $r=1,2$. Hence, the posterior that results after integrating G_1 and G_2 over their DP priors has normalizing constant

$$B=\int_{\beta_0}\int_{\boldsymbol{\sigma}_1}\int_{\boldsymbol{\sigma}_2}[\boldsymbol{\sigma}_1][\boldsymbol{\sigma}_2]\prod_{i=1}^{n}k_p(y_i-\beta_0;\sigma_{1i},\sigma_{2i})\,\mathrm{d}\boldsymbol{\sigma}_1\,\mathrm{d}\boldsymbol{\sigma}_2\,\mathrm{d}\beta_0,$$

where $\boldsymbol{\sigma}_r=(\sigma_{ri}:i=1,\ldots,n)$, $r=1,2$. We need to show that $0<B<\infty$. Note that, with a single observation, $\boldsymbol{\sigma}_r\equiv\sigma_{r1}\sim G_{r0}$, $r=1,2$, and hence the result is immediate when $n=1$.

Key to the proof is the Pólya urn structure of the joint priors $[\boldsymbol{\sigma}_r]$, $r=1,2$. One possible expression for $[\boldsymbol{\sigma}_r]$ involves a sum over all partitions of $\{1,\ldots,n\}$; the distribution for each sum component is driven by σ^*_{rj} i.i.d. from G_{r0} for $j=1,\ldots,n^*_r$, where n^*_r is the number of distinct elements of $\boldsymbol{\sigma}_r$ corresponding to the specific partition (see, e.g. Quintana & Iglesias, 2003). Hence, a lower bound for B emerges by considering the specific partition with only one distinct component, i.e. with $n^*_1=n^*_2=1$ and, say, $\sigma_{1i}=\sigma^*_1$ and $\sigma_{2i}=\sigma^*_2$, for each $i=1,\ldots,n$. Specifically,

$$B\geq E\int_{\beta_0}\int_{\sigma^*_1}\int_{\sigma^*_2}g_{10}(\sigma^*_1)g_{20}(\sigma^*_2)\prod_{i=1}^{n}k_p(y_i-\beta_0;\sigma^*_1,\sigma^*_2)\,\mathrm{d}\sigma^*_1\,\mathrm{d}\sigma^*_2\,\mathrm{d}\beta_0,$$

where E is a positive finite constant (a function of α_1 and α_2). Using the form of $k_p(\cdot)$, we obtain

$$\prod_{i=1}^{n}k_p(y_i-\beta_0;\sigma^*_1,\sigma^*_2)\geq\sum_{\ell=1}^{n-1}p^{\ell}(1-p)^{n-\ell}\sigma_1^{*-\ell}\ \sigma_2^{*\ell-n}1_{(\tilde{y}_\ell,\tilde{y}_{\ell+1}]}(\beta_0),$$

where the $\tilde{y}_\ell$ denote the ordered observations. Therefore,

$$B\geq E\int_{\sigma^*_1}\int_{\sigma^*_2}g_{10}(\sigma^*_1)g_{20}(\sigma^*_2)\sum_{\ell=1}^{n-1}p^{\ell}(1-p)^{n-\ell}\sigma_1^{*-\ell}\ \sigma_2^{*\ell-n}(\tilde{y}_{\ell+1}-\tilde{y}_\ell)\,\mathrm{d}\sigma^*_1\,\mathrm{d}\sigma^*_2,$$

and thus $B>0$, since the integral above can be readily shown to be strictly positive.

To prove $B<\infty$, first, note that

$$\int_{\beta_0}\prod_{i=1}^{n}k_p(y_i-\beta_0;\sigma_{1i},\sigma_{2i})\,\mathrm{d}\beta_0\le\prod_{i=1}^{n}\left\{\int_{\beta_0}(k_p(y_i-\beta_0;\sigma_{1i},\sigma_{2i}))^n\,\mathrm{d}\beta_0\right\}^{1/n}$$
$$\le\prod_{i=1}^{n}\left(\left\{\int_{\beta_0}(p\sigma_{1i}^{-1}1_{(-\sigma_{1i},0)}(y_i-\beta_0))^n\,\mathrm{d}\beta_0\right\}^{1/n}\right.$$
$$\left.+\left\{\int_{\beta_0}((1-p)\sigma_{2i}^{-1}1_{[0,\sigma_{2i})}(y_i-\beta_0))^n\,\mathrm{d}\beta_0\right\}^{1/n}\right).$$

The first bound results from the standard inequality, $\int\prod_{i=1}^{n}f_i^{a_i}(x)\,\mathrm{d}x\le\prod_{i=1}^{n}\{\int f_i(x)\,\mathrm{d}x\}^{a_i}$, for non-negative integrable functions f_i and constants $a_i\ge0$ with $\sum_{i=1}^{n}a_i=1$. The second bound is obtained by applying Minkowski's inequality to each term in the product resulting from the first bound. Now evaluating the integrals, we obtain $\int_{\beta_0}\prod_{i=1}^{n}k_p(y_i-\beta_0;\sigma_{1i},\sigma_{2i})\,\mathrm{d}\beta_0\le\prod_{i=1}^{n}(p\sigma_{1i}^{-(n-1)/n}+(1-p)\sigma_{2i}^{-(n-1)/n})$, and therefore

$$B\le\Delta_n=\int_{\boldsymbol{\sigma}_1}\int_{\boldsymbol{\sigma}_2}[\boldsymbol{\sigma}_1][\boldsymbol{\sigma}_2]\prod_{i=1}^{n}\left(p\sigma_{1i}^{-(n-1)/n}+(1-p)\sigma_{2i}^{-(n-1)/n}\right)\,\mathrm{d}\boldsymbol{\sigma}_1\,\mathrm{d}\boldsymbol{\sigma}_2$$

for any $n\ge2$. For $n=2$ it is easy to evaluate Δ_2 and verify that $\Delta_2<\infty$. A substantially more lengthy, albeit straightforward, induction argument establishes that $\Delta_n<\infty$ for any n. The approach uses the representation of $[\boldsymbol{\sigma}_r]$ in terms of successive complete conditionals, with $\sigma_{r1}\sim G_{r0}$, and for each $i=2,\ldots,n$, $[\sigma_{ri}\mid\sigma_{r1},\ldots,\sigma_{r,i-1}]$ given by a mixed distribution with point masses $(\alpha_r+i-1)^{-1}$ at the σ_{rj}, for $j=1,\ldots,i-1$, and continuous mass $\alpha_r(\alpha_r+i-1)^{-1}$ on G_{r0}. Hence, all the integrals that are involved in the evaluation of Δ_n reduce to integrals of the form $\int_0^\infty\sigma_r^{-q}g_{r0}(\sigma_r)\,\mathrm{d}\sigma_r$, for $q>0$, which are finite, since G_{r0}, $r=1,2$, are inverse Gamma distributions.