A Partially Collapsed Gibbs sampler for Bayesian quantile regression

Craig Reed and Keming Yu*

Abstract

We introduce a set of new Gibbs sampler for Bayesian analysis of quantile regression model. The new algorithm, which partially collapsing an ordinary Gibbs sampler, is called Partially Collapsed Gibbs (PCG) sampler. Although the Metropolis-Hastings algorithm has been employed in Bayesian quantile regression, including median regression, PCG has superior convergence properties to an ordinary Gibbs sampler. Moreover, Our PCG sampler algorithm, which is based on a theoretic derivation of an asymmetric Laplace as scale mixtures of normal distributions, requires less computation than the ordinary Gibbs sampler and can significantly reduce the computation involved in approximating the Bayes Factor and marginal likelihood. Like the ordinary Gibbs sampler, the PCG sample can also be used to calculate any associated marginal and predictive distributions. The quantile regression PCG sampler is illustrated by analysing simulated data and the data of length of stay in hospital. The latter provides new insight into hospital performance. C-code along with an R interface for our algorithms is publicly available on request from the first author.

JEL classification: C11, C14, C21, C31, C52, C53.

Keywords: Bayesian inference; Gibbs sampler; Partially collapsed Gibbs sampler; Quantile regression.

1 Introduction

Quantile regression is used when an estimate of the various quantiles (such as the median) of a conditional distribution is desired. Quantile regression can be

*corresponding author: Keming Yu, Department of Mathematical Sciences, Brunel University, Uxbridge, UB8 3PH, UK, tel: 44-1895-266128, fax: 44-1895-269732, email: keming.yu@brunel.ac.uk

seen as a natural analogue in regression analysis to the practice of using different measures of central tendency and statistical dispersion to obtain a more comprehensive and robust analysis (Koenker, 2005). Asymmetry as well as long tails (which means very extreme outcomes from a distribution have non-negligible probabilities), are common to not only in economics but also a number of other disciplines such as social sciences, medicine, public health, financial return, environment and engineering. For example, Levin (2001) studies a panel survey of the performance of Dutch school children and finds some evidence of positive peer effects in the lower tail of the achievement distribution. Many asymmetric and long-tailed distributions have been used to model the innovation in autoregressive conditional heteroscedasticity (ARCH) models in finance. In particular, the conditional Autoregressive Value at Risk (CAViaR) model introduced by Engle and Manganelli (2004) is a very popular time series model for estimating the Value at Risk in finance. Based on simulations, Min and Kim (2004) claim that over a wide-class of non Gaussian errors, with asymmetric and long-tailed distributions, simple mean regression cannot satisfactorily capture the key properties of the data; even the conditional mean estimation can be misleading. The need for and success of quantile regression in ecology has been attributed to the complexity of interactions between different factors leading to data with unequal variation of one variable for different ranges of another variable (Cade and Noon, 2003). In the study of maternal and child health and occupational and environmental risk factors, Abrevaya (2001) investigates the impact of various demographic characteristics and maternal behaviour on the birthweight of infants born in the U.S. Low birthweight is known to be associated with a wide range of subsequent health problems and developmental markers. Chamberlain (1994) infers that for manufacturing workers, the union wage premium, which is at 28 percent at the first decile, declines continuously to 0.3 percent at the upper decile. The author suggests that the location shift model estimate (least squares estimate) which is 15.8 percent, gives a misleading impression of the union effect. In fact, this mean union premium of 15.8 percent is captured primarily by the lower tail of the conditional distribution.

These examples demonstrate that the quantile regression approach is more appropriate when the underlying model is nonlinear, when the error term follows a non-normal distribution or when the tails of underlying distributions are of interest for modelling extreme behaviour. For more details we refer the interested reader to Koenker and Hallock (2004) and Yu *et al.* (2003).

Bayesian inference on quantile regression has attracted much interest recently. A few of the different models and sampling algorithms for Bayesian quantile regression include MCMC (Markov chain Monte Carlo) or RJMCMC (Reversible Jump Markov Chain Monte Carlo) methods via an asymmetric Laplace distribution for the likelihood function (Yu and Moyeed, 2001; Yu and Stander, 2007; Chen and Yu, 2008; Tsionas, 2003), Dirichlet process mixing based nonparametric median zero distribution for the regression model error (Kottas and Gelfand, 2001), an MCMC algorithm using Jeffreys' (Jeffreys, 1961) substitution posterior

for the median (Dunson and Taylor, 2005), the expectation-maximising (EM) algorithm using the asymmetric Laplace distribution (Geraci and Bottai, 2007), an empirical likelihood based algorithm (Lancaster and Jun, 2008) and density-based quantile curve estimation (Dunson, 2008).

The ordinary Gibbs sampler (introduced in the context of image processing by Geman and Geman (1984)), is a special case of Metropolis-Hastings sampling wherein the random value is always accepted. The task remains to specify how to construct a Markov Chain whose values converge to the target distribution. The key to the Gibbs sampler is that only univariate conditional distributions are considered - the distribution when all of the random variables except one are assigned fixed values. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms (often being normals, inverse gamma, or other common distributions). Thus we can simulate m random variables sequentially from the m conditionals rather than generating a single m-dimensional vector in a single pass using the full joint distribution. The PCG sampler, like blocking, takes this one step further in that some conditionals may be "reduced" in the sense that they are conditional on fewer variables being fixed. Such samplers tend to converge more quickly to the target distribution (van Dyk and Park, 2008). To develop Gibbs sampler for quantile regression we need to develop proper distribution theory. In this paper, via a theoretic derivation of an AL as a scale mixtures of normal distributions and by augmenting the data, we first propose a Partially Collapsed Gibbs (PCG) sampler which converges to the joint posterior distribution of all unknown parameters and latent variables, then extend the approach to nonparametric Bayes. We prefer PCG to an ordinary Gibbs sampler for Bayesian quantile regression is due to soem key advantages of our approach. Firstly, all involved conditional distributions are simple distributions and are easy to sample from. This reduces the computation involved and allows us to use Chib's method (Chib, 1995) to approximate the Bayes Factor (see Section 2.5). Secondly, as a consequence of using the PCG sampler, we reduce the number of steps required to calculate marginal likelihood even further and the sampler will have superior convergence. These properties are particularly appealing when several quantile regressions are required at one time.

The paper is organized as follows. In Section 2, we first derive a lemma of an asymmetric Laplace as scale mixtures of normal distributions, then construct a PCG sampler for quantile regression, including the details on calculating and summarising the appropriate marginal distributions, predictive distributions and then approximating the Bayes Factor as well as semiparametric extension. In Section 3, we illustrate the PCG sampler by analysing first simulated data and then a real dataset used to explore hospital performance via the length of stay of patients. Section 4 concludes.

2 Partially Collapsed Gibbs Sampler for Quantile Regression

2.1 Preliminaries

Consider the regression model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \tau^{-1}\epsilon_i,\ i = 1, \cdots, n. \tag{1}$$

where y_i denotes the ith observation, $\boldsymbol{x}_i = [1\ x_{i1}\ x_{i2}\ \cdots\ x_{ik}]^T$ represent the $k+1$ known covariates associated with subject i, $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ \cdots\ \beta_k]^T$ are the $k+1$ unknown parameters, τ is the inverse scale parameter, and ϵ_i, $i = 1, \cdots, n$ are identically and independently distributed (i.i.d.) error terms. The distribution of the error is assumed to exist, but it is assumed unknown. This model in matrix form is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tau^{-1}\boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is a vector of observations y_i, $\boldsymbol{\epsilon}$ is a vector of error terms ϵ_i and $\boldsymbol{X}$ is an $n \times (k+1)$ design matrix $\boldsymbol{X} = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots\ \boldsymbol{x}_n]^T$.

Suppose that all conditional quantiles $Q_p(\boldsymbol{y}|\boldsymbol{X})$ for the regression model (1) are given by $Q_p(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}(p)$ for all $p \in (0,1)$, then classic quantile regression theory (Koenker, 2005) seeks estimating $\boldsymbol{\beta(p)}$ by minimising $\sum_{i=1}^n \rho_p(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})$, where

$$\rho_p(u) := \begin{cases} p\,|u| & \text{if} \quad u \geq 0 \\ (1-p)|u| & \text{if} \quad u < 0, \end{cases} \tag{2}$$

Under Bayesian quantile regression, we follow Yu and Moyeed (2001) considering the likelihood based on the asymmetric Laplace (AL) distribution and then show how to construct a PCG sampler. The probability density function (pdf) of the AL distribution with location parameter μ, inverse scale parameter $\tau \in (0, \infty)$ and skewness parameter $p \in (0,1)$, is given by

$$f(x|\mu, \tau, p) = \tau p(1-p)\exp(-\tau\rho_p(x-\mu)), \tag{3}$$

where $\rho_p(u)$ is defined in (2). If X has the AL distribution, we denote it as $X \sim AL(\mu, \tau, p)$.

The following lemma extend the result of a asymmetric Laplace as scale mixtures of normal distributions (Park and Casella, 2008) to allow us to simulate draws from the AL distribution by first drawing from an exponential distribution then a normal distribution.

Lemma

Let $X \sim AL(\mu, \tau, p)$, Z be a standard normal random variable and ξ be an exponential random variable. Then

$$X =^d \sqrt{\frac{2\xi}{\tau p(1-p)}}Z + \frac{1-2p}{p(1-p)}\xi + \mu, \tag{4}$$

where $=^d$ denotes equality in distribution. This reduces to the well known scale mixture of normals representation of the symmetric Laplace distribution if $p = 0.5$.

The details of proof appear in Appendix. This lemma is also useful in the development of Bayesian variable selection in for quantile regression (Dunson, Reed and Yu, 2009), and Bayesian skewed Lesso for high-dimensional predictors (Yu, Dunson and Reed, 2009).

2.2 Deriving the PCG sampler for Bayesian quantile regression

Under the quantile regression model in Yu and Moyeed(2001), and allowing for an inverse scale parameter τ, the likelihood $l(\boldsymbol{y}|\boldsymbol{\beta}, \tau)$ for a fixed p is proportional to

$$\tau^n \exp\left(-\tau \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})\right), \tag{5}$$

where here and for the rest of this section, we suppress dependecne on $\boldsymbol{X}$ and p to make notation clearer.

With the likelihood specified, all that is needed for Bayesian inference is a prior on the unknown parameters $\boldsymbol{\beta}$ and τ, which may or may not be specific to the particular value of p. We will choose the following priors for simplicity, independent of the value of p:

$$\tau \sim \Gamma(c_0, d_0), \tag{6}$$

$$\boldsymbol{\beta}|\tau \sim N_k(\boldsymbol{b}_0, \boldsymbol{B}_0), \tag{7}$$

with $c_0, d_0, \boldsymbol{b}_0, \boldsymbol{B}_0$ known. Typically a vague prior will be used on τ because it is regarded as a nuisance parameter. We can get an improper prior by setting $\boldsymbol{b}_0 = \boldsymbol{0}, \boldsymbol{B}_0 = C\boldsymbol{I}$, $C \to \infty$, and $c_0 = d_0 = 0$ as this gives

$$\pi(\boldsymbol{\beta}, \tau) \propto \tau^{-1}.$$

To enable the development of a PCG sampler, we now augment the data with the latent random weights $w_i, i = 1, \ldots n$. We suppose that the full likelihood for a particular observation y_i conditional on $\boldsymbol{\beta}, \tau$ **and** the latent weight w_i is

$$y_i|w_i, \boldsymbol{\beta}, \tau \sim N\left(\frac{1-2p}{p(1-p)}w_i + \boldsymbol{x_i}^T\boldsymbol{\beta}, \frac{2w_i}{\tau p(1-p)}\right),$$

and each weight w_i conditional on $\boldsymbol{\beta}$ and τ is independently and identically distributed (i.i.d) exponentially with rate parameter τ. Using the lemma, it can be seen that the marginal distribution of y_i marginalised over the latent weight is $AL(\boldsymbol{x_i}^T\boldsymbol{\beta}, \tau, p)$.

In what follows, it is more convenient to work with $v_i = 1/w_i$. This means that each v_i given $\boldsymbol{\beta}$ and τ is i.i.d. inverse $\Gamma(1, \tau)$, and this can be viewed as the prior distribution on v_i. The full likelihood $l(\boldsymbol{y}|\boldsymbol{\beta}, \tau, v_1, v_2, \cdots, v_n)$ is then proportional to

$$\tau^{n/2} \left(\prod_{i=1}^{n} v_i^{1/2} \right) \exp\left(-\frac{\tau p(1-p)}{4} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta}) \right). \tag{8}$$

Here, $\boldsymbol{u} = [u_i]_{i=1}^n$ with

$$u_i = y_i - \frac{1-2p}{p(1-p)v_i}$$

and $\boldsymbol{V} = \text{diag}[v_i]_{i=1}^n$. The posterior distribution can then be calculated using Bayes theorem

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \tau, v_1, v_2, \cdots, v_n|\boldsymbol{y}) &\propto l(\boldsymbol{y}|\boldsymbol{\beta}, \tau, v_1, v_2, \cdots, v_n)\pi(v_1, v_2, \cdots, v_n|\boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}|\tau)\pi(\tau) \\
&\propto \tau^{n/2} \left(\prod_{i=1}^{n} v_i^{1/2} \right) \exp\left(-\frac{\tau p(1-p)}{4} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta}) \right) \\
&\quad \times \left(\prod_{i=1}^{n} v_i^{-2} \right) \tau^n \exp\left(-\tau \sum_{i=1}^{n} v_i^{-1} \right) \\
&\quad \times \exp\left(-\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b}_0)^T \boldsymbol{B}_0^{-1} (\boldsymbol{\beta} - \boldsymbol{b}_0) \right) \\
&\quad \times \tau^{c_0 - 1} \exp(-d_0 \tau).
\end{aligned} \tag{9}$$

The key to constructing a PCG sampler is that we can obtain a reduced conditional posterior for τ whose parameters are easier to calculate by integrating out (or partially collapsing) the latent weights. This is equivalent to multiplying the reduced likelihood in (5) by the prior for τ in (6). This gives

$$\pi(\tau|\boldsymbol{\beta}, y) \propto \tau^n \exp\left(-\tau \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}) \right) \times \tau^{c_0 - 1} \exp(-b_0 \tau),$$

hence

$$\tau|\boldsymbol{\beta}, \boldsymbol{y} \sim \Gamma\left(c_0 + n, d_0 + \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x_i}^T \boldsymbol{\beta}) \right).$$

To help identify the full conditional distribution for $\boldsymbol{\beta}$, we can decompose the term $(\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta})$ in (9) into a sum of squares

$$(\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta}^*)^T \boldsymbol{V} (\boldsymbol{u} - \boldsymbol{X}\boldsymbol{\beta}^*) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} (\boldsymbol{\beta} - \boldsymbol{\beta}^*) \tag{10}$$

provided we choose $\boldsymbol{\beta}^*$ to satisfy

$$\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} \boldsymbol{\beta}^* = \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{u}.$$

Using (10) and the usual trick of completing the square, we can deduce that

$$\boldsymbol{\beta}|\tau, v_1 \cdots v_n, \boldsymbol{y} \sim N_k\left(\hat{\boldsymbol{\beta}}, \left(\frac{\tau p(1-p)}{2} \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} + \boldsymbol{B}_0^{-1} \right)^{-1} \right),$$

where

$$\hat{\boldsymbol{\beta}} = \left(\frac{\tau p(1-p)}{2}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}\left(\frac{\tau p(1-p)}{2}\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{u} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right).$$

To complete the PCG sampler, we require the full conditional posterior of $v_1, v_2, \cdots, v_n$. For a particular value of i, we have

$$\begin{aligned}
\pi(v_i|\boldsymbol{\beta}, \tau, \boldsymbol{y}) \quad &\propto v_i^{-3/2}\exp\left(-\frac{\tau p(1-p)v_i}{4}\left(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta} - \frac{1-2p}{p(1-p)v_i}\right)^2 + \frac{\tau}{v_i}\right) \\
&\propto v_i^{-3/2}\exp\left(-\frac{\tau p(1-p)(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 v_i}{4} - \frac{1}{v_i}\left(\frac{\tau p(1-p)}{4}\frac{(1-2p)^2}{p^2(1-p)^2} + \tau\right)\right) \\
&= v_i^{-3/2}\exp\left(-\frac{\tau p(1-p)(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta})^2 v_i}{4} - \frac{\tau}{4p(1-p)v_i}\right).
\end{aligned}$$

This can be recognised as an inverse Gaussian (IG) distribution with pdf

$$\begin{aligned}
f(x|\lambda, \nu) \quad &\propto \quad x^{-3/2}\exp\left(\frac{-\lambda(x-\nu)^2}{2\nu^2 x}\right), x, \nu, \lambda > 0 \\
&\propto \quad x^{-3/2}\exp\left(-\frac{\lambda x}{2\nu^2} - \frac{\lambda}{2x}\right),
\end{aligned}$$

from which we can deduce that

$$\lambda = \frac{\tau}{2p(1-p)}, \nu = \frac{1}{p(1-p)|y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}|}.$$

Hence, v_i given $\boldsymbol{\beta}, \tau$ and $\boldsymbol{y}$ is distributed as

$$IG\left(\frac{1}{p(1-p)|y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}|}, \frac{\tau}{2p(1-p)}\right).$$

We then have

$$\pi(v_1, v_2, \cdots, v_n|\boldsymbol{\beta}, \tau, \boldsymbol{y}) = \prod_{i=1}^{n}\pi(v_i|\boldsymbol{\beta}, \tau, \boldsymbol{y}).$$

The Inverse Gaussian distribution can be simulated using results from Michael *et al.* (1976).

Since we are using a PCG sampler, the order in which the parameters are updated is crucial to ensuring the Markov Chain converges to the desired stationary distribution (see van Dyk and Park (2008)). We therefore construct the PCG sampler for quantile regression using the priors in (6) and (7) as follows:

1. Fix the value of p so that the pth quantile is modelled. Generate initial values $\boldsymbol{\beta}^{(0)}$.

2. Generate

$$\tau^{(1)} \sim \Gamma\left(c_0 + n, d_0 + \sum_{i=1}^{n}\rho_p(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}^{(0)})\right).$$

3. Generate $v_i^{(1)}, i = 1, \cdots, n$ independently from an inverse Gaussian distribution
$$v_i^{(1)} \sim IG\left(\frac{1}{p(1-p)|y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}^{(0)}|}, \frac{\tau^{(1)}}{2p(1-p)}\right).$$

4. Calculate
$$\hat{\boldsymbol{\beta}}^{(1)} = \left(\frac{\tau^{(1)}p(1-p)}{2}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}\left(\frac{\tau^{(1)}p(1-p)}{2}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{u}^{(1)} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0\right),$$
where $\boldsymbol{V}^{(1)} = \text{diag}(v_i^{(1)})_{i=1}^n$ and $\boldsymbol{u}^{(1)} = [u_i^{(1)}]_{i=1}^n$ with
$$u_i^{(1)} = y_i - \frac{1-2p}{p(1-p)v_i^{(1)}}.$$
Finally draw $\boldsymbol{\beta}^{(1)}$ from a normal distribution
$$\boldsymbol{\beta}^{(1)} \sim N_k\left(\hat{\boldsymbol{\beta}}^{(1)}, \left(\frac{\tau^{(1)}p(1-p)}{2}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}\right).$$

5. Repeat steps 2 to 4 N times until convergence and remove the first M iterations as burn in.

Updating the parameters in that order will ensure that the posterior distribution in (9) is the stationary distribution. This is because combining steps 2 and 3 essentially produces draws from the conditional posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{y})$ where $\boldsymbol{\theta} = [v_1\, v_2\, \cdots\, v_n\, \tau]^T$. Hence the PCG sampler for quantile regression is really a blocked version of the ordinary Gibbs sampler as it alternates between $\pi(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{y})$ and $\pi(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y})$.

It is worth noting that if p=0.5, i.e. we are modelling the median, then since $\rho_{0.5}(z) = |z|/2$ and $u_i^{(1)} = \cdots = u_i^{(N)} = y_i$, the PCG sampler simplifies a little. The PCG sampler for median regression using the priors in (6) and (7) can be summarised as follows:

1. Generate initial values $\boldsymbol{\beta}^{(0)}$.

2. Generate
$$\tau^{(1)} \sim \Gamma\left(c_0 + n, d_0 + \frac{1}{2}\sum_{i=1}^n |y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}^{(0)}|\right).$$

3. Generate $v_i^{(1)}, i = 1, \cdots, n$ independently from an inverse Gaussian distribution
$$v_i^{(1)} \sim IG\left(\frac{4}{|y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}^{(0)}|}, 2\tau^{(1)}\right).$$

4. For $p = 0.5$,

$$\hat{\boldsymbol{\beta}}^{(1)} = (\frac{\tau^{(1)}}{8}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{X} + \boldsymbol{B}_0^{-1})^{-1}(\frac{\tau^{(1)}}{8}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{y} + \boldsymbol{B}_0^{-1}\boldsymbol{b}_0).$$

   Finally draw $\boldsymbol{\beta}^{(1)}$ from a normal

$$\boldsymbol{\beta}^{(1)} \sim N_k\left(\hat{\boldsymbol{\beta}}^{(1)}, \left(\frac{\tau^{(1)}}{8}\boldsymbol{X}^T\boldsymbol{V}^{(1)}\boldsymbol{X} + \boldsymbol{B}_0^{-1}\right)^{-1}\right).$$

5. Repeat N times until convergence and discard the first M iterations as burn in.

2.3 Marginal distribution of $\boldsymbol{\beta}|\boldsymbol{y}$

Inference in quantile regression usually focuses on the unknown parameter $\boldsymbol{\beta}$. We can obtain the marginal distribution of $\boldsymbol{\beta}|\boldsymbol{y}$ by integrating out τ and the latent weights $v_1, \cdots, v_n$ from the joint posterior distribution, using

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) = \int \cdots \int \pi(\boldsymbol{\beta}, \tau, v_1 \cdots v_n|\boldsymbol{y}) d\tau dv_1 \cdots dv_n. \tag{11}$$

This integral cannot be solved analytically, but we can implement Monte Carlo integration. A Monte Carlo estimate for (11) is given by

$$\frac{1}{N-M}\sum_{g=M+1}^{N}\pi(\boldsymbol{\beta}|\tau^{(g)}, v_1^{(g)} \cdots v_n^{(g)}, \boldsymbol{y}),$$

where the index g runs over post convergence iterations. We can also use Monte carlo integration to obtain numerical summaries of the marginal posterior. For the marginal posterior mean, we have

$$E[\boldsymbol{\beta}|\boldsymbol{y}] \approx \frac{1}{N-M}\sum_{g=M+1}^{N}E[\boldsymbol{\beta}|\tau^{(g)}, v_1^{(g)} \cdots v_n^{(g)}, \boldsymbol{y}] = \frac{1}{N-M}\sum_{g=M+1}^{N}\hat{\boldsymbol{\beta}}^{(g)},$$

where $\hat{\boldsymbol{\beta}}^{(g)}$ is defined in section 2.2. As Casella and George (1992) point out, this estimate is better than just using the sample from $\boldsymbol{\beta}|\tau, v_1 \cdots v_n, \boldsymbol{y}$ because the conditional distributions conditional on other simulated parameters carry more information about the marginal distribution than just the simulated values from the parameter of interest conditional on the others.

2.4 Prediction

The results from section 2.3 can be used to obtain accurate predictive densities of the pth quantile of a new observation $\boldsymbol{y}_{new}$ at a given new design matrix $\boldsymbol{X}_{new}$.

Since $Q_p(\boldsymbol{y}_{new}|\boldsymbol{X}_{new}, \boldsymbol{y}) = \boldsymbol{X}_{new}\boldsymbol{\beta}|\boldsymbol{y}$, an estimate for the predictive density is given by

$$\pi(Q_p(\boldsymbol{y}^\dagger|\boldsymbol{X}^\dagger, \boldsymbol{y})) \approx \frac{1}{N-M}\sum_{g=M+1}^{N} \pi(\boldsymbol{X}_{new}\boldsymbol{\beta}|\tau^{(g)}, v_1^{(g)} \cdots v_n^{(g)}, \boldsymbol{y}).$$

An estimate of the predictive mean pth quantile can be easily calculated from the marginal posterior mean $E[\boldsymbol{\beta}|\boldsymbol{y}]$ using

$$E[Q_p(\boldsymbol{y}_{new}|\boldsymbol{X}_{new}, \boldsymbol{y})] = \boldsymbol{X}_{new}E[\boldsymbol{\beta}|\boldsymbol{y}].$$

The predictive median, if required, and a 95% credible interval for the pth quantile of $\boldsymbol{y}_{new}$ can be obtained from the sample $[\boldsymbol{X}_{new}\boldsymbol{\beta}|\tau^{(g)}, v_1^{(g)}, v_2^{(g)}, \cdots, v_n^{(g)}]_{g=M+1}^{N}$, using the empirical 2.5%, 50% and 97.5% quantiles.

2.5 The Bayes Factor

The Bayes Factor B_{12} for model M_1 vs. model M_2 is defined as

$$B_{12} := \frac{l(y|M_1)}{l(y|M_2)}.$$

Given prior odds $\pi(M_1)/\pi(M_2)$, the posterior odds can be calculated using Bayes Theorem giving

$$\frac{\pi(M_1|y)}{\pi(M_2|y)} = \frac{\pi(M_1)}{\pi(M_2)}B_{12}$$

The prior probabilities $\pi(M_1)$ and $\pi(M_2)$ are usually taken to be equal. In that case, the posterior odds depend solely on the Bayes factor. The term $l(y|M_j)$ for $j = 1, 2$ is the likelihood density marginalised over the parameters and is crucial to evaluating the Bayes Factor. Unfortunately, for most cases, the marginal likelihood is extremely difficult to calculate analytically. Chib (1995) suggests a method to approximate the logarithm of this value based on the Gibbs sample and we can adapt this to our case. For a model M_j, the approximate marginal likelihood can be calculated as follows (we have omitted the dependence on model M_j on the right hand side of the equations to make notation clearer):

1. Using Bayes theorem, we have

$$l(\boldsymbol{y}|M_j) = \frac{l(\boldsymbol{y}|\boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}, \tau)}{\pi(\boldsymbol{\beta}, \tau|\boldsymbol{y})}.$$

2. Use an estimator $\tilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and $\tilde{\tau}$ for τ that has a high posterior density. Replacing the unknown parameters by their estimates gives the logarithm of $\hat{l}(\boldsymbol{y}|M_j)$ as

$$\log(\hat{l}(\boldsymbol{y}|M_j)) = \log(l(\boldsymbol{y}|\tilde{\boldsymbol{\beta}}, \tilde{\tau})) + \log(\pi(\tilde{\boldsymbol{\beta}}, \tilde{\tau})) - \log(\pi(\tilde{\boldsymbol{\beta}}, \tilde{\tau}|\boldsymbol{y}))$$

3. We can express the last term $\log(\pi(\tilde{\boldsymbol{\beta}}, \tilde{\tau}|\boldsymbol{y}))$ as

$$\log(\pi(\tilde{\boldsymbol{\beta}}, \tilde{\tau}|\boldsymbol{y})) = \log(\pi(\tilde{\tau}|\tilde{\boldsymbol{\beta}}, \boldsymbol{y})) + \log(\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{y})).$$

Finally, approximate $\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{y})$ by

$$\frac{1}{N-M} \sum_{g=M+1}^{N} \pi(\tilde{\boldsymbol{\beta}}|\tau^{(g)}, v_1^{(g)} \cdots v_n^{(g)}, \boldsymbol{y}).$$

Under the prior assumption that the models are equally likely, if $\log B_{12}$ is positive (negative), then model 1 (model 2) is preferred. Note however that this method cannot be implemented if an improper prior is used, because the prior density $\pi(\tilde{\boldsymbol{\beta}}, \tilde{\tau})$ does not exist.

2.6 Semiparametric extension

In order to construct a model with more flexible tail behaviour, a general scale mixture of *AL* distribution can be used. Following Kottas and Gelfand (2001) among others, a general nonparametric such mixture with a Dirichlet process (DP) prior for the mixing distribution, which is supported on R^+, then a semiparametric regression model extension can be constructed. Specifically, denoting by $DP(\theta G_0)$ be the DP with precision parameter θ and base distribution G_0, define the nonparametric scale mixture as

$$\int ALD(y - \boldsymbol{x}^T\boldsymbol{\beta}, \theta\,\tau, p)\, dG(\theta), \quad G \sim DP(\alpha G_0).$$

Then the PCG algorithm developed in Section 3.1 can be extended by simply adding one more step:

$$y_i|\theta_i \sim ALD(y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}, \theta_i\,\tau, p), \quad i = 1, ..., n,$$

in which

$$\theta_i|G \sim G \quad i = 1, ..., n,$$
$$G|\alpha, d \sim \mathrm{DP}(\alpha G_0),$$

and $\alpha \sim \Gamma(a, b)$, and the basic distribution G_0 is taken to be an inverse $\Gamma(s, t)$, with mean $t/(s-1)$ if $s > 1$.

3 Applications

In this section, we apply Bayesian quantile regression using the PCG sampler firstly to an artificial example and secondly to a real dataset which investigates how the patients' admission age, admission method and gender can affect length of stay (LOS) in hospital emergency. All analyses were done in R (R Development Core Team, 2008) and the convergence and number of burn in samples to exclude was assessed using the package CODA (Plummer *et al.*, 2008).

Table 1: Marginal posterior means of $\boldsymbol{\beta}|\boldsymbol{y}$

p	$E[\beta_0(p)\|\boldsymbol{y}]$	True value of $\beta_0(p)$	$E[\beta_1(p)\|\boldsymbol{y}]$	True value of $\beta_1(p)$
0.05	8.4984	8.3551	-1.1566	-1.1495
0.25	9.2764	9.3255	-1.0418	-1.0613
0.50	9.9312	10.0000	-0.9808	-1.0000
0.75	10.7433	10.6745	-0.9545	-0.9387
0.95	11.9287	11.6449	-0.8785	-0.8505

Table 2: Predictive mean quantiles for y_{new} at $x_{new} = 5$

p	$E[Q_p(y_{new}\|x_{new}, \boldsymbol{y})]$	True value of $Q_p(y_{new}\|x_{new})$
0.05	2.7154	2.6075
0.25	4.0674	4.0189
0.50	5.0272	5.0000
0.75	5.9708	5.9811
0.95	7.5362	7.3925

3.1 Simulation

To construct an artificial example, we used 50 random uniform numbers on the interval $(0, 10)$ as the covariates. We then generated 5 observations at each covariate from the model $y_i = \beta_0 + \beta_1 x_i + 1/11(11 + x_i)\epsilon_i$, $\epsilon_i \sim N(0, 1)$ making 250 observations in total. We chose $\beta_0 = 10, \beta_1 = -1$. We then carried out quantile regression at 5 different quantiles, namely 5%, 25%, 50% 75% and 95%. We assumed no prior knowledge and used independent $N(0, 10^6)$ priors on all regression parameters and $\Gamma(10^{-3}, 10^{-3})$ on all inverse scale parameters. Following the CODA analysis, we ran the Gibbs sampler for 11,000 iterations and discarded the first 1,000 as burn in. Table 1 compares the posterior mean of the marginal distribution of $\beta_0(p)|\boldsymbol{y}$ against the true quantile, given by $10 + Q_p(N(0, 1))$ and compares the posterior mean of $\beta_1(p)|\boldsymbol{y}$ against the true quantile, given by $1/11Q_p(N(0.1)) - 1$. Figure 1 superimposes the quantile lines onto the data. Finally, table 2 compares the means of the 5 predictive quantile distributions of a new observation at $x_{new} = 5$. against the true value, given by $5 + 16/11Q_p(N(0, 1))$.

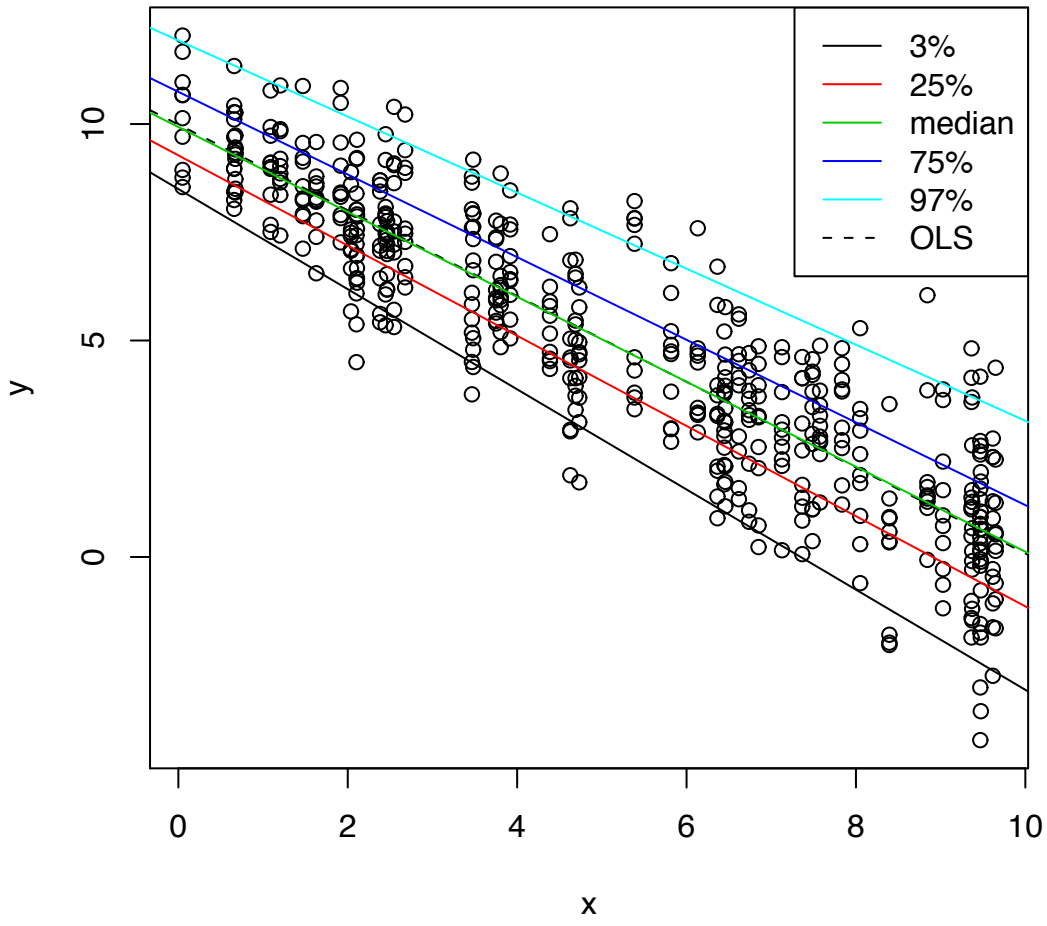


Figure 1: Fitted linear quantiles using marginal posterior mean of $\boldsymbol{\beta}|\boldsymbol{y}$

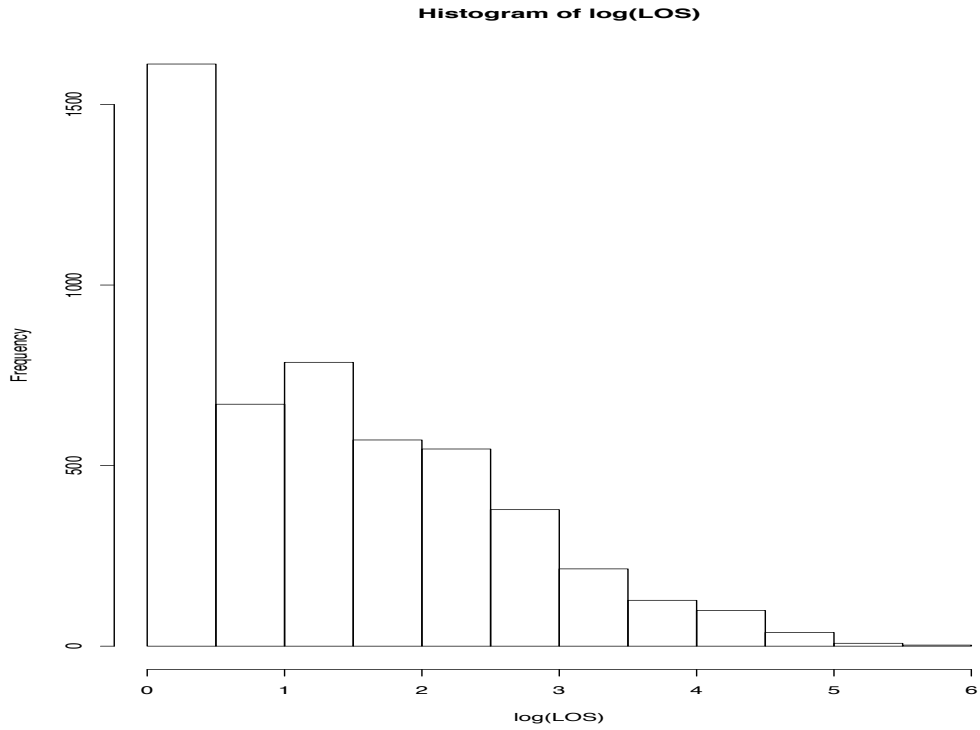
3.2 Length of stay as a performance indicator: quantile regression approach

Length of stay in hospital (LOS) is a crucial variable for the quality of life of all patients and their families. Furthermore, it is the single most important component in the consumption of hospital resources. It is also very important for hospital planning since it is a direct determinant of the number of beds to be provided. Moreover, LOS is a frequent point of comparison between patients, hospitals and countries. We say that a patient's LOS at the pth $(0 < p < 1)$ quantile of a LOS distribution if his/her LOS is longer than the proportion p of the reference group of patients and shorter than the proportion $(1-p)$. Thus, half of patients stay longer than the median patient and half stay shorter. Similarly, the quartiles divide the patient population into four segments with equal proportions of the reference population in each segment. The quintiles divide the population into five parts; the deciles into ten parts. Due to the strong skewness of the distribution (see Figure 2), we model the conditional quantiles of log(LOS) as a linear combination of admission age, admission method (a coded value dependent on how the patient was admitted) and gender (a binary variable taking 0 if the patient is female and 1 if the patient is male). We also fit a reduced model to see whether the gender of a patient has an effect on their log(LOS). We used informative priors on the regression parameters by assuming that all quantile planes would be parallel. We centred the corresponding normal distributions on the least squares solution (plus an additional $Q_p(N(0,1))$ for the intercept terms) with prior covariance matrix equal to the identity matrix. We placed uninformative priors on the inverse scale parameters. In order to see whether gender influences the log(LOS), we provide the approximate marginal likelihood $\hat{l}(\log(LOS)|\text{Model})$. We also provide the deviance information criterion (DIC) (see Spiegelhalter *et al.*, 2002), a statistic familiar to users of WinBUGS (Lunn *et al.*, 2000).

Like the previous section, 11,000 samples were generated with the first 1,000 samples rejected as burn in. Table 3 shows the posterior mean and 95% credible intervals for each of the 0.25, 0.5 and 0.75 quantile regression parameters. From this table, it can be seen that as the age of a patient increases, so does the log length of stay. This holds for all quantiles that were measured. What is also apparent is that as the "code" value associated with the method of admission increases, the log length of stay also increases. The maximum code value for this dataset was 28, which corresponds to a patient being transferred from another hospitals A&E, suggesting that this patient had the longest time in hospital. This was consistent across all measured quantiles.

At the median and 75th percentile, both the DIC and Bayes Factors favoured the full model. The Bayes Factors comparing the full model with the reduced model were 93.69 (median) and 343.78 (75th percentile) indicating strong evidence on the Jeffreys (Jeffreys, 1961) that the full model was better. The evidence was not as strong at the 25th percentile with the DIC still preferring the

Figure 2: the histogram of LOS



full model though not as convincingly. The Bayes Factor was 0.50 suggesting weak evidence for the reduced model. However, the 95% credible interval of the gender coefficient does not cross zero, suggesting that we should prefer the full model.

Table 3: Results of quantile regression analysis on log length of stay in hospital

p	Model Covariates	Marginal posterior mean	95% credible interval
0.25	Intercept	-2.2037	(-2.4877,-1.8984)
	Admission age	0.0072	(0.0058,0.0085)
	Admission method	0.0989	(0.0842,0.1122)
	Gender	-0.0580	(-0.1092,-0.0093)
0.50	Intercept	-2.6639	(-3.0039,-2.3184)
	Admission age	0.0205	(0.0196,0.0213)
	Admission method	0.1268	(0.1107,0.1427)
	Gender	-0.1166	(-0.1721,-0.0623)
0.75	Intercept	-3.2594	(-3.6365,-2.8325)
	Admission age	0.0234	(0.0224,0.0244)
	Admission method	0.1832	(0.1640,0.2001)
	Gender	-0.1411	(-0.2060,-0.0764)

Table 4: Comparison statistics for investigating whether length of stay depends on gender

p	Model	$\log \hat{l}$(log(LOS)|Model)	DIC
0.25	Full	-7714.82	15377.58
	Reduced	-7714.13	15381.14
0.50	Full	-7898.64	15745.73
	Reduced	-7903.18	15760.24
0.75	Full	-8367.01	16682.80
	Reduced	-8372.85	16699.46

4 Discussions

We have introduced a PCG sampler for Bayesian quantile regression which uses simple conditional distributions to simulate the joint posterior distributions of all the unknown parameters in the regression models, including the latent variables. We have also seen how it can be used to obtain marginal and predictive distributions and to carry out model testing for a particular quantile.

Using the location-scale mixture of normals representation of the AL distribution also permits more complicated quantile regression models to be analysed. In particular, the semiparametric model extension via nonparametric mixtures of AL distributions using a DP prior (section 2.6) allows the data to drive the shape of the error distribution. Nevertheless, Richardson (1999) pointed out that popular forms of priors tend to be those which have parameters that can be set straightforwardly and which lead to posterior with a relatively straightforward form.

Acknowledgements

The authors' research were partially supported by an EPSRC doctorial Training Grant.

Appendix

Proof of Lemma

The proof proceeds in a similar fashion to the proof of Proposition 3.2.1. in Kotz *et al.* (2001).

If $X \sim AL(\mu, \tau, p)$, the moment generating function $M_X(t) = \exp(tX)$ is given by

$$M_X(t) = \frac{\tau^2 p(1-p)\exp(\mu t)}{(\tau p - t)(\tau(1-p)+t)}, \quad -\tau(1-p) < t < \tau p \tag{12}$$

(Yu and Zhang, 2005).

Let $Y = \sqrt{\frac{2\xi}{\tau p(1-p)}}Z + \frac{1-2p}{p(1-p)}\xi + \mu$. Conditioning on ξ, we have

$$M_Y(t) = E[\exp(tY)] = E[E[\exp(tY)|\xi]]$$
$$= \int_0^\infty \tau \exp(-\tau\xi) \times \exp\left(\left(\frac{1-2p}{p(1-p)}\xi + \mu\right)t\right) M_Z\left(\sqrt{\frac{2\xi}{\tau p(1-p)}}t\right) d\xi. \tag{13}$$

Now since Z is standard normal, $M_Z(t) = \exp(t^2/2)$. Substituting this into equation (13), we have

$$M_Y(t) = \tau \exp(\mu t) \int_0^\infty \exp\left(-\xi\left(\tau - \frac{1-2p}{p(1-p)}t - \frac{1}{\tau p(1-p)}t^2\right)\right) d\xi \tag{14}$$

$$
\begin{aligned}
&= \tau \exp(\mu t) \left(\tau - \frac{1-2p}{p(1-p)} t - \frac{1}{\tau p(1-p)} t^2 \right)^{-1} \\
&= \frac{\tau^2 p(1-p) \exp(\mu t)}{p(1-p)\tau^2 - (1-2p)\tau t - t^2}. \qquad (15)
\end{aligned}
$$

Note that the denominator in (15) factorises to $(\tau p - t)(\tau(1-p) + t)$, which is the same as the denominator in (12). Hence $M_Y(t) = M_X(t)$, and therefore $Y =^d X$. Note also that the conditions for the integral in (14) to converge are exactly those for which $M_X(t)$ exists.

REFERENCES

Abrevaya, J. (2001), The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes, *Empirical Economics*, 26, 247–57.

Cade, B.S. and Noon, B.R. (2003), A gentle introduction to quantile regression for ecologists, *Frontiers in Ecology and the Environment,* 1, 412–420.

Casella, G. and George, E.I. (1992), Explaining the Gibbs Sampler, *The American Statistician,* 46, 167–174.

Chamberlain, G. (1994), *Quantile Regression, Censoring and the Structure of Wages,* Elsevier, New York.

Chen, L. and Yu, K. (2008), Automatic Bayesian Quantile Regression Curve, forthcoming in *Statistics and Computing.*

Chib, S. (1995), Marginal Likelihood from the Gibbs Output, *Journal of the American Statistical Association,* 90, 1313–1321.

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

Dunson,D.B., Pillai, N. and Park, J. (2007), Bayesian density regression, *J. R. Statist. Soc. B*, 69, 163-183.

Dunson, D.B., Reed, C. and Yu, K. (2009), Bayesian variable selection in quantile regression, *Technical Report, Brunel University, UK.*

Engle, R.F. & S. Manganelli (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantile. *Journal of Business and Economic Statistics*, 22, 367–381.

Geman, S. and Geman, D. (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution, *Biostatistics*, 8, 140–154.

Jeffreys, H. (1961), Theory of Probability. Clarendon Press: Oxford.

Koenker, R. and Hallock, K.F. (2001), Quantile regression, *Journal of Economic Perspectives*, 15, 143-156.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press, London.

Kottas, A. and Gelfand, A.E. (2001), Bayesian semiparametric median regression modeling, *Journal of the American Statistical Association*, 96, 1458–1468

Kottas, A., and Krnjajić, M., (2009), Bayesian Semiparametric Modeling in Quantile Regression, To appear in Scandinavian Journal of Statistics.

Lancaster, T. and Jun, S.J. (2009), Bayesian quantile regression methods, forthcoming in *Journal of Applied Econometrics*.

Levin, J. (2001), For Whom the Reduction Counts: A Quantile Regression Analysis of Class size on Scholastic Achievement, *Empirical Economics*, 26, 221–246.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.

Michael, J. R., Schucany, W.R. and Haas, R.W. (1976) Generating Random Variates Using Transformations with Multiple Roots, *The American Statistician*, 30, 88–90.

Min, I. and Kim, I. (2004), A Monte Carlo comparison of parametric and nonparametric quantile regressions, *Applied Economic Letters* 11, 71-74.

Plummer M., Best N., Cowles K. and Vines, K. (2008). coda: Output analysis and diagnostics for MCMC. R package version 0.13-3.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Richardson, S. (1999) "Contribution to the Discussion of Walker *et al.*, Bayesian Nonparametric Inference for Random Distribution and Related Functions" *J. R. Statist. Soc.* B, **61**, 485–527.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde A. (2002) Bayesian measures of model complexity and fit (with discussion), *J Roy Statist Soc B,* 64, 583–640.

Tsionas, E.G. (2003) Bayesian Quantile Inference, *Journal of statistical compu-*

tation and simulation, 73, 659–674

van Dyk, D. A. and Park, T. (2008). Partially Collapsed Gibbs Samplers: Theory and Methods. *Journal of the American Statistical Association* 103, 790–796.

Yu, K. and 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.

Yu, K. and Stander, J. (2007), Bayesian Analysis of a Tobit Quantile Regression Model, *Journal of Econometrics*, 137, 260–276

Yu, K. and Zhang, J. (2005), A three-parameter asymmetric laplace distribution and its extension *Communications in Statistics - Theory and Methods,* 34, 1867–1879.

Yu, K., Dunson, D.B. and Reed, C. (2009), Bayesian skewed Lesso, *Technical Report, Brunel University, UK.*