Statistics and Its Interface Volume 13 (2020) 237–249

Sparse signal shrinkage and outlier detection in high-dimensional quantile regression with variational Bayes

Daeyoung Lim, Beomjo Park, David Nott, Xueou Wang, and Taeryon Choi*

Model misspecification can compromise valid inference in conventional quantile regression models. To address this issue, we consider two flexible model extensions for high-dimensional data. The first is a Bayesian quantile regression approach with variable selection, which uses a sparse signal shrinkage prior on the high-dimensional regression coefficients. The second extension robustifies conventional parametric quantile regression methods by including observation specific mean shift terms. Since the number of outliers is assumed to be small, the vector of mean shifts is sparse, which again motivates the use of a sparse signal shrinkage prior. Specifically, we exploit the horseshoe+ prior distribution for variable selection and outlier detection in the high-dimensional quantile regression models. Computational complexity is alleviated using fast mean field variational Bayes methods, and we compare results obtained by variational methods with those obtained using Markov chain Monte Carlo (MCMC).

AMS 2000 subject classifications: Primary 62G35, 62F15; secondary 62J07.
Keywords and phrases: Asymmetric Laplace distribution, Horseshoe+ prior, Outlier detection, Quantile regression, Variational Bayes.

1. INTRODUCTION

Regression modeling of high-dimensional data, where the number of covariates p is much larger than the number of observations n, is increasingly common in modern statistical applications. Fitting realistic models to such data can involve considerable statistical and computational complexity. Furthermore, many of the covariates are often uninformative, and this may adversely affect the parameter estimation with error accumulation known as the "curse of dimensionality" [17]. In view of this, the development of a methodology for detecting insignificant variables and outliers has become more important. This paper explores the performance of shrinkage priors in high-dimensional Bayesian linear quantile regression models, with applications to outlier detection and variable selection, all implemented using fast variational Bayes algorithms. We show that the methodology works well in a variety of examples, from both a statistical and a computational point of view.

First proposed by [57], Bayesian quantile regression adopts the asymmetric Laplace distribution as the additive error distribution, that is, $y = \mu(\boldsymbol{x}) + \epsilon$, $\epsilon \sim \mathrm{AL}_\tau(\sigma)$ where $\mu(\boldsymbol{x})$ denotes the conditional mean of y given the vector of covariate $\boldsymbol{x}$, τ denotes the desired quantile level, and $\mathrm{AL}_\tau(\sigma)$ is the asymmetric Laplace (AL) distribution with scale parameter σ and $\tau \in (0, 1)$ a prespecified quantile level. Quantile regression is somewhat more robust to outliers than conventional mean regression models, and it also provides a richer way to describe the relationship between the response y and covariates $\boldsymbol{x}$. Bayesian quantile regression has thus become increasingly popular. Some recent contributions include [26], who proposed an efficient Gibbs sampler based on an analytically tractable location-scale mixture representation of the AL distribution, and [22] who implemented Bayesian quantile regression for a semiparametric model with a Dirichlet process mixture error distribution similar to that of [25]. Although the AL distribution is used only to obtain a log-likelihood function equivalent to the check loss function $\rho_\tau(u) = u(\tau - I(u < 0))$ (where I denotes the indicator function), it has been proven empirically and theoretically that the model produces correct estimates [41].

Variable selection in Bayesian quantile regression is a field with a growing literature. [56] rely on point mass spike and slab priors to probabilistically remove noise variables. [1] adopts the adaptive LASSO penalty, which is then translated to the normal-exponential-gamma hierarchical prior, except that the gamma distribution is replaced with the inverse gamma distribution. [28] give a unified treatment of some common penalties, including the LASSO, group LASSO and elastic net. A wide variety of other continuous shrinkage priors have been suggested in the literature to select significant variables and improve predictive performance in high-dimensional regression problems [35, 20, 7, 2, 3, 18, 11].

*Corresponding author.

Methods for model and variable selection are potentially a useful tool for outlier detection. It is a common practice to detect suspected outliers in post-analysis diagnostics using studentized residuals or generalized Cook's distances [14] in a leave-one-out approach [51]. However, model-based approaches adopting mean shift parameters for every observation have also been used by [30] and [40], among others. Motivated by the penalized regression approach of [40], [54] recently proposed a sparse signal regression and outlier detection method using the horseshoe+ prior [5] in Bayesian linear regression. A Bayesian approach to outlier detection in Bayesian quantile regression was considered in [39], making use of the latent variables in the scale mixture representation of the AL distribution. Here we generalize the work of [54] using mean shift terms with sparsity priors for quantile regression, and this provides a widely applicable strategy for outlier detection in a range of models.

The sparse signal shrinkage priors we consider are the horseshoe [11] and horseshoe+ [5] priors. The horseshoe prior has recently gained popularity because of its good computational and statistical performance, and the horseshoe+ prior is an extension involving an additional hierarchical layer. The extra layer of mixing variables makes local shrinkage effects marginally dependent after integrating out the additional hierarchical layer, which is helpful in the case of ultra-sparse effects. Computations with these models can be carried out conveniently using a Gibbs sampler in conditional conjugate models with a hierarchical representation of the half-Cauchy distribution as a mixture of inverse-gamma distributions ([47] and [31]). However, the location-scale mixture representation of the AL distribution involves an additional parameter for each observation, which renders the Markov chain Monte Carlo (MCMC) algorithm infeasible for massive datasets or data with the number of covariates much larger than the number of observations ($p > n$). Thus, we instead consider a faster approximate estimation method, namely variational Bayes (VB).

There are many uses of the horseshoe prior applied to variable selection in linear regression [37, 19, 36] but none in quantile regression or the approaches considered here, as far as we know. Formally, extending linear regression methods to linear quantile regression is not difficult. However, the augmented model we consider in this paper for outlier detection in quantile regression results in a $p > n$ problem, and an alternative model variant we consider with distinct shrinkage parameters for outlier terms and coefficients also requires a fine-tuned initialization to effectively capture the signals from the high dimensional covariates.

The rest of this paper is organized as follows. Section 2 describes Bayesian quantile regression models for variable selection and outlier detection. In Section 3, we briefly review VB and its implementation for the proposed models. Section 4 applies the methods to synthetic and real data, comparing them with MCMC methods in terms of CPU time, and with existing methods in terms of accuracy, sensitivity, and specificity. All MCMC samplers are based on the same likelihood and prior distributions as those of VB. Finally, we discuss possible extensions and the limitations of our approach in Section 6.

2. MODELS

2.1 Linear quantile regression model

Given a set of n observations consisting of the response variable y_i and p dimensional covariates $\boldsymbol{x}_i$, $i = 1, \ldots, n$, the linear quantile regression model is

$$y_i = \beta_{0,\tau} + \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \text{AL}_\tau(\sigma) \quad i = 1, \ldots, n \tag{1}$$

where $\beta_{0,\tau}$ is the intercept and $\boldsymbol{\beta}_\tau = (\beta_{1,\tau}, \ldots, \beta_{p,\tau})^T \in \mathbb{R}^p$ is a vector of regression coefficients. The assumed model (1) implies that the τth conditional quantile of y_i given $\boldsymbol{x}_i$ is $Q_\tau(y_i \mid \boldsymbol{x}_i) = \beta_{0,\tau} + \boldsymbol{x}_i^T \boldsymbol{\beta}_\tau$. The subscript τ indicates that $\beta_{0,\tau}$ and $\boldsymbol{\beta}_\tau$ depend on the quantile level τ; however, we omit the subscript for notational convenience hereafter.

The location-scale reparameterization of the AL distribution [26] allows convenient hierarchical modeling

$$\text{AL}_\tau(\epsilon \mid \sigma) = \int_0^\infty \frac{1}{\kappa_2 \sqrt{2\pi\sigma z}} \exp\left(-\frac{1}{2\kappa_2^2 \sigma z}(\epsilon - \kappa_1 z)^2\right) \pi(z)\, \mathrm{d}z \tag{2}$$

where $\kappa_1 = (1-2\tau)/(\tau(1-\tau))$, $\kappa_2^2 = 2/(\tau(1-\tau))$, and $\pi(z)$ is an exponential mixing density with mean σ. The hierarchical model becomes

$$\begin{aligned} y_i \mid \beta_0, \boldsymbol{\beta}, \sigma, \boldsymbol{z} &\overset{\text{ind}}{\sim} N(\beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \kappa_1 z_i, \kappa_2^2 \sigma z_i) \\ z_i \mid \sigma &\overset{\text{iid}}{\sim} \text{Exp}\left(\frac{1}{\sigma}\right) \\ \sigma &\sim \text{InvGam}\left(\frac{r_0}{2}, \frac{s_0}{2}\right) \end{aligned}$$

We place a normal prior on the intercept $\beta_0 \sim N\left(0, \sigma_0^2\right)$ and place an ultra-sparse *horseshoe+* prior [5] on the regression coefficients $\boldsymbol{\beta}$ for high-dimensional variable selection. The horseshoe+ prior on $\beta_1, \ldots, \beta_p$ is

$$\begin{aligned} (\beta_j \mid \lambda_j, \eta_j, A_\beta) &\sim N\left(0, \lambda_j^2\right) \\ (\lambda_j \mid \eta_j, A_\beta) &\sim \text{C}^+\left(0, A_\beta \eta_j\right) \\ \eta_j &\sim \text{C}^+(0, 1) \end{aligned} \tag{3}$$

where C^+ is a half-Cauchy distribution whose density is

$$p(\lambda_j \mid A_\beta) = \frac{2}{\pi A_\beta \left(1 + (\lambda_j / A_\beta)^2\right)}, \quad \lambda_j > 0.$$

In the original horseshoe prior $\lambda_j \mid A_\beta \sim \text{C}^+(0, A_\beta)$, and the horseshoe+ modification adds an extra hierarchical layer so that the local shrinkage effects are not independent after integrating out $\eta = (\eta_1, \ldots, \eta_p)^T$.

To allow for conditional conjugacy in models with half-Cauchy prior components, [31] introduced a hierarchical representation of the half-Cauchy (3) as a mixture of inverse-gamma distributions. Applying this here gives

$$
\begin{aligned}
(\beta_j \,|\, \lambda_j) &\sim N\left(0, \lambda_j^2\right) \\
\left(\lambda_j^2 \,|\, \zeta_j\right) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{\zeta_j}\right) \\
(\zeta_j \,|\, \eta_j) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{(A_\beta \eta_j)^2}\right) \\
\left(\eta_j^2 \,|\, \zeta_\eta\right) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{\zeta_\eta}\right) \\
\zeta_\eta &\sim \text{InvGam}\left(\frac{1}{2}, 1\right)
\end{aligned} \tag{4}
$$

Thus, the horseshoe+ prior reduces to a conditionally multivariate normal prior for $\boldsymbol{\beta}$, with multiple layers of inverse-gamma mixing variables.

2.2 Outlier detection in linear quantile regression models

As discussed in the introduction, we extend the mean-shift normal model for outlier detection ([40] and [54]) to Bayesian sparse signal quantile regression by including the shift parameter in Equation (1). This extension adds robustness against the distributional assumptions for normality. Specifically, the model is

$$
y_i = \beta_0 + \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_i + \epsilon_i,\ \epsilon_i \overset{\text{iid}}{\sim} \text{AL}_\tau(\sigma),\ i = 1, \cdots, n \tag{5}
$$

where γ_i indicates the quantile shift parameter and $\gamma = (\gamma_1, \ldots, \gamma_n)^T$ is assumed to be sparse. Since γ is sparse, a small number of outliers can be accommodated in the model through the mean-shift parameters. Sparsity also motivates the use of a horseshoe+ prior on γ. That is,

$$
\begin{aligned}
(\gamma_i \,|\, \lambda_{\gamma_i}, \eta_{\gamma_i}, A_\gamma) &\sim N\left(0, \lambda_{\gamma_i}^2\right) \\
(\lambda_{\gamma_i} \,|\, \eta_{\gamma_i}, A_\gamma) &\sim \text{C}^+\left(0, A_\gamma \eta_{\gamma_i}\right) \\
\eta_{\gamma_i} &\sim \text{C}^+(0, 1),
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
(\gamma_i \,|\, \lambda_{\gamma_i}) &\sim N\left(0, \lambda_{\gamma_i}^2\right) \\
\left(\lambda_{\gamma_i}^2 \,|\, \zeta_{\gamma_i}\right) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{\zeta_{\gamma_i}}\right) \\
(\zeta_{\gamma_i} \,|\, \eta_{\gamma_i}) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{(A_\gamma \eta_{\gamma_i})^2}\right) \\
\left(\eta_{\gamma_i}^2 \,|\, \zeta_{\eta,\gamma}\right) &\sim \text{InvGam}\left(\frac{1}{2}, \frac{1}{\zeta_{\eta,\gamma}}\right) \\
\zeta_{\eta,\gamma} &\sim \text{InvGam}\left(\frac{1}{2}, 1\right).
\end{aligned}
$$

We call the above model the full horseshoe+ quantile OD (outlier detection) model. This model can be represented as in (1) by augmenting the design matrix and concatenating the parameters, i.e., $\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{I}_n)$, $\tilde{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$. The augmented model is more restrictive compared to the full horseshoe+ quantile OD model, in that the former utilizes a single shrinkage parameter A_β, whereas the latter implements the separate shrinkage parameters A_β and A_γ on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively.

3. ESTIMATION

3.1 An overview of variational Bayes

Bayesian computations are most often carried out using MCMC methods. However, here we use VB as a fast alternative, since the primary focus of our paper is a setting where the dimension is high and MCMC is computationally burdensome. For a further background, see [34], [9] and [46].

VB considers a family of variational distributions $\mathcal{F}$ and minimizes within the family the Kullback-Leibler (KL) divergence to the true posterior. If we write q to denote the variational density function over the parameters, then

$$
\begin{aligned}
\log p\left(\boldsymbol{y}\right) &= \int \log \frac{\pi\left(\Theta\right) p\left(\boldsymbol{y} \,|\, \Theta\right)}{q\left(\Theta\right)} q\left(\Theta\right)\, d\Theta \\
&\quad + \int \log \frac{q\left(\Theta\right)}{\pi\left(\Theta \,|\, \boldsymbol{y}\right)} q\left(\Theta\right)\, d\Theta,
\end{aligned} \tag{6}
$$

where the first integral is the so-called variational lower bound, and the second integral is the KL divergence between the true posterior $\pi(\Theta \,|\, \boldsymbol{y})$ and the variational posterior $q(\Theta)$. Note that $\log p\left(\boldsymbol{y}\right)$ does not depend on the choice of q and thus minimizing the KL divergence in Equation (6) over $q \in \mathcal{F}$ is equivalent to maximizing the lower bound over the variational family. Thus a best approximation $q \in \mathcal{F}$ to the posterior is found as

$$
\hat{q}(\Theta) = \underset{q(\Theta) \in \mathcal{F}}{\text{argmax}} \left[\text{E}_{q(\Theta)} \left\{\log p(\boldsymbol{y}, \Theta)\right\} - \text{E}_{q(\Theta)} \left\{\log q(\Theta)\right\}\right], \tag{7}
$$

where we have written $p(\boldsymbol{y}, \Theta) = \pi\left(\Theta\right) p\left(\boldsymbol{y} \,|\, \Theta\right)$.

With conditionally conjugate priors and factorizing assumptions for the variational family $\mathcal{F}$, the optimal form of the variational posterior distribution can be derived analytically and the estimation can be done using a coordinate-ascent optimization [15]. This corresponds to so-called mean-field VB. Suppose the family $\mathcal{F}$ consists of densities of the form $\hat{q}(\Theta) = \prod_{j=1}^{B} q_j(\theta_j)$ for some partition $\theta = (\theta_1^T, \ldots, \theta_B^T)^T$ of θ. Let $\theta_j \,|\, \Theta_{-j}, \boldsymbol{y}$ be the full conditional of θ_j. Fixing $q_k(\theta_k)$, $k \neq j$, it can be shown that the optimal $q_j(\theta_j)$ is

$$
q_j^\star(\theta_j) \propto \exp\left(\text{E}_{\Theta \setminus \theta_j} \log p(\theta_j \,|\, \Theta_{-j}, \boldsymbol{y})\right)
$$

Variational message passing [53, 46] is a general algorithm used to implement mean-field VB in conjugate-exponential

family models. In non-conjugate models, an extension called non-conjugate variational message passing (NCVMP) is very useful [23, 45]. Assume $q_j(\theta_j)$ takes the exponential family form

$$q_j(\theta) = \exp\left\{\lambda_j^T S_j(\theta_j) - Z_j(\lambda)\right\}$$

where λ_j is the vector of natural parameters, $S_j(\theta_j)$ is the vector of sufficient statistics and $Z_j(\lambda_j)$ is the log normalizing constant. Then the NCVMP update is

$$\lambda_j \leftarrow \left(\text{cov}_{q(\theta)}\left\{S_j(\theta_j)\right\}\right)^{-1}\left(\frac{\partial}{\partial \lambda_j}\,\text{E}_{q(\theta)}\left\{\log p(\boldsymbol{y},\theta)\right\}\right).$$

See [23] and [45] for further details.

In general, the posterior independence assumption in the VB approximation can bias the point estimates and underestimate variability compared with the MCMC algorithm in which the true posterior is approximated with arbitrary precision. Our empirical findings in Section 4 also confirm that the MCMC results are consistently less biased than those obtained using VB. These drawbacks are not only confined to our application but also apply to VB approaches more generally (e.g., [48] and [43]). The error in the VB approximation depends on how well the assumed posterior distribution reflects reality. We refer interested readers to [44, 16, 8, 29] for technical details and further discussions.

3.2 Linear quantile model

For the case of Bayesian linear quantile regression, [31] considered the implementation of mean-field VB methods. However, the use of the horseshoe-type priors or outlier detections in combination with quantile regression was not considered. Mean-field methods for the horseshoe and other sparse signal shrinkage priors for mean regression were considered by [31] and [32]. Both the linear quantile regression model and the full horseshoe+ quantile OD model we have described allow efficient mean-field variational updates for all factors thanks to conditional conjugacy. In particular, consider the linear quantile regression model in Section 2.1 with the set of parameters denoted by $\Theta_1 = \left(\beta_0, \boldsymbol{\beta}, \sigma, \{z_i\}_{i=1}^n, \left\{\lambda_j^2, \zeta_j, \eta_j^2\right\}_{j=1}^p, \zeta_\eta\right)$ where the subscript in Θ_1 indicates that these parameters belong to the first model, namely the linear quantile model in Section 2.1, as opposed to the full horseshoe+ quantile OD model in Section 2.2.

Writing $\doteq$ to denote equality up to an additive constant, the log-posterior is given by

$$\begin{aligned}\ell(\Theta_1 \mid \boldsymbol{y}) + \log \pi(\Theta_1) \doteq & -\frac{n}{2}\log\sigma - \frac{1}{2}\sum_{i=1}^n \log z_i \\ & - \frac{1}{2\kappa_2^2\sigma}\sum_{i=1}^n \frac{(y_i - \beta_0 - \boldsymbol{x}_i^T\boldsymbol{\beta} - \kappa_1 z_i)^2}{z_i} \\ & - n\log\sigma - \frac{1}{\sigma}\sum_{i=1}^n z_i - \left(\frac{r_\sigma^0}{2}+1\right)\log\sigma \\ & - \frac{s_\sigma^0}{2\sigma} - \frac{1}{2}\sum_{j=1}^p \frac{\beta_j^2}{\lambda_j^2} - \frac{1}{2}\log\lambda_j^2 - \frac{(\beta_0-\mu_{\beta_0})^2}{2\sigma_{\beta_0}^2} \\ & + \sum_{j=1}^p -\frac{3}{2}\log\lambda_j^2 - \frac{1}{2\zeta_j\lambda_j^2} \\ & + \sum_{j=1}^p -\frac{3}{2}\log\zeta_j - \frac{1}{2(A_\beta\eta_j)^2\zeta_j} \\ & + \sum_{j=1}^p -\frac{3}{2}\log\eta_j^2 - \frac{1}{\zeta_\eta\eta_j^2} - \frac{3}{2}\log\zeta_\eta - \frac{1}{2\zeta_\eta}\end{aligned}$$

We decompose the variational posterior into

$$\begin{aligned}(8)\qquad q(\Theta_1) = & \; q_1(\beta_0)q_2(\boldsymbol{\beta})q_3(\sigma)\left[\prod_{i=1}^n q_4(z_i)\right] \\ & \times \left[\prod_{j=1}^p q_5(\lambda_j^2)q_6(\zeta_j)q_7(\eta_j^2)\right] q_8(\zeta_\eta).\end{aligned}$$

The optimal variational distributions based on mean-field VB are given as follows: $q_1(\beta_0)$ is a univariate normal distribution, parameterized as $N(\mu_{\beta_0}^q, \sigma_{\beta_0}^{2q})$; $q_2(\boldsymbol{\beta})$ is a multivariate normal distribution denoted as $N_p(\mu_\beta^q, \Sigma_\beta^q)$; $q_3(\sigma)$ is an inverse gamma distribution denoted as $\text{InvGam}(r_\sigma^q/2, s_\sigma^q/2)$; and $q_4(z_i)$ is a generalized inverse Gaussian (GIG) distribution parameterized as $\text{GIG}\left(\frac{1}{2}, \chi_{z_i}^q, \psi_z^q\right)$. Here, the superscript q is used for the variational parameters to avoid notational confusion with the parameters for their prior distributions given in Section 2. The remaining optimal variational distributions $q_5(\lambda_j^2)$, $q_6(\zeta_j)$, $q_7(\eta_j^2)$, $q_8(\zeta_\eta)$ are inverse gamma distributions denoted as $\text{InvGam}\,(1, \varpi_{1j})$, $\text{InvGam}\,(1, \varpi_{2j})$, $\text{InvGam}\,(1, \varpi_{3j})$, and $\text{InvGam}\left(\frac{p+1}{2}, \varpi_4\right)$, respectively. $\varpi_{1j}, \ldots, \varpi_{3j}$, $j = 1, \ldots, p$, and ϖ_4 denote the variational parameters to be updated in the VB algorithms. Specifically, Algorithm 1 provides a full description of the derived updates.

The variational lower bound consists of three components: the expectation of the log-likelihood, log-prior distributions, and variational entropies.

$$\begin{aligned}\text{E}(\ell(\Theta_1 \mid \boldsymbol{y}) + \pi(\Theta)) \doteq & \left(\frac{3n + r_0 + 2}{2}\right)\left(\psi\left(\frac{r_q}{2}\right) - \log\left(\frac{s_q}{2}\right)\right) \\ & - \frac{1}{2\kappa_2^2}\frac{r_\sigma^q}{s_\sigma^q}\left[\sum_{i=1}^n \kappa_1^2\,\text{E}\,(z_i) - 2\kappa_1\left(y_i - \mu_{\beta_0}^q - \boldsymbol{x}_i^T\mu_\beta^q\right)\right. \\ & \left. + \text{E}\left(\frac{1}{z_i}\right)\left(\left(y_i - \mu_{\beta_0}^q - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 + \boldsymbol{x}_i^T\Sigma_\beta^q\boldsymbol{x}_i + \sigma_{\beta_0}^{2q}\right)\right] \\ & - \frac{r_\sigma^q}{s_\sigma^q}\sum_{i=1}^n \text{E}\,(z_i) - \frac{s_\sigma^q r_\sigma^q}{2s_\sigma^q} - \sum_{j=1}^p\left[\frac{\mu_{\beta,j}^q + \Sigma_{\beta,jj}^q}{2\varpi_{1j}} + \frac{1}{2}\log\varpi_{1j}\right]\end{aligned}$$

Algorithm 1 VB algorithm for linear quantile regression

procedure VBLinMod($\boldsymbol{y}, \boldsymbol{X}, \tau$)
  Initialize variational parameters.
  $\text{LB} = -\infty$
  $\text{dif} = 1$
  for i in 1 : maxIter **do**
    $\sigma^{2q}_{\beta_0} \leftarrow \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \mu_{\frac{1}{z_i}} + \frac{1}{\sigma_0^2} \right)^{-1}$
    $\mu^q_{\beta_0} \leftarrow \sigma^{2q}_{\beta_0} \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \left(\sum_{i=1}^n \mu_{\frac{1}{z_i}} \left(y_i - \boldsymbol{x}_i^T \mu^q_\beta \right) - \kappa_1 \right) + \frac{\mu_0}{\sigma_0^2} \right)$
    $\Sigma^q_\beta \leftarrow \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \mu_{\frac{1}{z_i}} + \text{diag}\left(\frac{1}{\varpi_{i1}}, \ldots, \frac{1}{\varpi_{1p}} \right) \right)^{-1}$
    $\mu^q_\beta \leftarrow \Sigma^q_\beta \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \boldsymbol{x}_i \left(\mu_{\frac{1}{z_i}} \left(y_i - \mu^q_{\beta_0} \right) - \kappa_1 \right) \right)$
    $r^q_\sigma \leftarrow r^0_\sigma + 3n$
    $s^q_\sigma \leftarrow \frac{1}{\kappa_2^2} \Big[\sum_{i=1}^n \kappa_1^2 \,\text{E}(z_i) - 2\kappa_1 \left(y_i - \mu^q_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta \right)$
      $+ \mu_{\frac{1}{z_i}} \left(\left(y_i - \hat{\mu}_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta \right)^2 + \hat{\sigma}^2_{\beta_0} + \boldsymbol{x}_i^T \Sigma^q_\beta \boldsymbol{x}_i \right) \Big]$
      $+ 2 \sum_{i=1}^n \text{E}(z_i) + s^0_\sigma$
    $\chi^q_{z_i} \leftarrow \frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \left[\left(y_i - \hat{\mu}_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta \right)^2 + \hat{\sigma}^2_{\beta_0} + \boldsymbol{x}_i^T \Sigma^q_\beta \boldsymbol{x}_i \right]$
      for $i = 1, \ldots, n$
    $\psi^q_z \leftarrow \frac{r^q_\sigma}{s^q_\sigma} \left(2 + \frac{\kappa_1^2}{\kappa_2^2} \right)$
    $\varpi_{1j} \leftarrow \frac{\mu^p_{\beta,j}{}^2 + \Sigma_{\beta,jj}}{2} + \frac{1}{\varpi_{2j}}$ for $j = 1, \ldots, p$
    $\varpi_{2j} \leftarrow \frac{1}{\varpi_{1j}} + \frac{1}{A_\beta^2 \varpi_{3j}}$ for $j = 1, \ldots, p$
    $\varpi_{3j} \leftarrow \frac{1}{A_\beta^2 \varpi_{2j}} + \frac{p+1}{2\varpi_4}$ for $j = 1, \ldots, p$
    $\varpi_4 \leftarrow \sum_{j=1}^p \frac{1}{\varpi_{3j}} + 1$
    Compute LB_{new}.
    Compute $\text{dif} = \text{LB} - \text{LB}_{\text{new}}$
    if dif $<$ TOL **then**
      break
    else
      $\text{LB} \leftarrow \text{LB}_{\text{new}}$
  return $(\sigma^{2q}_{\beta_0}, \mu^q_{\beta_0}, \Sigma^q_\beta, \mu^q_\beta, r^q_\sigma, s^q_\sigma, \psi^q_z, \{\chi^q_{z_i}\}_{i=1}^n,$
    $\{\varpi_{1j}, \varpi_{2j}, \varpi_{3j}\}_{j=1}^p, \varpi_4)$

$$- \frac{3}{2} \sum_{j=1}^p \log(\varpi_{1j} \varpi_{2j} \varpi_{3j}) - \log \varpi_4 - \frac{p+1}{2\varpi_4}$$

The variational entropies are summed to be

$$\begin{aligned}
\text{H}(\Theta_1) \doteq\ & \frac{1}{2} \log\det \Sigma^q_\beta + \frac{r^q_\sigma}{2} + \log\left(\frac{s^q_\sigma}{2} \right) + \log\Gamma\left(\frac{r^q_\sigma}{2} \right) \\
& - \left(1 + \frac{r^q_\sigma}{2} \right) \psi\left(\frac{r^q_\sigma}{2} \right) + \frac{1}{2} \log \hat{\sigma}^2_{\beta_0} \\
& + \sum_{i=1}^n \Bigg[- \frac{1}{4} \left(\log\left(\psi^q \right) - \log\left(\chi^q_{z_i} \right) \right) \\
& + \log\left(2K_{1/2}\left(\sqrt{\psi^q_z \chi^q_{z_i}} \right) \right) + \frac{1}{2} \left(\chi_i \,\text{E}\left(\frac{1}{z_i} \right) + \psi^q \,\text{E}\left(z_i \right) \right) \Bigg] \\
& + \frac{1}{2} \sum_{j=1}^p \log(\varpi_{1j} \varpi_{2j} \varpi_{3j}) + \log \varpi_4
\end{aligned}$$

where $K_\nu(x)$ is the modified Bessel function of the second kind. Then, the lower bound becomes

$$\text{LB} = \text{E}\left(\ell(\Theta_1 \mid \boldsymbol{y}) + \log \pi(\Theta_1) \right) + \text{H}(\Theta_1)$$

3.3 Full horseshoe+ quantile OD model

In the full horseshoe+ quantile OD model, let the set of parameters be denoted by $\Theta_2 = \left(\Theta_1, \left\{ \gamma_i, \lambda^2_{\gamma_i}, \zeta_{\gamma_i}, \eta^2_{\gamma_i} \right\}_{i=1}^n \right)$. We decompose the variational posterior into

$$q(\Theta_2) = q(\Theta_1) q_9(\boldsymbol{\gamma}) \left[\prod_{i=1}^n q_{10}(\lambda^2_{\gamma_i}) q_{11}(\zeta_{\gamma_i}) q_{12}(\eta^2_{\gamma_i}) \right] \tag{9}$$

with $q(\Theta_1)$ defined as in (8). The optimal form of the additional factors in (9) are: $q_9(\boldsymbol{\gamma})$ is multivariate normal, $N_n(\mu^q_\gamma, \Sigma^q_\gamma)$; $q_{10}(\lambda^2_{\gamma_i})$, $q_{11}(\zeta_{\gamma_i})$, $q_{12}(\eta^2_{\gamma_i}), q_{13}(\zeta_{\eta,\gamma})$ are inverse gamma, $\text{InvGam}\left(1, \varpi^\gamma_{1i}\right)$, $\text{InvGam}\left(1, \varpi^\gamma_{2i}\right)$, $\text{InvGam}\left(1, \varpi^\gamma_{3i}\right)$, $\text{InvGam}\left(\frac{n+1}{2}, \varpi^\gamma_4 \right)$, respectively. Algorithm 2 describes the coordinate-ascent algorithm on the updates of all the variational parameters Θ_2. Algorithm 2 describes the derived updates.

4. EMPIRICAL STUDY

4.1 High-dimensional linear quantile regression

In this section, we examine the performance of VB in linear quantile regression for different quantile levels. We generate data $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, where $n = 100$ and $p = 300$ by $\boldsymbol{x}_{ij} \overset{\text{iid}}{\sim} \text{Unif}(0, 1)$ and $\boldsymbol{\beta} = (5\mathbf{1}_{10}, \mathbf{0}_{p-10})^T$. The response variables are obtained through $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$.

Since quantile regression does not reflect the true data generating process but rather attempts to capture quantile information using the check loss function, we assess the performance under various error distributions. We refer to the simulation setting of [12] for the error distribution, which varies from a single component normal distribution to non-normal multimodal distributions, yet has at least close-to-zero or exactly zero medians.

- Normal: $\epsilon_i \overset{\text{iid}}{\sim} N(0, 1)$
- Skewed: $\epsilon_i \overset{\text{iid}}{\sim} 0.2N(-0.88, 1) + 0.2N\left(-0.392, 1.5^2\right) + 0.6N\left(0.196, \frac{5}{9}^2\right)$
- Kurtotic: $\epsilon_i \overset{\text{iid}}{\sim} \frac{2}{3}N(0, 1) + \frac{1}{3}N\left(0, 0.1^2\right)$
- Bimodal: $\epsilon_i \overset{\text{iid}}{\sim} 0.5N\left(-1, \frac{2}{3}^2\right) + 0.5N\left(1, \frac{2}{3}^2\right)$
- Bimodal with separate modes: $\epsilon_i \overset{\text{iid}}{\sim} 0.5N\left(-1.5, 0.5^2\right) + 0.5N\left(1.5, 0.5^2\right)$
- Skewed bimodal: $\epsilon_i \overset{\text{iid}}{\sim} 0.75N(-0.43, 1) + 0.25N\left(1.07, \frac{1}{3}^2\right)$
- Trimodal: $\epsilon_i \overset{\text{iid}}{\sim} 0.45N\left(-1.2, 0.6^2\right) + 0.45N\left(1.2, 0.6^2\right) + 0.1N\left(0, 0.25^2\right)$

Algorithm 2 VB algorithm for the full quantile OD model

procedure VBFULL($\boldsymbol{y}, \boldsymbol{X}, \tau$)
  Initialize variational parameters.
  $\text{LB} = -\infty$
  $\text{dif} = 1$
  for *iter* in 1 : maxIter **do**
    $\sigma^{2q}_{\beta_0} \leftarrow \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \mu_{\frac{1}{z_i}} + \frac{1}{\sigma_0^2} \right)^{-1}$
    $\mu^q_{\beta_0} \leftarrow \hat{\sigma}^2_{\beta_0} \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \left(\sum_{i=1}^n \mu_{\frac{1}{z_i}} (y_i - \boldsymbol{x}_i^T \mu_\beta - \mu_{\gamma_i}) - \kappa_1 \right) + \frac{\mu_0}{\sigma_0^2} \right)$
    $\Sigma^q_\beta \leftarrow \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \mu_{\frac{1}{z_i}} + \text{diag}\left(\frac{1}{\varpi_{i1}}, \ldots, \frac{1}{\varpi_{1p}} \right) \right)^{-1}$
    $\mu^q_\beta \leftarrow \Sigma^q_\beta \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \sum_{i=1}^n \boldsymbol{x}_i \left(\mu_{\frac{1}{z_i}} \left(y_i - \mu^q_{\beta_0} - \mu_{\gamma_i} \right) - \kappa_1 \right) \right)$
    $r^q_\sigma \leftarrow r^0_\sigma + 3n$
    $s^q_\sigma \leftarrow \frac{1}{\kappa_2^2} \left[\sum_{i=1}^n \kappa_1^2 \, \text{E}(z_i) - 2\kappa_1 \left(y_i - \mu^q_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta - \mu^q_{\gamma_i} \right) + \mu_{\frac{1}{z_i}} \left(\left(y_i - \mu^q_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta - \mu^q_{\gamma_i} \right)^2 + \sigma^{2q}_{\beta_0} + \boldsymbol{x}_i^T \Sigma^q_\beta \boldsymbol{x}_i + \Sigma^q_{\gamma,ii} \right) \right] + 2 \sum_{i=1}^n \text{E}(z_i) + s^0_\sigma$
    $\chi^q_{z_i} \leftarrow \frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \left[\left(y_i - \mu^q_{\beta_0} - \boldsymbol{x}_i^T \mu^q_\beta - \mu^q_{\gamma_i} \right)^2 + \sigma^{2q}_{\beta_0} + \boldsymbol{x}_i^T \Sigma^q_\beta \boldsymbol{x}_i + \Sigma^q_{\gamma,ii} \right]$ for $i = 1, \ldots, n$
    $\psi^q_z \leftarrow \frac{r^q_\sigma}{s^q_\sigma} \left(2 + \frac{\kappa_1^2}{\kappa_2^2} \right)$
    $\Sigma^q_\gamma \leftarrow \left(\frac{r^q_\sigma}{\kappa_2^2 s^q_\sigma} \text{diag}\left(\left\{ \mu_{\frac{1}{z_i}} \right\}_{i=1}^n \right) + \text{diag}\left(\left\{ \frac{1}{\varpi^\gamma_{1i}} \right\}_{i=1}^n \right) \right)^{-1}$
    $\mu^q_\gamma \leftarrow \Sigma^q_\gamma \left(\text{diag}\left(\left\{ \mu_{\frac{1}{z_i}} \right\}_{i=1}^n \right) \left(\boldsymbol{y} - \mu^q_{\beta_0} \mathbf{1}_n - \boldsymbol{X} \mu^q_\beta \right) - \kappa_1 \right)$
    $\varpi^\gamma_{1i} \leftarrow \frac{\mu^q_{\gamma,i} + \Sigma^q_{\gamma,ii}}{2} + \frac{1}{\varpi_{1i}}$ for $i = 1, \ldots, n$
    $\varpi^\gamma_{2i} \leftarrow \frac{1}{\varpi^\gamma_{1i}} + \frac{1}{A^2_\gamma \varpi_{3i}}$ for $i = 1, \ldots, n$
    $\varpi^\gamma_{3i} \leftarrow \frac{1}{A^2_\gamma \varpi^\gamma_{2i}} + \frac{n+1}{2\varpi^\gamma_4}$ for $i = 1, \ldots, n$
    $\varpi^\gamma_4 \leftarrow \sum_{i=1}^n \frac{1}{\varpi^\gamma_{3i}} + 1$
    $\varpi_{1j} \leftarrow \frac{{\mu^p_{\beta,j}}^2 + \Sigma_{\beta,jj}}{2} + \frac{1}{\varpi_{2j}}$ for $j = 1, \ldots, p$
    $\varpi_{2j} \leftarrow \frac{1}{\varpi_{1j}} + \frac{1}{A^2_\beta \varpi_{3j}}$ for $j = 1, \ldots, p$
    $\varpi_{3j} \leftarrow \frac{1}{A^2_\beta \varpi_{2j}} + \frac{p+1}{2\varpi_4}$ for $j = 1, \ldots, p$
    $\varpi_4 \leftarrow \sum_{j=1}^p \frac{1}{\varpi_{3j}} + 1$
    Compute LB_{new}.
    Compute $\text{dif} = \text{LB} - \text{LB}_{\text{new}}$
    if dif < TOL **then**
      break
    else
      $\text{LB} \leftarrow \text{LB}_{\text{new}}$
  return $(\sigma^{2q}_{\beta_0}, \mu^q_{\beta_0}, \Sigma^q_\beta, \mu^q_\beta, r^q_\sigma, s^q_\sigma, \Sigma^q_\gamma, \mu^q_\gamma, \psi^q_z, \{\chi^q_{z_i}\}_{i=1}^n, \{\varpi^\gamma_{1i}, \varpi^\gamma_{2i}, \varpi^\gamma_{3i}\}_{i=1}^n, \varpi^\gamma_4, \{\varpi_{1j}, \varpi_{2j}, \varpi_{3j}\}_{j=1}^p, \varpi_4)$

- Cauchy: $\epsilon_i \overset{\text{iid}}{\sim} \text{C}(0, 1)$

We measure the accuracy of the estimation using absolute deviation error (AD) and average check loss (ACL), which are defined by

$$\text{AD} = \frac{1}{n} \sum_{i=1}^n \left| \boldsymbol{x}_i^T \boldsymbol{\beta} - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \right|,$$

$$\text{ACL}_\tau = \frac{1}{n} \sum_{i=1}^n \rho_\tau(\boldsymbol{x}_i^T \boldsymbol{\beta} - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}), \quad \rho_\tau(x) = x(\tau - I(x < 0)).$$

Further, we measure the shrinkage effects using two variable selection metrics, the true positive rate (TPR) and false positive rate (FPR). Although these variable selection measures are not directly applicable to our methods for the shrinkage priors, we provide them to illustrate the high-dimensional variable selection with horseshoe+ priors that shrink the coefficients to near zero. Since our shrinkage priors do not force the coefficient to zero, we instead check if the true value is included in the 95% credible interval(CI) for each coefficient.

Our numerical implementation of all the proposed methods using horseshoe+priors with both VB (VBQR-HS) and MCMC (MCQR-HS) is written in `R`, and the MCMC algorithms for MCQR-HS and the one with outlier detection are given in the Appendix. All the simulations presented in Section 4 are performed on a PC with an Intel i5-3470 CPU. In the implementation of VBQR-HS, we set the tolerance level to 10^{-5} and the hyperparameters for the priors are set as $\sigma_0^2 = 10$, $A_\beta = 10^{-2}$, $r^0_\sigma = 4$, and $s^0_\sigma = 1$. The convergence of the VB algorithm is monitored by the relative increase in the lower bound in each iteration. For MCQR-HS, we discard the initial 5000 iterations and collect 10000 posterior samples without thinning.

Table 1 shows the performance measures for VBQR-HS and MCQR-HS and various error distributions for the 500 simulation iterations. VBQR-HS yields valid estimates comparable to those of MCQR-HS on a reasonable quantile interval from 0.3 to 0.7 but with some instability at extreme quantile levels $\tau = 0.1$ and 0.9. However, such instabilities at extreme quantile levels are expected, as the number of observations in the tails is small (e.g., [38, 49, 13]) and the AL distribution does not impose certain structures on tail fatness by construction (e.g, [52, 4]). Aside from accuracy, VB is shown to be numerically stable and its statistical performance is robust to the assumption of different error distributions. One notable exception is the Cauchy error distribution whose expected value is infinity, and the heavy tailed behavior can lead to an erratic estimation and low accuracy, which is observed in both MCQR-HS and VBQR-HS. This requires an additional exploration beyond the scope of the current paper. As a caveat, the CI method for checking variable selection is penalized for VBQR-HS since the uncertainty of the VB approximation is known to be underestimated, resulting in narrower intervals. The averaged elapsed time of VBQR-HS was an order of magnitude faster than that of MCQR-HS, (e.g., 6.40 seconds for VBQR-HS and 72.51 seconds for MCQR-HS with 100 simulation iterations), even though the MCMC algorithm was implemented in such a way that the time complexity of sampling $\boldsymbol{\beta}$ is reduced from $O(p^3)$ to $O(np^2)$ when $p \gg n$ [6]. This finding also validates the proposed VB approximation as an attractive alternative to the MCMC approach in high-dimensional applications where computation time is critical.

Table 1. The performance measures of VBQR-HS and MCQR-HS (in parentheses) averaged over the 500 simulation iterations under various error distributions

	τ	AD	ACL	TPR	FPR		τ	AD	ACL	TPR	FPR
Normal	0.1	1.71(0.96)	0.18(0.10)	0.37(0.88)	0.00(0.00)	Bimod SepMod	0.1	2.62(1.75)	0.28(0.18)	0.44(0.92)	0.00(0.00)
	0.3	0.61(0.55)	0.20(0.18)	0.57(0.91)	0.00(0.00)		0.3	1.23(1.04)	0.41(0.34)	0.54(0.90)	0.01(0.00)
	0.5	0.38(0.34)	0.19(0.17)	0.60(0.93)	0.01(0.00)		0.5	0.96(0.72)	0.48(0.36)	0.47(0.89)	0.02(0.00)
	0.7	0.60(0.55)	0.19(0.18)	0.58(0.92)	0.00(0.00)		0.7	1.22(1.04)	0.41(0.34)	0.54(0.90)	0.01(0.00)
	0.9	1.56(0.96)	0.16(0.10)	0.37(0.87)	0.00(0.00)		0.9	2.32(1.74)	0.25(0.18)	0.51(0.92)	0.00(0.00)
Skewed	0.1	1.98(1.10)	0.21(0.12)	0.32(0.85)	0.00(0.00)	Skewed Bimodal	0.1	1.99(1.14)	0.21(0.12)	0.36(0.85)	0.00(0.00)
	0.3	0.63(0.61)	0.20(0.19)	0.57(0.91)	0.00(0.00)		0.3	0.75(0.67)	0.25(0.22)	0.53(0.89)	0.00(0.00)
	0.5	0.32(0.29)	0.16(0.14)	0.66(0.94)	0.01(0.00)		0.5	0.51(0.42)	0.25(0.21)	0.56(0.91)	0.01(0.00)
	0.7	0.46(0.41)	0.15(0.13)	0.66(0.94)	0.00(0.00)		0.7	0.69(0.61)	0.22(0.20)	0.59(0.92)	0.00(0.00)
	0.9	1.28(0.79)	0.14(0.09)	0.41(0.89)	0.00(0.00)		0.9	1.56(1.04)	0.17(0.11)	0.47(0.91)	0.00(0.00)
Kurtotic	0.1	1.50(0.76)	0.16(0.08)	0.33(0.87)	0.00(0.00)	Trimodal	0.1	2.22(1.32)	0.24(0.14)	0.40(0.89)	0.00(0.00)
	0.3	0.40(0.39)	0.13(0.13)	0.64(0.93)	0.00(0.00)		0.3	0.89(0.79)	0.29(0.26)	0.51(0.90)	0.01(0.00)
	0.5	0.19(0.20)	0.10(0.10)	0.77(0.96)	0.00(0.00)		0.5	0.67(0.52)	0.33(0.26)	0.50(0.90)	0.01(0.00)
	0.7	0.41(0.40)	0.13(0.13)	0.63(0.93)	0.00(0.00)		0.7	0.90(0.78)	0.30(0.25)	0.53(0.90)	0.01(0.00)
	0.9	1.37(0.76)	0.15(0.08)	0.33(0.87)	0.00(0.00)		0.9	1.96(1.31)	0.21(0.14)	0.42(0.88)	0.00(0.00)
Bimodal	0.1	2.00(1.23)	0.21(0.13)	0.40(0.88)	0.00(0.00)	Cauchy	0.1	11.44(8.25)	1.33(0.85)	0.03(0.22)	0.02(0.00)
	0.3	0.83(0.73)	0.27(0.23)	0.54(0.90)	0.00(0.00)		0.3	5.37(3.83)	1.82(1.26)	0.15(0.49)	0.02(0.00)
	0.5	0.60(0.48)	0.30(0.24)	0.53(0.90)	0.01(0.00)		0.5	4.47(2.44)	2.24(1.22)	0.16(0.62)	0.02(0.00)
	0.7	0.82(0.72)	0.27(0.24)	0.55(0.91)	0.00(0.00)		0.7	4.87(2.93)	1.90(1.06)	0.15(0.61)	0.01(0.00)
	0.9	1.79(1.22)	0.19(0.13)	0.44(0.89)	0.00(0.00)		0.9	11.47(7.76)	1.46(0.82)	0.04(0.31)	0.02(0.00)

Table 2. The average number of outliers found after 10 iterations for each true number of outliers $O \in \{5, 10, 20\}$ and quantile level $\tau = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ when $p > n$

	Number of Outliers				Number of Outliers		
τ	5	10	20	τ	5	10	20
0.1	0	0	0	0.1	0	0	0.1
0.3	5.0	9.7	3.6	0.3	4.6	10.0	2.4
0.5	5.0	10.0	7.6	0.5	4.9	10.0	6.2
0.7	5.0	8.8	5.8	0.7	5.0	8.6	6.6
0.9	4.7	7.9	3.1	0.9	2.5	7.2	2.5

(a) $p = 100,\ n = 50$ (b) $p = 200,\ n = 50$

4.2 Outlier detection in linear quantile regression

In this section, we evaluate the outlier detection performance of the proposed horseshoe+ quantile OD model in comparison with the model studied by [54]. For clarity, we denote the model of [54] as the *mean OD model*. The simulated data was generated by

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_i + \epsilon_i, \quad \epsilon_i \sim N(0, 1.1^2) \tag{10}$$

where $x_{ij} \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, $\beta_j \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, $\gamma_i = 10$ for O (the given number of outliers) randomly selected observations and $\gamma_i = 0$ otherwise.

We consider $n = 50$ with $p = 100$ and $p = 200$. Simulations are conducted with every combination of the true number of outliers $O \in \{5, 10, 20\}$ and quantile levels $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The initialization mechanism is modified accordingly to robustly single out the signals in incorporating the ultra-sparsity [27, 54].

Table 2 describes the simulation results under high-dimensional circumstances. It is observed that, aside from the extreme quantiles, the performance of the quantile outlier detection model is very stable even in high-dimensional regression. The low detection rate for $O = 20$ is not due to the vulnerability of the model itself but rather stems from the low signal-to-noise ratio. Since outliers account for nearly half the data, the fixed component of the conditional regression function, $\boldsymbol{x}^T\boldsymbol{\beta}$, shifts accordingly to accommodate the variation.

The performance of the quantile OD model stands out for non-normal error distributions. Specifically, we generate the simulated data with an $\text{AL}_\tau(1)$ error distribution:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_i + \epsilon_i, \quad \epsilon_i \sim \text{AL}_\tau(1) \tag{11}$$

where τ now affects the data generation and the other true parameters are equal to that of the normal error model. The dimension for the data set is set to $n = 1000$ and $p = 15$ following [54]. Table 3 demonstrates that the quantile OD model uniformly outperforms the mean OD model for quantile levels $0.3 \sim 0.9$. Moreover, the quantile OD model performs worse for the extreme quantile level 0.1. In general, this result implies that the quantile OD model captures outliers in a more robust fashion compared to the mean OD model.

Table 3. Average number of outliers detected with the quantile OD model and mean OD model (in parentheses) under $\text{AL}_\tau(1)$ data

	Number of Outliers			
τ	10	20	50	100
0.1	0.3 (0.0)	0 (1.0)	0 (2.3)	0 (12.0)
0.3	10 (0.3)	20.0 (0.0)	50.0 (2.7)	100 (13.7)
0.5	10 (0.0)	19.7 (0.0)	49.7 (1.3)	99.7 (7.70)
0.7	9.7 (0.0)	20.0 (1.3)	49.3 (1.3)	98 (9.70)
0.9	8.3 (0.7)	17.3 (1.3)	40.0 (0.7)	80 (9.70)

5. REAL DATA APPLICATIONS

5.1 Boston housing data

The Boston housing data have been used as a benchmark real data application for variable selection in quantile regression [21, 56]. We use the dataset available from the `R` package `mlbench`, of which the corrected median value of owner-occupied homes in USD 1000s (`cmedv`) is the response variable and the remaining 15 variables are included as covariates after standardization.

We evaluate the performance of VBQR-HS together with MCQR-HS and adopt the mean-weighted absolute residuals considered in [56] in comparison with the QR-SSVS algorithm of [56]. Here, the mean-weighted absolute residual corresponds to the ACL between the observed y_i and estimated $\hat{y}_i$, i.e., $\frac{1}{s_k}\sum_{i=1}^{s_k}\rho_\tau(y_i-\hat{y}_i)$, where s_k is the size of the kth validation set for the 10-fold cross validation. Table 4 reports these estimates. The QR-SSVS based on MCMC is reported to have taken approximately 17 seconds for each quantile level with the whole Boston housing data whereas our VBQR-HS is completed in a mere 4.2 seconds. In terms of the predictive accuracy, the VB approach makes a small sacrifice for a tremendous gain in computation time. Note that the estimates of QR-SSVS M.A.(model averaged) in Table 4 are those given in Table 7 of [56].

Table 4. Mean weighted absolute deviations for 10-fold cross validation

τ	VBQR-HS	MCQR-HS	QR-SSVS M.A.
0.05	0.351	0.309	0.306
0.1	0.561	0.560	0.550
0.25	1.094	1.103	1.087
0.5	1.631	1.654	1.576
0.75	1.556	1.590	1.521
0.9	1.043	1.084	1.026
0.95	0.712	0.693	0.670

Table 5 shows the estimated regression coefficients of the VBQR-HS with three quantile levels in comparison with the other two MCMC methods. The values in parentheses in Table 5 are the 95% CI for each parameter. MCQR-HS indicates the posterior mean and corresponding 95% CI of the variable written above, based on the same model as VBQR-HS but estimated via MCMC, whereas QR-SSVS indicates those from the stochastic search variable selection algorithm of [56] using the R function `ssvsquantreg` available in the R package `MCMCpack`. We set $A_\beta = 0.005$, while all else remains identical to the setting in the simulation study. The results are, to a large degree, consistent in terms of model selection with those of MCQR-HS and QR-SSVS, which indicates that the proposed VBQR-HS mimics the behavior of the MCMC algorithms well, as expected.

Table 5. Regression coefficients of VBQR-HS.

	Quantile level (τ)		
	0.05	**0.5**	**0.95**
`lon`	−0.595(−0.597, −0.591)	−0.619(−0.640, −0.597)	0.000(−0.000, 0.000)
(MCQR-HS)	−0.792(−1.037, −0.565)	−0.589(−1.064, 0.001)	−0.002(−0.167, 0.156)
(QR-SSVS)	−0.712(−1.305, −0.063)	−0.565(−0.953, −0.170)	−0.034(−0.798, 0.643)
`lat`	0.242(0.239, 0.244)	0.003(0.002, 0.004)	0.000(−0.000, 0.001)
(MCQR-HS)	0.119(−0.022, 0.377)	0.030(−0.045, 0.258)	0.003(−0.138, 0.168)
(QR-SSVS)	0.152(−0.218, 0.638)	0.195(−0.058, 0.515)	0.329(−0.485, 1.308)
`crim`	−0.617(−0.627, −0.608)	−0.357(−0.389, −0.314)	−0.000(−0.001, −0.000)
(MCQR-HS)	−0.716(−1.273, −0.368)	−0.264(−1.168, −0.044)	−0.427(−2.336, 0.083)
(QR-SSVS)	−0.993(−2.317, −0.22)	−0.906(−1.391, −0.271)	−0.817(−2.730, 0.994)
`zn`	−0.002(−0.002, −0.001)	0.664(0.630, 0.696)	0.001(0.001, 0.001)
(MCQR-HS)	0.021(−0.070, 0.276)	0.265(−0.039, 0.997)	0.109(−0.063, 0.960)
(QR-SSVS)	0.032(−0.742, 0.764)	0.721(0.197, 1.207)	0.820(−0.083, 1.967)
`indus`	0.003(0.002, 0.003)	−0.001(−0.001, −0.000)	0.000(−0.000, 0.001)
(MCQR-HS)	0.032(−0.080, 0.322)	−0.042(−0.519, 0.089)	−0.002(−0.063, 0.960)
(QR-SSVS)	0.058(−0.846, 0.952)	−0.068(−0.569, 0.348)	0.009(−1.345, 1.627)
`chas`	0.000(0.000, 0.001)	−0.001(−0.001, −0.000)	−0.000(−0.001, −0.000)
(MCQR-HS)	0.100(−0.121, 0.862)	−0.029(−0.590, 0.214)	−0.246(−1.712, 0.071)
(QR-SSVS)	0.049(−2.113, 2.403)	−1.044(−2.228, 0.025)	−3.540(−9.071, 0.000)
`nox`	−0.001(−0.001, 0.000)	−0.698(−0.750, −0.639)	−2.626(−2.684, −2.562)
(MCQR-HS)	−0.259(−0.862, 0.032)	−0.310(−1.415, 0.050)	−2.635(−3.604, −1.588)
(QR-SSVS)	−0.461(−1.665, 0.462)	−0.654(−1.322, 0.000)	−1.791(−3.856, 0.000)
`rm`	1.713(1.704, 1.723)	3.373(3.328, 3.423)	3.421(3.402, 3.453)
(MCQR-HS)	1.793(1.344, 2.218)	3.736(3.023, 4.441)	3.517(2.875, 4.220)
(QR-SSVS)	1.801(0.875, 2.787)	3.541(2.901, 4.197)	3.694(2.664, 4.750)
`age`	−0.003(−0.003, −0.002)	−0.002(−0.003, 0.002)	0.000(−0.000, 0.001)
(MCQR-HS)	−0.492(−1.015, 0.005)	−0.276(−1.103, 0.039)	−0.002(−0.167, 0.155)
(QR-SSVS)	−0.463(−1.382, 0.257)	−0.606(−1.166, −0.007)	−0.058(−1.343, 1.241)
`dis`	−0.006(−0.007, −0.006)	−1.442(−1.491, −1.389)	−2.563(−2.599, −2.527)
(MCQR-HS)	−0.985(−1.548, −0.447)	−1.156(−1.880, −0.354)	−2.674(−3.599, −1.892)
(QR-SSVS)	−1.019(−2.214, 0.002)	−1.788(−2.415, −1.151)	−2.985(−4.548, −1.505)
`rad`	0.000(0.000, 0.001)	1.243(1.144, 1.334)	5.328(5.279, 5.376)
(MCQR-HS)	0.141(−0.073, 1.063)	0.358(−0.063, 1.723)	5.481(4.121, 7.347)
(QR-SSVS)	0.480(−0.658, 1.965)	1.494(0.589, 2.362)	5.793(3.187, 8.263)
`tax`	−1.619(−1.631, −1.606)	−1.977(−2.069, −1.885)	−0.000(−0.001, 0.000)
(MCQR-HS)	−2.260(−2.896, −1.723)	−1.171(−2.452, 0.003)	−0.095(−1.291, 0.129)
(QR-SSVS)	−2.153(−3.618, −0.739)	−1.912(−2.728, −0.985)	−1.187(−3.481, 0.674)
`ptratio`	−0.499(−0.504, −0.494)	−1.478(−1.501, −1.452)	−3.415(−3.438, −3.383)
(MCQR-HS)	−0.559(−0.928, −0.197)	−1.444(−2.032, −0.955)	−3.424(−4.168, −2.517)
(QR-SSVS)	−0.615(−1.327, 0.045)	−1.433(−1.838, −1.022)	−2.503(−3.985, −0.978)
`b`	0.451(0.447, 0.455)	1.120(1.100, 1.140)	1.359(1.308, 1.399)
(MCQR-HS)	0.573(0.322, 0.832)	1.185(0.812, 1.559)	0.725(−0.028, 1.962)
(QR-SSVS)	0.572(0.000, 1.198)	1.107(0.757, 1.468)	0.737(−0.564, 2.449)
`lstat`	−3.209(−3.223, −3.195)	−2.695(−2.749, −2.647)	−4.169(−4.194, −4.135)
(MCQR-HS)	−2.571(−3.182, −1.974)	−2.612(−3.384, −1.842)	−4.011(−4.688, −3.218)
(QR-SSVS)	−2.705(−3.900, −1.545)	−2.283(−2.971, −1.605)	−3.728(−4.893, −2.538)

Table 6. TPRs for identifying noise variables with the VBQR-HS, MCQR-HS, and QR-SSVS methods

	VBQR-HS	MCQR-HS	QR-SSVS
		$\tau = 0.05$	
Uncorrelated	0.990	0.632	0.917
Correlated	0.972	0.947	1.000
		$\tau = 0.5$	
Uncorrelated	0.989	0.719	0.804
Correlated	0.993	0.997	0.999
		$\tau = 0.95$	
Uncorrelated	0.996	0.544	0.536
Correlated	0.942	0.609	0.998

Further, we consider a high-dimensional analysis for the performance of the proposed VBQR-HS algorithm using a dataset obtained from the Boston housing data by augmenting the original 15 predictors with some noise predictors. Cases with a large number of noise predictors are considered, resulting in a genuinely high-dimensional setting. Then, we explore the assumed model to examine whether the proposed VBQR-HS successfully filters out the randomly generated predictors in comparison with MCQR-HS and QR-

SSVS. For each number of noise predictors, we consider both scenarios where the noises are correlated and uncorrelated. The uncorrelated noises are generated from $N(0,1)$, whereas the correlated ones are generated from a multivariate normal distribution with zero mean and a covariance matrix sampled from the Wishart distribution with the degrees of freedom as the number of noise predictors plus one and the scale matrix as a Gram matrix $\boldsymbol{X}^T\boldsymbol{X}$, where $X_{ij} \sim N(0,1)$. Specifically, we include 1000 additional noise variables, and thus the estimation problem becomes the so called "large p, small n" ($p = 1015 > n = 506$) situation. The non-zero coefficient estimates of the original 15 variables turn out to have little difference from those in Table 5, with the vast majority of the coefficient estimates for the noise predictors having magnitudes of or below 10^{-4}. Table 6 provides the TPRs for their coefficient estimates assuming that the true values are zero based on both the uncorrelated and the correlated noise variables, and Figure 1 shows the posterior mean estimates for all the coefficients with their 95% CIs based on the uncorrelated noise variables for $\tau = 0.5$ by way of an illustration. As shown in Table 6 and Figure 1, the proposed VBQR-HS also works well for variable selection with the noise variables in terms of its accuracy with the TPR. Specifically, by forcing the regression coefficients to zero if their magnitudes are smaller than 0.001 adjusted with $A_\beta = 0.0001$ for strong sparsity, 99% of the noise variables are correctly identified by VBQR-HS for the uncorrelated noise variables, considerably better than MCQR-HS and QR-SSVS, whereas the variable selection accuracy of VBQR-HS is comparable to or slightly worse than that of QR-SSVS.

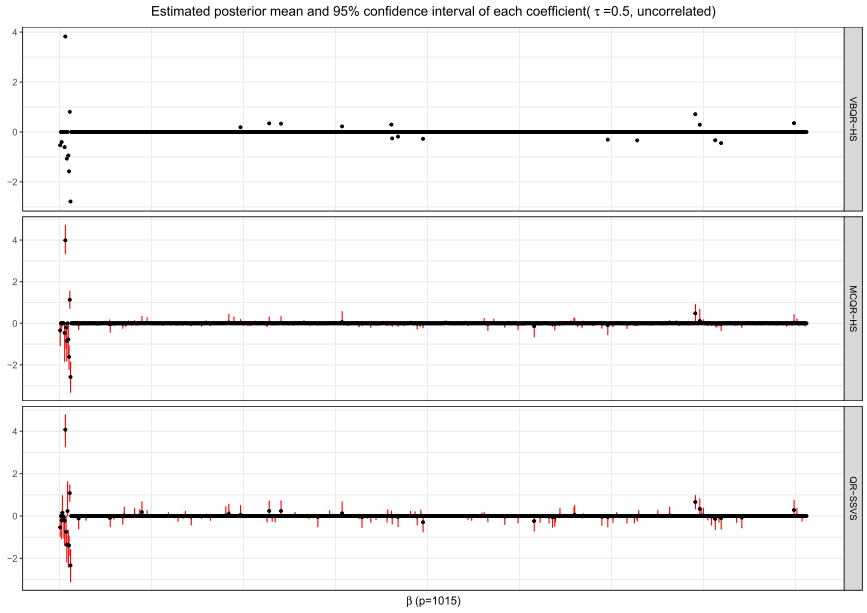

Figure 1. Regression coefficients of the augmented Boston housing data with 1000 uncorrelated noise predictors for $\tau = 0.5$.

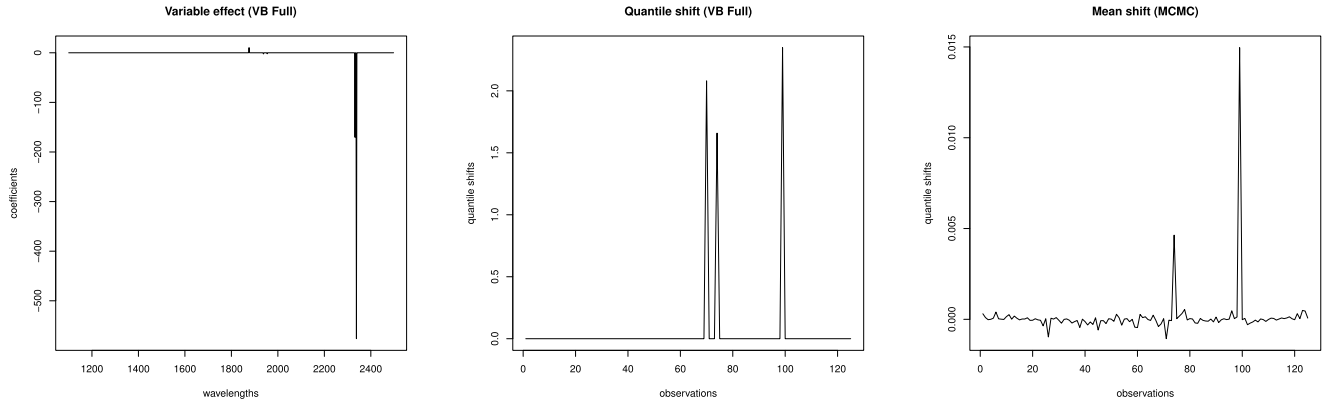

Figure 2. The estimated coefficients and quantile shifts of the sugar data. Left: Coefficients of the quantile, Middle: Quantile shift obtained from the VB full model, and Right: Quantile shift obtained from MCMC.

5.2 Sugar data for outlier detection

We next analyzed the sugar data in [10] using the VB full horseshoe+ quantile OD model. The data consist of infrared spectrograph data of three types of sugar: sucrose, glucose, and fructose whose absorbances in aqueous solutions with differing levels of concentration are of interest. The training set involves 125 observations on 700 wavelengths ranging from 1100 to 2498 nanometers. Instead of performing the mean regression to detect outliers, we adopt the median regression, $\tau = 0.5$. We follow the settings of [54] which utilized a history matching process [50] for prior elicitation where $A_\beta = 0.000013$ and $A_\gamma = 0.000045$. As has been explained in the full model, we use the robust initialization with `rlm` in the R package `MASS`. For the quantile model parameters such as σ, their hyperparameters were selected following [22].

[54] reported the 74th and 99th observations to be outliers in this example using the mean OD model. The VB full model identified 3 outliers, namely the 70th, 74th, and 99th observations from the quantile shift parameters, whereas only the 74th and 99th observations are considered to be outliers in the case of MCMC which coincides with the results of [54].

6. DISCUSSION

In this paper, we proposed two flexible Bayesian quantile regression models that can address model misspecification in particular for high-dimensional data. The first is a Bayesian quantile regression approach with variable selection using the horseshoe+ prior on the high-dimensional regression coefficients. The second is a robust quantile regression model incorporating outlier detection. For the implementation, we have developed computationally efficient methodology in terms of variational Bayes approximation as a faster alternative to conventional Bayesian methods with a substantially reduced computation time.

The proposed models have limitations and can potentially be improved. For both models, we observed poor performance at extreme quantile levels. To alleviate the worse tail behavior for the AL distribution, alternative distributions have recently been proposed for quantile regression models, such as the skew-normal distribution, skew-exponential power distribution [4], generalized AL distribution [55], and power-exponential distribution [24]. Our models may adopt these alternative distributions to improve the tail behavior. For more flexible modeling, it is natural to consider the semiparametric counterparts of the proposed model, but we need to deal with the complicated geometry of the parameter space, requiring further regularization on nonparametric components, in particular for outlier detection. Following [33], the method may be extended to various shape restrictions including monotone and convex/concave restrictions, with the computations in these cases performed using the non-conjugate variational message passing algorithm. In addition, it would be interesting to see how well our methodology scales up to real applications involving ultra-high dimensional and large sample sizes. For example, variational approaches with stochastic gradient algorithms and subsampling can be considered for large datasets [42]. A more detailed discussion of these topics will be left for future work.

APPENDIX A. MCMC ALGORITHM FOR LINEAR QUANTILE REGRESSION

- $(\beta_0 \mid -) \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$ where

$$\sigma^2_{\beta_0} = \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^n \frac{1}{z_i} + \frac{1}{\sigma_0^2} \right)^{-1}$$
$$\mu_{\beta_0} = \sigma^2_{\beta_0} \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^n \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - k_1 z_i}{z_i} + \frac{\mu_0}{\sigma_0^2} \right)$$

- $(\boldsymbol{\beta} \mid -) \sim N(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ where

$$\boldsymbol{\Sigma}_b = \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^n \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{z_i} + \text{diag}(\lambda_1^{-1}, \ldots, \lambda_p^{-1}) \right)^{-1}$$
$$\boldsymbol{\mu}_b = \boldsymbol{\Sigma}_b \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^n \frac{\boldsymbol{x}_i^T (y_i - \beta_0 - k_1 z_i)}{z_i} \right)$$

- $(z_i \mid -) \sim \text{GIG}\left(0.5, \frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{k_2^2 \sigma}, (\frac{k_1^2}{k_2^2} + 2)/(2\sigma)\right)$ for $i = 1, \ldots, n$
- $(\sigma \mid -) \sim \text{InvGam}\left(\frac{r_\sigma^0 + 3n}{2}, s_\sigma^n\right)$ where

$$s_\sigma^n = s_\sigma^0 + \sum_{i=1}^n \frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta} - k_1 z_i)^2}{k_2^2 z_i} + 2 \sum_{i=1}^n z_i$$

- For $j = 1, \ldots, p$

$$(\lambda_j^2 \mid -) \sim \text{InvGam}\left(1, \frac{\beta_j^2}{2} + \frac{1}{\zeta_j}\right)$$
$$(\zeta_j \mid -) \sim \text{InvGam}\left(1, \frac{1}{\lambda_j^2} + \frac{1}{(A_\beta \eta_j)^2}\right)$$
$$(\eta_j^2 \mid -) \sim \text{InvGam}\left(1, \frac{1}{A_\beta^2 \zeta_j} + \frac{1}{\zeta_\eta}\right)$$

- $(\zeta_\eta \mid -) \sim \text{InvGam}((p+1)/2, \sum_{j=1}^p \eta_j^{-2} + 1)$.

APPENDIX B. MCMC ALGORITHM FOR FULL QUANTILE OD MODEL

- $(\beta_0 \mid -) \sim N(\mu_{\beta_0}, \sigma^2_{\beta_0})$ where

$$\sigma^2_{\beta_0} = \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^{n} \frac{1}{z_i} + \frac{1}{\sigma_0^2} \right)^{-1}$$
$$\mu_{\beta_0} = \sigma^2_{\beta_0} \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^{n} \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - \gamma_i - k_1 z_i}{z_i} + \frac{\mu_0}{\sigma_0^2} \right)$$

- $(\boldsymbol{\beta} \mid -) \sim N(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ where

$$\boldsymbol{\Sigma}_b = \left(\frac{1}{k_2^2 \sigma} \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T}{z_i} + \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_p^{-1}) \right)^{-1}$$
$$\boldsymbol{\mu}_b = \boldsymbol{\Sigma}_b \left(\sum_{i=1}^{n} \frac{\boldsymbol{x}_i^T (y_i - \beta_0 - \gamma_i - k_1 z_i)}{k_2^2 \sigma z_i} \right)$$

- For $i = 1, \ldots, n$, $(z_i \mid -) \sim \mathrm{GIG}(0.5, \chi_i, (\frac{k_1^2}{k_2^2} + 2)/(2\sigma))$ where

$$\chi_i = \frac{(y_i - \beta_0 - \gamma_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{k_2^2 \sigma}$$

- $(\sigma \mid -) \sim \mathrm{InvGam}\left(\frac{r_\sigma^0 + 3n}{2}, s_\sigma^n \right)$ where

$$s_\sigma^n = s_\sigma^0 + \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta} - \gamma_i - k_1 z_i)^2}{k_2^2 z_i} + 2 \sum_{i=1}^{n} z_i$$

- The Gibbs samplers for $(\lambda_j^2, \zeta_j, \eta_j^2, \zeta_\eta)$ are the same as in Section A.
- Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^T$. We can sample $\boldsymbol{\gamma}$ in one sampler with $(\boldsymbol{\gamma} \mid -) \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ where

$$\boldsymbol{\Sigma}_g = \left(\frac{1}{k_2^2 \sigma} \mathrm{diag}(z_1^{-1}, \ldots, z_n^{-1}) + \mathrm{diag}(\lambda_{\gamma_1}^{-2}, \ldots, \lambda_{\gamma_n}^{-2}) \right)^{-1}$$
$$\boldsymbol{\mu}_g = \boldsymbol{\Sigma}_b \boldsymbol{b}$$
$$\boldsymbol{b}_i = \frac{y_i - \beta_0 - \boldsymbol{x}_i^T \boldsymbol{\beta} - k_1 z_i}{k_2^2 \sigma z_i}, \quad \text{for } i = 1, \ldots, n$$

- $(\lambda^2_{\gamma_i} \mid -) \sim \mathrm{InvGam}(1, \gamma_i^2/2 + 1/\zeta_{\gamma_i})$.
- $(\zeta_{\gamma_i} \mid -) \sim \mathrm{InvGam}(1, 1/\lambda^2_{\gamma_i} + 1/(A_\gamma \eta_{\gamma_i})^2)$.
- $(\eta^2_{\gamma_i} \mid -) \sim \mathrm{InvGam}(1, 1/(A^2_\gamma \zeta_{\gamma_i}) + 1/\zeta_{\eta,\gamma})$.
- $(\zeta_{\eta,\gamma} \mid -) \sim \mathrm{InvGam}((n+1)/2, \sum_{i=1}^n \eta_{\gamma_i}^{-2} + 1)$.

ACKNOWLEDGEMENTS

We greatly appreciate the Editor, the Associate Editor and Reviewer for their helpful comments that improve the style and substance of the paper. Research of Taeryon Choi was supported by a Korea University Grant (K1807851) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1A2C1010018).

Received 3 March 2019

REFERENCES

[1] Alhamzawi, R., Yu, K. and Benoit, D. F. (2012). Bayesian adaptive Lasso quantile regression. *Statistical Modelling* **12** 279–297. MR3179503

[2] Armagan, A., Clyde, M. and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In *Advances in Neural Information Processing Systems* 523–531.

[3] Armagan, A., Dunson, D. B. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23** 119. MR3076161

[4] Bernardi, M., Bottone, M. and Petrella, L. (2016). Bayesian Robust Quantile Regression. *arXiv preprint arXiv:1605.05602*.

[5] Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **12** 1105–1131. MR3724980

[6] Bhattacharya, A., Chakraborty, A. and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika* **103** 985–991. MR3620452

[7] Bhattacharya, A., Pati, D., Pillai, N. S. and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490. MR3449048

[8] Bishop, C. M. (2006). *Pattern recognition and machine learning, Chapter 10*. Springer. MR2247587

[9] Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776

[10] Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 627–641. MR1626005

[11] Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[12] Chen, C. W., Dunson, D. B., Reed, C. and Yu, K. (2013). Bayesian variable selection in quantile regression. *Statistics and its Interface* **6** 261–274. MR3066690

[13] Chernozhukov, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160

[14] Cook, R. D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. Ser. B* **48** 133–169. With discussion. MR0867994

[15] Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. *Advances in Neural Information Processing Systems* **13** 507–513.

[16] Giordano, R., Broderick, T. and Jordan, M. I. (2018). Covariances, robustness and variational bayes. *The Journal of Machine Learning Research* **19** 1981–2029. MR3874159

[17] Giraud, C. (2014). *Introduction to High-Dimensional Statistics. Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Taylor & Francis.

[18] Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440

[19] Hahn, P. R., He, J. and Lopes, H. (2018). Bayesian factor model shrinkage for linear IV regression with many instruments. *J. Bus. Econom. Statist.* **36** 278–287. MR3790214

[20] Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association* **106** 1383–1393. MR2896843

[21] Harrison Jr, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* **5** 81–102.

[22] Jo, S., Roh, T. and Choi, T. (2016). Bayesian spectral analysis models for quantile regression with Dirichlet process mixtures. *Journal of Nonparametric Statistics* **28** 177–206. MR3463556
[23] Knowles, D. A. and Minka, T. (2011). Non-conjugate Variational Message Passing for Multinomial and Binary Regression. In *Advances in Neural Information Processing Systems* **24** (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 1701–1709. Curran Associates, Inc.
[24] Kobayashi, G. et al. (2017). Bayesian Endogenous Tobit Quantile Regression. *Bayesian Analysis* **12** 161–191. MR3597571
[25] Kottas, A. and Kranjajic, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics* **36** 297–319. MR2528986
[26] Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of statistical computation and simulation* **81** 1565–1578. MR2851270
[27] Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* 1846–1877. MR3015046
[28] Li, Q., Xi, R., Lin, N. et al. (2010). Bayesian regularized quantile regression. *Bayesian Analysis* **5** 533–556. MR2719666
[29] MacKay, D. J. (2003). *Information theory, inference and learning algorithms, Chapter 33*. Cambridge University Press. MR2012999
[30] McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis* **52** 249–257. MR2409979
[31] Neville, S. E. (2013). Elaborate distribution semiparametric regression via mean field variational Bayes. PhD thesis, University of Wollongong.
[32] Neville, S. E., Ormerod, J. T. and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electron. J. Stat.* **8** 1113–1151. MR3263115
[33] Ong, V. H., Mensah, D. K., Nott, D. J., Jo, S., Park, B. and Choi, T. (2017). A variational Bayes approach to a semiparametric regression using Gaussian process priors. *Electron. J. Stat.* **to appear**. MR3720915
[34] Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician* **64** 140–153. MR2757005
[35] Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103** 681–686. MR2524001
[36] Peltola, T., Havulinna, A. S., Salomaa, V. and Vehtari, A. (2014). Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop-Volume* **1218** 79–88. CEUR-WS.org.
[37] Pong-Wong, R. (2014). Estimation of genomic breeding values using the Horseshoe prior. *BMC Proceedings* **8** S6.
[38] Portnoy, S. and Jurečková, J. (1999). On extreme regression quantiles. *Extremes* **2** 227–243 (2000). MR1781938
[39] Santos, B. and Bolfarine, H. (2016). On Bayesian quantile regression and outliers. *arXiv preprint arXiv:1601.07344*.
[40] She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106** 626–639. MR2847975
[41] Sriram, K., Ramamoorthi, R., Ghosh, P. et al. (2013). Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density. *Bayesian Analysis* **8** 479–504. MR3066950
[42] Tan, L. S. L. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Stat. Comput.* **28** 259–275. MR3747562
[43] Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In *Bayesian Time Series Models* (D. Barber, T. Cemgil and S. Chiappa, eds.) 5, 109–130. Cambridge University Press. MR2894235
[44] Turner, R. E. and Sahani, M. (2011). *Two problems with variational expectation maximisation for time series models*. In *Bayesian Time Series Models* 104–124. Cambridge University Press. MR2894235
[45] Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research* **15** 1351–1369. MR3214787
[46] Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* **112** 137–168. MR3646558
[47] Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Anal.* **6** 847–900. MR2869967
[48] Wang, B. and Titterington, D. M. (2004). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics* 373–380.
[49] Wang, K. and Wang, H. J. (2016). Optimally combined estimation for tail quantile regression. *Statist. Sinica* **26** 295–311. MR3468354
[50] Wang, X., Nott, D. J., Drovandi, C. C., Mengersen, K. and Evans, M. (2018). Using history matching for prior choice. *Technometrics* **60** 445–460. MR3878100
[51] Weisberg, S. (2014). *Applied linear regression*, fourth ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ. MR3237496
[52] Wichitaksorn, N., Choy, S. and Gerlach, R. (2014). A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics* **42** 579–596. MR3281462
[53] Winn, J. and Bishop, C. M. (2005). Variational message passing. *J. Mach. Learn. Res.* **6** 661–694. MR2249835
[54] Xueou, W. (2016). Improved Bayesian Model Specification Via Robust Methods and Checking for Prior-Data Conflict PhD thesis, National University of Singapore.
[55] Yan, Y. and Kottas, A. (2017). A new family of error distributions for Bayesian quantile regression. *arXiv preprint arXiv:1701.05666*.
[56] Yu, K., Chen, C. W., Reed, C. and Dunson, D. B. (2013). Bayesian variable selection in quantile regression. *Stat Interface* **6** 261–274. MR3066690
[57] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters* **54** 437–447. MR1861390

Daeyoung Lim
Department of Statistics
University of Connecticut
U.S.A.
E-mail address: daeyoung.lim@uconn.edu

Beomjo Park
Department of Statistics
Carnegie Mellon University
U.S.A.
E-mail address: beomjop@andrew.cmu.edu

David Nott
Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: standj@nus.edu.sg

Xueou Wang
Department of Statistics and Applied Probability
National University of Singapore
Singapore
E-mail address: xueou.wang@gmail.com

Taeryon Choi
Department of Statistics
Korea University
Republic of Korea
E-mail address: trchoi@korea.ac.kr