Bayesian Scale Mixtures of Normals Linear Regression and Bayesian Quantile Regression with Big Data and Variable Selection

Yuanqi Chu[a], Zhouping Yin[b], Keming Yu[a,*]

[a] *Department of Mathematics, Brunel University London, Kingston Lane, Uxbridge, Middlesex, UB8 3PH London, United Kingdom*
[b] *College of Mathematics and Physics, Anqing Normal University, Anqing 246133, People's Republic of China*

Abstract

Quantile regression, which estimates various conditional quantiles of a response variable, including the median (0.5th quantile), is particularly useful when the conditional distribution is asymmetric or heterogeneous or fat-tailed or truncated. Bayesian methods for the inference of quantile regression have been receiving increasing attention from both theoretical and empirical viewpoints but facing the challenge of scaling up when the data are too large to be processed by a single machine under many big data environments nowadays. In this paper, we develop a structure link between Bayesian scale mixtures of normals linear regression and Bayesian quantile regression (BQR) via normal-inverse-gamma (NIG) distribution type of likelihood function, prior distribution and posterior distribution. We further explore the detailed methods of BQR for big data, variable selection and posterior prediction. The performance of the proposed techniques is evaluated via simulation studies and a real data analysis.

Keywords: Scale Mixtures of Normals, Quantile Regression (QR), Bayesian Inference, Big Data, Normal-Inverse-Gamma (NIG), Variable Selection

1. Introduction

Quantile regression (QR) estimates various conditional quantiles of a response or dependent random variable, including the median (0.5th quantile). Putting different quantile regressions together provides a more complete description of the underlying conditional distribution of the response than a simple mean regression. This is particularly useful when the conditional distribution is asymmetric or heterogeneous or fat-tailed or truncated. Quantile regression has been widely used in statistics and numerous application areas (Cole and Green

*Corresponding author
Email address: keming.yu@brunel.ac.uk (Keming Yu)

[1]; Koenker and Hallock [2]; Yu et al. [3]; Briollais and Durrieu [4], among others). In the "big data" era for statistical science, the richness of data sources with many complicated data structures and the increase of extreme values and heterogeneity may see quantile regression methods more relevant than mean regression to dig deep into the data and grab information from it. In particular, with advanced power of computer, complicated quantile regression-based models could be developed under a Bayesian framework, and Bayesian quantile regression (BQR) has received increasing attention from both theoretical and empirical viewpoints with wide applications (see Bernardi et al. [5]; Wang et al. [6]; Rodrigues and Fan [7]; Petrella and Raponi [8], among others). So far, several methods have been developed to quantile regression for big data analysis (Wu and Yin [9]; Yu et al. [10]; Gu et al. [11]; Chen et al. [12], among others), but little attention has been paid to such methodology under Bayesian inference paradigm.

In this paper, we propose a new approach of BQR for big data. This approach has its posterior distribution on the whole data as a joint posterior from M sub-datasets split from the whole data. Section 2 and Section 3 give details of the normal-inverse-gamma (NIG) expressions of the prior and posterior distributions for Bayesian scale mixtures of normals linear regression and BQR respectively. Section 4 presents the posterior predictive distributions. Section 5 develops big data based algorithms for Bayesian scale mixtures of normals model and BQR via the introduction of NIG summation operator. Section 6 provides big data based algorithms for Bayesian $LASSO$ scale mixtures of normals regression and Bayesian $LASSO$ quantile regression. Section 7 demonstrates the proposed algorithms via simulations and a real data analysis.

2. Bayesian scale mixtures of normals linear regression for big data

2.1. Model and likelihood

Consider the scale mixtures of normals linear model

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \sigma\epsilon_i, \;\; i = 1, \ldots, n,$$

where $\boldsymbol{x}_i$ is a $k \times 1$ vector of predictors for observation y_i, $\boldsymbol{\beta}$ is a $k \times 1$ unknown vector of regression coefficients, $\epsilon_1, \ldots, \epsilon_n$ are $i.i.d.$ random variables distributed as scale mixtures of normals. That is, $\epsilon_i \stackrel{d}{=} \sqrt{\zeta_i} z_i$ where z_i follows a standard normal distribution and ζ_i is an independent random variable with some known probability distribution f_{ζ_i} on $(0, \infty)$. σ is an unknown scaling factor. We aim to model the conditional mean $E[y_i|\boldsymbol{x}_i, \zeta_i]$ under Bayesian estimation paradigm. Our primary interest is in inference of the unknown parameters $\boldsymbol{\beta}$ and σ. More compactly, the scale mixtures of normals linear regression in matrix format is specified as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{Y} = (y_1, \dots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$ is an $n \times k$ predictor matrix and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is an $n \times 1$ scale mixtures of normals disturbances with a mean vector of zeros and $n \times n$ positive definite covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\zeta_1, \dots, \zeta_n)$. Then the conditional likelihood of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}) \propto (\sigma^2)^{-\frac{n}{2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\}. \tag{2}$$

Consider the formulation

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$, we can thus rewrite likelihood (2) as

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}) &\propto (\sigma^2)^{-\frac{n-k}{2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\} \\ &\quad (\sigma^2)^{-\frac{k}{2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\} \qquad (3) \\ &= (\sigma^2)^{-(a+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[b + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{\beta} - \boldsymbol{\mu})]\} \\ &\propto IG(a, b) N_k(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Lambda}^{-1}), \qquad (4) \end{aligned}$$

where $IG(a, b)$ denotes the inverse-gamma distribution with shape parameter a and scale parameter b. $N_k(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Lambda}^{-1})$ denotes the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\sigma^2\boldsymbol{\Lambda}^{-1}$. The represented likelihood (4) is a typical structure of a k-dimensional normal-inverse-gamma distribution $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$ in terms of parameters $(\boldsymbol{\beta}, \sigma^2)$. Here $\boldsymbol{\mu} = \hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda} = \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X}, a = \frac{n-k-2}{2}, b = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$.

2.2. NIG expressions of posterior distribution

2.2.1. Posterior distribution under non-informative prior

The conjugate non-informative prior $f(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}$ suggests a specific case of the NIG distribution which is denoted as $NIG_k(\mathbf{0}_k, \mathbf{0}_{k\times k}, -\frac{k}{2}, 0)$. Under this prior, the posterior distribution $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma})$ is given by

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) &= f(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \\ &= IG(\widetilde{a}, \widetilde{b}) N_k(\widetilde{\boldsymbol{\mu}}, \sigma^2\widetilde{\boldsymbol{\Lambda}}^{-1}) \\ &\propto (\sigma^2)^{-(\widetilde{a}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[\widetilde{b} + \frac{1}{2}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\mu}})^T \widetilde{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \widetilde{\boldsymbol{\mu}})]\}. \end{aligned}$$

Then we denote the joint posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ as $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma})$ $= NIG_k(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}, \widetilde{a}, \widetilde{b})$. Here $\widetilde{\boldsymbol{\mu}} = (\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}, \widetilde{\boldsymbol{\Lambda}} = \mathbf{X}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{X}, \widetilde{a} = \frac{n-k}{2}, \widetilde{b} = \frac{1}{2}\boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{Y} - \frac{1}{2}\widetilde{\boldsymbol{\mu}}^T \widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\mu}}$.

2.2.2. *Posterior distribution under informative prior*

Consider a form of conjugate informative prior for $(\boldsymbol{\beta}, \sigma^2)$:

$$
\begin{aligned}
f(\boldsymbol{\beta}, \sigma^2) &= f(\sigma^2) f(\boldsymbol{\beta}|\sigma^2) \\
&\propto (\sigma^2)^{-(a_0+1)} \exp\{-\frac{b_0}{\sigma^2}\} (\sigma^2)^{-\frac{k}{2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\boldsymbol{\beta}-\boldsymbol{\mu}_0)\} \\
&= (\sigma^2)^{-(a_0+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[b_0 + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)^T \boldsymbol{\Lambda}_0 (\boldsymbol{\beta}-\boldsymbol{\mu}_0)]\},
\end{aligned}
$$

where $f(\sigma^2)$ is $IG(a_0, b_0)$ with prior values a_0, b_0 and $f(\boldsymbol{\beta}|\sigma^2)$ is $N_k(\boldsymbol{\mu}_0, \sigma^2 \boldsymbol{\Lambda}_0^{-1})$ with prior values $\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0$. We can thus calibrate the joint prior as an NIG distribution $f(\beta, \sigma^2) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$. Under this prior, the posterior distribution is given by

$$
\begin{aligned}
f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) &= f(\sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) f(\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \\
&= IG(\bar{a}, \bar{b}) N_k(\bar{\boldsymbol{\mu}}, \sigma^2 \bar{\boldsymbol{\Lambda}}^{-1}) \\
&\propto (\sigma^2)^{-(\bar{a}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})^T \bar{\boldsymbol{\Lambda}} (\boldsymbol{\beta}-\bar{\boldsymbol{\mu}})]\},
\end{aligned}
$$

which can be denoted as $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) = NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$. Here $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda}_0 + \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1} (\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}), \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}_0 + \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}, \bar{a} = a_0 + \frac{n}{2}, \bar{b} = b_0 + \frac{1}{2} \boldsymbol{Y}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{Y} + \frac{1}{2} \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \frac{1}{2} \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}$.

3. Bayesian quantile regression for big data

3.1. *Model and likelihood*

Let y_i be a continuous response variable and $\boldsymbol{x}_i$ a $k \times 1$ vector of predictors for the ith observation, $i = 1, \ldots, n$. Denote $Q_p(y_i|\boldsymbol{x}_i)$ as the pth $(0 < p < 1)$ quantile regression function of y_i given $\boldsymbol{x}_i$. Suppose that all conditional quantiles $Q_p(y_i|\boldsymbol{x}_i)$ can be modelled as $Q_p(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_p$, where $\boldsymbol{\beta}_p$ is a $k \times 1$ vector of unknown parameters that depends on quantile p. Then the linear Quantile Regression (QR) model for the pth quantile can be denoted as

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_p + \epsilon_i, \;\; i = 1, \ldots, n,$$

where ϵ_i is the error term whose distribution is assumed to have zero pth quantile. Following Koenker and Bassett [13], the estimation for $\boldsymbol{\beta}_p$ proceeds by minimizing

$$\sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_p), \tag{5}$$

where $\rho_p(u) = u\{p - I(u < 0)\}$ is the check function and $I(\cdot)$ denotes the indicator function. Equivalently, we can express $\rho_p(u)$ as

$$\rho_p(u) = \frac{|u| + (2p-1)u}{2}. \tag{6}$$

According to Yu and Moyeed [14] and Yu and Stander [15], minimizing (5) is equivalent to maximizing a likelihood function that is based on the asymmetric Laplace distribution (ALD) at specific value of p. Assuming an ALD-based working model such that $\epsilon_i \sim ALD(\kappa, \sigma, p)$ with location parameter $\kappa = 0$, scale parameter $\sigma \in (0, \infty)$ and skewness parameter $p \in (0, 1)$, then the probability density function of ϵ_i is given by

$$f(\epsilon_i; \kappa = 0, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\{-\frac{\rho_p(\epsilon_i)}{\sigma}\}, \ \ i = 1, \ldots, n,$$

where $\rho_p(u)$ is defined in (6). Following Reed and Yu [16] and Kozumi and Kobayashi [17], we can represent ϵ_i as a scale mixture of normals with an exponential mixing density as follows:

$$\epsilon_i | v_i, \sigma \sim N((1-2p)v_i, 2\sigma v_i), \ \ v_i|\sigma \sim \text{Exp}(\sigma^{-1}p(1-p)),$$

where $\text{Exp}(\theta)$ denotes an exponential distribution with rate parameter θ. Consequently, the conditional distribution of y_i is normal with mean $\boldsymbol{x}_i^T\boldsymbol{\beta}_p + (1-2p)v_i$ and variance $2\sigma v_i$:

$$y_i | \boldsymbol{\beta}_p, \sigma, v_i, \boldsymbol{x}_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}_p + (1-2p)v_i, 2\sigma v_i), \ \ i = 1, \ldots, n. \tag{7}$$

The matrix form of (7) is as follows:

$$\boldsymbol{Y} | \boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}, \boldsymbol{V} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}_p + (1-2p)\boldsymbol{v}, 2\sigma\boldsymbol{V}),$$

where $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times k$ predictor matrix with ith row $\boldsymbol{x}_i^T$, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ and $\boldsymbol{V} = \text{diag}(\boldsymbol{v})$. Thus, the conditional likelihood of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{Y}|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}, \boldsymbol{V}) \propto \sigma^{-n/2} \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_p - (1-2p)\boldsymbol{v}]^T \frac{\boldsymbol{V}^{-1}}{2}[\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_p - (1-2p)\boldsymbol{v}]\}.$$

Let $\boldsymbol{Y}_p^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - (1-2p)\boldsymbol{v})$ and $\boldsymbol{X}^* = \frac{1}{\sqrt{2}}\boldsymbol{X}$, then $\boldsymbol{Y}_p^*$ follows a normal-type of conditional likelihood as

$$f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-n/2} \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p]^T\boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p]\}. \tag{8}$$

Denote further $\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*$, we can rewrite (8) as

$$
\begin{aligned}
f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-\frac{n-k}{2}} \exp\{-\frac{1}{2\sigma}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]^T \boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]\} \\
&\sigma^{-\frac{k}{2}} \exp\{-\frac{1}{2\sigma}(\boldsymbol{\beta}_p - \hat{\boldsymbol{\beta}}_p)^T(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)(\boldsymbol{\beta}_p - \hat{\boldsymbol{\beta}}_p)\} \\
&= (\sigma)^{-(a+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[b_p + \frac{1}{2}(\boldsymbol{\beta}_p - \boldsymbol{\mu}_p)^T \boldsymbol{\Lambda}(\boldsymbol{\beta}_p - \boldsymbol{\mu}_p)]\} \\
&\propto IG(a, b_p) N_k(\boldsymbol{\mu}_p, \sigma\boldsymbol{\Lambda}^{-1}), \qquad (9)
\end{aligned}
$$

where $\boldsymbol{\mu}_p = \hat{\boldsymbol{\beta}}_p, \boldsymbol{\Lambda} = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, a = \frac{n-k-2}{2}, b_p = \frac{1}{2}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]^T\boldsymbol{V}^{-1}[\boldsymbol{Y}_p^* - \boldsymbol{X}^*\hat{\boldsymbol{\beta}}_p]$. The reformulated likelihood (9) is a structure of a k-dimensional distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$ in terms of parameters $(\boldsymbol{\beta}_p, \sigma)$.

3.2. NIG expressions of posterior distribution

3.2.1. Posterior distribution under non-informative prior

The conjugate non-informative prior $f(\boldsymbol{\beta}_p, \sigma) \propto \sigma^{-1}$ suggests a form of $NIG_k(\mathbf{0}_k, \mathbf{0}_{k\times k}, -\frac{k}{2}, 0)$. Given this prior, the joint conditional posterior distribution $f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ can be written as

$$
\begin{aligned}
f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}|\boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto f(\boldsymbol{Y}_p^*|\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}) f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}) f(\boldsymbol{v}|\sigma) f(\sigma) \\
&\propto \sigma^{-(\frac{3n+2}{2})}(\prod_{i=1}^n v_i^{-1/2}) \\
&\times \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + 2p(1-p)\sum_{i=1}^n v_i]\}.
\end{aligned}
$$

The posterior distribution $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ is thus given by

$$
\begin{aligned}
f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto \sigma^{-(\frac{3n+2}{2})} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T\boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + 2p(1-p)\sum_{i=1}^n v_i]\} \\
&= \sigma^{-(\frac{3n-k}{2}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[\widetilde{b}_p + \frac{1}{2}(\boldsymbol{\beta}_p - \widetilde{\boldsymbol{\mu}}_p)^T\widetilde{\boldsymbol{\Lambda}}(\boldsymbol{\beta}_p - \widetilde{\boldsymbol{\mu}}_p)]\},
\end{aligned}
$$

which can be denoted as a k-dimensional distribution $NIG_k(\widetilde{\boldsymbol{\mu}}_p, \widetilde{\boldsymbol{\Lambda}}, \widetilde{a}, \widetilde{b}_p)$, where $\widetilde{\boldsymbol{\mu}}_p = (\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*, \widetilde{\boldsymbol{\Lambda}} = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, \widetilde{a} = \frac{3n-k}{2}, \widetilde{b}_p = \frac{1}{2}\boldsymbol{Y}_p^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^* - \frac{1}{2}Y_p^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*\widetilde{\boldsymbol{\mu}}_p + p(1-p)\sum_{i=1}^n v_i$. Furthermore, the full posterior distribution of each v_i conditional on $\boldsymbol{\beta}_p, \sigma$ and raw data $y_i, \boldsymbol{x}_i, i = 1, 2, \ldots, n$ is obtained by

$$
\begin{aligned}
f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2] - \frac{p(1-p)}{\sigma}v_i\} \\
&= v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + v_i]\} \\
&= v_i^{-1/2} \exp\{-\frac{1}{2}(v_i^{-1}\widetilde{\xi}_i^{\,2} + v_i\widetilde{\zeta}_i^{\,2})\},
\end{aligned}
$$

where $\widetilde{\xi}_i^{\,2} = (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_p)^2/2\sigma$ and $\widetilde{\zeta}_i^{\,2} = 1/2\sigma$. This conditional posterior can be recognized as a form of generalized inverse Gaussian distribution $GIG(\frac{1}{2}, \widetilde{\xi}_i, \widetilde{\zeta}_i)$. Recall that if $z \sim GIG(\varphi, \eta_1, \eta_2)$, then the probability density function of z is given by

$$f(z|\varphi, \eta_1, \eta_2) = \frac{(\eta_2/\eta_1)^{\varphi}}{2K_{\varphi}(\eta_1\eta_2)} z^{\varphi-1} \exp\{-\frac{1}{2}(z^{-1}\,\eta_1^2 + z\,\eta_2^2)\}, z > 0,\ -\infty < \varphi < \infty,\ \eta_1, \eta_2 \geq 0,$$

where $K_{\varphi}(\cdot)$ is a modified Bessel function of the third kind (Barndorff-Nielsen and Shephard [18]).

3.2.2. Posterior distribution under informative g-prior

For the informative prior setting, following Alhamzawi and Yu [19], a conjugate prior for $(\boldsymbol{\beta}_p, \sigma)$ with a modification of Zellner's informative g-prior (Zellner [20]) in QR could be provided as

$$\boldsymbol{\beta}_p | \sigma, \boldsymbol{X}^*, \boldsymbol{V} \sim N_k(\boldsymbol{0}_k, g\sigma(\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*)^{-1}), \ \ f(\sigma) \propto \sigma^{-1},$$

where $g > 0$ is a known scaling factor prescribed by the user. Smith and Kohn [21] proposed a Bayesian variable selection algorithm utilizing regression splines. They found that the choice of $g = 100$ works well and suggested to choose g between 10 and 1000. Following Smith and Kohn [21], the fixed setting of $g = 100$ has been considered by some other authors (see Lee et al. [22]; Gupta et al. [23], among others). Then we obtain the joint prior distribution of $(\boldsymbol{\beta}_p, \sigma)$ as

$$f(\boldsymbol{\beta}_p, \sigma | \boldsymbol{X}^*, \boldsymbol{V}) \propto \sigma^{-(\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[\frac{1}{2}\boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g} \boldsymbol{\beta}_p]\}, \tag{10}$$

which is a special case of $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0)$ with $\boldsymbol{\mu}_0 = \boldsymbol{0}_k, \boldsymbol{\Lambda}_{g0} = \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}$, $a_0 = 0, b_0 = 0$.

The joint conditional posterior distribution $f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v} | \boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ under prior (10) is given by

$$
\begin{aligned}
&f(\boldsymbol{\beta}_p, \sigma, \boldsymbol{v} | \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto f(\boldsymbol{Y}_p^* | \boldsymbol{\beta}_p, \sigma, \boldsymbol{v}) f(\boldsymbol{\beta}_p | \sigma, \boldsymbol{v}) f(\boldsymbol{v}|\sigma) f(\sigma) \\
&\propto \sigma^{-(\frac{3n+k+2}{2})} (\prod_{i=1}^{n} v_i^{-1/2}) |\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*|^{1/2} \\
&\times \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1} (\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g} \boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^{n} v_i]\}.
\end{aligned}
$$

The corresponding posterior $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*)$ is given as follows:

$$\begin{aligned}
f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto \sigma^{-(\frac{3n+k+2}{2})} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\
&\quad + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^n v_i]\} \\
&= \sigma^{-(\frac{3n}{2}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma}[\bar{b}_p + \frac{1}{2}(\boldsymbol{\beta}_p - \bar{\boldsymbol{\mu}}_p)^T \bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta}_p - \bar{\boldsymbol{\mu}}_p)]\},
\end{aligned}$$

which has an expression of $NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$, where $\bar{\boldsymbol{\mu}}_p = [(1+\frac{1}{g})\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*]^{-1}$ $\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^*, \bar{\boldsymbol{\Lambda}} = (1+\frac{1}{g})\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, \bar{a} = \frac{3n}{2}, \bar{b}_p = \frac{1}{2}\boldsymbol{Y}_p^{*T}\boldsymbol{V}^{-1}\boldsymbol{Y}_p^* - \frac{1}{2}\bar{\boldsymbol{\mu}}_p^T\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}_p + p(1-p)\sum_{i=1}^n v_i$. Moreover, the full conditional marginal distributions of $\boldsymbol{\beta}_p$ and σ can be obtained respectively by

$$f(\boldsymbol{\beta}_p|\sigma, \boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p]\},$$

which can be expressed as an $N_k(\bar{\boldsymbol{\mu}}_p, \sigma\bar{\boldsymbol{\Lambda}}^{-1})$, and

$$\begin{aligned}
f(\sigma|\boldsymbol{\beta}_p, \boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) &\propto \sigma^{-(\frac{3n+k}{2}+1)} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\
&\quad + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^n v_i]\},
\end{aligned}$$

which is an IG distribution with shape $\frac{3n+k}{2}$ and scale $\frac{1}{2}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \frac{\boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*}{g}\boldsymbol{\beta}_p + 2p(1-p)\sum_{i=1}^n v_i]$. The full posterior distribution of each $v_i, i = 1, 2, \ldots, n$ is also tractable:

$$\begin{aligned}
f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1} \exp\{-\frac{1}{4\sigma}[v_i^{-1}((y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \frac{\boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p}{g})] - \frac{p(1-p)}{\sigma}v_i\} \\
&= v_i^{-1} \exp\{-\frac{1}{4\sigma}[v_i^{-1}((y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \frac{\boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p}{g}) + v_i]\} \\
&= v_i^{-1} \exp\{-\frac{1}{2}(v_i^{-1}\bar{\xi}_i^2 + v_i\bar{\zeta}_i^2)\},
\end{aligned}$$

where $\bar{\xi}_i^2 = [(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + \boldsymbol{\beta}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_p/g]/2\sigma$ and $\bar{\zeta}_i^2 = 1/2\sigma$, which can be recognized as a $GIG(0, \bar{\xi}_i, \bar{\zeta}_i)$.

4. Posterior predictive distributions

4.1. Posterior predictive distribution for Bayesian scale mixtures of normals regression

Given a new $n \times k$ predictor matrix $\boldsymbol{X}^{\text{new}}$, one may be interested in the Bayesian prediction of a new response outcome $\boldsymbol{Y}^{\text{new}}$ under the current posterior calibration of $(\boldsymbol{\beta}, \sigma^2)$ with the observations $\boldsymbol{X}, \boldsymbol{Y}$. To obtain the analytic expression of $f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y})$, we first derive the following computation result of integrating out σ^2 from the joint posterior $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}) = NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$, where the expressions for $\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}$ and $\bar{b}$ are given in Section 2.2.2.

$$
\begin{aligned}
\int_0^\infty NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})\, d\sigma^2 &= \frac{\bar{b}^{\bar{a}}}{(2\pi)^{\frac{k}{2}}|\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}\Gamma(\bar{a})} \\
&\quad \int_0^\infty (\sigma^2)^{-(\bar{a}+\frac{k}{2}+1)} \exp\{-\frac{1}{\sigma^2}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T \bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]\}\, d\sigma^2 \\
&= \frac{\bar{b}^{\bar{a}}\Gamma(\bar{a} + \frac{k}{2})}{(2\pi)^{\frac{k}{2}}|\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}\Gamma(\bar{a})}[\bar{b} + \frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T \bar{\boldsymbol{\Lambda}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]^{-(\bar{a}+\frac{k}{2})} \\
&= \frac{\Gamma(\frac{2\bar{a}+k}{2})}{\Gamma(\frac{2\bar{a}}{2})(2\bar{a})^{\frac{k}{2}}\pi^{\frac{k}{2}}|\frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1}|^{\frac{1}{2}}}[1 + \frac{1}{2\bar{a}}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})^T (\frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1})^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\mu}})]^{-(\frac{2\bar{a}+k}{2})} \\
&= \frac{\Gamma(\frac{v_t+k}{2})}{\Gamma(\frac{v_t}{2})v_t^{\frac{k}{2}}\pi^{\frac{k}{2}}|\boldsymbol{\Sigma}_t|^{\frac{1}{2}}}[1 + \frac{1}{v_t}(\boldsymbol{\beta} - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_t)]^{-\frac{v_t+k}{2}} \\
&= \boldsymbol{t}_{v_t}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t).
\end{aligned} \tag{11}
$$

That is, the marginal posterior $f(\boldsymbol{\beta}|\boldsymbol{Y})$ is a k-dimensional multivariate t-distribution $\boldsymbol{t}_{v_t}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ with location vector $\boldsymbol{\mu}_t = \bar{\boldsymbol{\mu}}$, shape matrix $\boldsymbol{\Sigma}_t = \frac{\bar{b}}{\bar{a}}\bar{\boldsymbol{\Lambda}}^{-1}$ and degrees of freedom $v_t = 2\bar{a}$. Then the computation of the posterior predictive distribution of $\boldsymbol{Y}^{\text{new}}$ can be proceeded as follows:

$$
\begin{aligned}
f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}) &= \int_0^\infty \int_{-\infty}^\infty f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta}, \sigma^2) f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y})\, d\boldsymbol{\beta}\, d\sigma^2 \\
&= \int_0^\infty \int_{-\infty}^\infty N_n(\boldsymbol{X}^{\text{new}}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}) NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})\, d\boldsymbol{\beta}\, d\sigma^2 \\
&= \int_0^\infty NIG_k(\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}, (\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})^{-1}, \bar{a}, \bar{b})\, d\sigma^2.
\end{aligned} \tag{12}
$$

Applying the integral result (11) to (12), the computation of the density $f(\mathbf{Y}^{\text{new}}|\mathbf{Y})$ is given by

$$
f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}) = \boldsymbol{t}_{2\bar{a}}(\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}, \frac{\bar{b}}{\bar{a}}(\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})),
$$

which is an n-dimensional multivariate t-distribution with location $\boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\mu}}$, shape matrix $\frac{\bar{b}}{\bar{a}}(\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})$ and degrees of freedom $2\bar{a}$. Furthermore, by the law of total conditional variance (Bowsher and Swain [24]), we can obtain the variance of the future observation $\mathbf{Y}^{\text{new}}$ conditional on σ^2

$$
\begin{aligned}
var(\boldsymbol{Y}^{\text{new}}|\sigma^2) &= E[var(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta}, \sigma^2)|\sigma^2] + var[E(\boldsymbol{Y}^{\text{new}}|\boldsymbol{\beta}, \sigma^2)|\sigma^2] \\
&= E[\sigma^2\boldsymbol{\Sigma}|\sigma^2] + var[\boldsymbol{X}^{\text{new}}\boldsymbol{\beta}|\sigma^2] \\
&= (\boldsymbol{\Sigma} + \boldsymbol{X}^{\text{new}}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}T})\sigma^2.
\end{aligned}
$$

Therefore, given σ^2, the posterior predictive distribution has two constituents of uncertainty: (1) the model variability induced by the term σ^2 in $\boldsymbol{Y}$ and (2) the posterior uncertainty within the current calibration of $(\boldsymbol{\beta}, \sigma^2)$ due to the finite sample size of $\boldsymbol{Y}$.

4.2. Posterior predictive distribution for Bayesian quantile regression

In the context of the Bayesian quantile regression model, we carry out the prediction of a new measurement $\boldsymbol{Y}^{\text{new}}$ given a new predictor matrix $\boldsymbol{X}^{\text{new}}$ along with the current estimated parameters $(\boldsymbol{\beta}_p, \sigma)$ as follows. Consider the linear QR model for the pth quantile and observations $\boldsymbol{X}$ and $\boldsymbol{Y}$, and follow the notations for $\boldsymbol{X}^*, \boldsymbol{Y}_p^*, \boldsymbol{v}$ and $\boldsymbol{V}$ presented in Section 3.1. Under the joint posterior $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \mathbf{Y}_p^*, \mathbf{X}^*) = NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$, where $\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}$ and $\bar{b}_p$ are given in Section 3.2.2, we can proceed with the prediction of $\boldsymbol{Y}^{\text{new}}$ in two steps: (1) let $\boldsymbol{X}^{\text{new}*} = \frac{1}{\sqrt{2}}\boldsymbol{X}^{\text{new}}$ and compute the corresponding conditional density $f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{Y}_p^*)$ (with conditioning on $\boldsymbol{X}^{\text{new}*}$ implicit), where $\boldsymbol{Y}_p^{\text{new}*} = \frac{1}{\sqrt{2}}(\boldsymbol{Y}^{\text{new}} - (1-2p)\boldsymbol{v})$ is a linear transformation of variable $\boldsymbol{Y}^{\text{new}}$; (2) derive the target density $f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}_p^*)$. The conditional distribution of $\boldsymbol{Y}_p^{\text{new}*}$ is given by

$$
\begin{aligned}
f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{Y}_p^*) &= \int_0^\infty \int_{-\infty}^\infty f(\boldsymbol{Y}_p^{\text{new}*}|\boldsymbol{\beta}_p, \sigma) f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{Y}_p^*)\, d\boldsymbol{\beta}_p\, d\sigma \\
&= \int_0^\infty \int_{-\infty}^\infty N_n(\boldsymbol{X}^{\text{new}*}\boldsymbol{\beta}_p, \sigma\boldsymbol{V}) NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)\, d\boldsymbol{\beta}_p\, d\sigma \\
&= \int_0^\infty NIG_k(\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p, (\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})^{-1}, \bar{a}, \bar{b}_p)\, d\sigma \\
&= \mathbf{t}_{2\bar{a}}(\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p, \frac{\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})).
\end{aligned} \tag{13}
$$

The conditional of $\boldsymbol{Y}^{\text{new}} = \sqrt{2}\boldsymbol{Y}_p^{\text{new}*} + (1-2p)\boldsymbol{v}$ is a linear combination of the deduced distribution (13). Following the affine transformation property of the multivariate t-distribution (see Roth [25] for more details), the new response outcome $\boldsymbol{Y}^{\text{new}}$ is distributed as

$$
f(\boldsymbol{Y}^{\text{new}}|\boldsymbol{Y}_p^*) = \boldsymbol{t}_{2\bar{a}}(\sqrt{2}\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p + (1-2p)\boldsymbol{v}, \frac{2\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})), \tag{14}
$$

which is an n-dimensional multivariate t-distribution with location $\sqrt{2}\boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\mu}}_p + (1-2p)\boldsymbol{v}$, shape matrix $\frac{2\bar{b}_p}{\bar{a}}(\boldsymbol{V} + \boldsymbol{X}^{\text{new}*}\bar{\boldsymbol{\Lambda}}^{-1}\boldsymbol{X}^{\text{new}*T})$ and degrees of freedom $2\bar{a}$. Accordingly, the posterior predictive distribution sampling for BQR can be achieved as below. For each $l = 1, \ldots, L$, we draw samples $\sigma^{(l)} \sim IG(\bar{a}, \bar{b}_p)$ and $\boldsymbol{\beta}_p^{(l)} \sim N_k(\bar{\boldsymbol{\mu}}_p, \sigma^{(l)}\bar{\boldsymbol{\Lambda}}^{-1})$. The obtained samples $\{\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\}_{l=1}^L$ give L replicates from the joint posterior distribution $f(\boldsymbol{\beta}_p, \sigma|\boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) = NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p)$. For each sample $\{\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\}$, we generate $\boldsymbol{Y}_p^{\text{new}*(l)} \sim N_n(\boldsymbol{X}^{\text{new}*}\boldsymbol{\beta}_p^{(l)}, \sigma^{(l)}\boldsymbol{V})$. The resulting $\{\boldsymbol{Y}_p^{\text{new}*(l)}\}_{l=1}^L$ provide draws for the conditional distribution (13). Then the corresponding samples $\{\boldsymbol{Y}^{\text{new}(l)}\}_{l=1}^L = \{\sqrt{2}\boldsymbol{Y}_p^{\text{new}*(l)} + (1-2p)\boldsymbol{v}\}_{l=1}^L$ give L replicates from the target posterior predictive density (14).

5. Big data based algorithms for Bayesian scale mixtures of normals regression and BQR

In this section, we propose two divide-and-conquer algorithms to facilitate the calculation of full data posterior distribution in big data settings for Bayesian scale mixtures of normals regression and BQR respectively. We first introduce the concept of NIG multiplication operator as follows.

5.1. NIG multiplication operator of posterior distribution

Given the linear regression model (1) with $n\times 1$ response vector $\boldsymbol{Y}$, observed $n\times k$ design matrix $\boldsymbol{X}$ and $n\times n$ positive definite covariance matrix $\boldsymbol{\Sigma}$, where the sample size n is so large that the data cannot be stored on a single computer. If we partition the big data into M subsets, such that $\boldsymbol{Y} = (\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_M)^T$, $\boldsymbol{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_M)^T$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_M)$, where $\boldsymbol{Y}_m$ is an $n_m\times 1$ vector, $\boldsymbol{X}_m$ is an $n_m\times k$ matrix, $\boldsymbol{\Sigma}_m$ is an $n_m\times n_m$ diagonal matrix and $\sum_{m=1}^M n_m = n$, then following (3) and given the sub-datasets, the conditional likelihood function (2) can be written as

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Sigma}) \propto &(\sigma^2)^{-(\sum_{m=1}^M n_m-k)/2}\exp\{-\frac{1}{2\sigma^2}\sum_{m=1}^M(\boldsymbol{Y}_m-\boldsymbol{X}_m\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{Y}_m-\boldsymbol{X}_m\hat{\boldsymbol{\beta}})\}\\ &\times(\sigma^2)^{-\frac{k}{2}}\exp\{-\frac{1}{2\sigma^2}\sum_{m=1}^M(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X}_m)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\}, \end{aligned} \tag{15}$$

where $\hat{\boldsymbol{\beta}} = (\sum_{m=1}^M \boldsymbol{X}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X}_m)^{-1}\sum_{m=1}^M \boldsymbol{X}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{Y}_m$. The reformulated expression (15) with regard to parameters of interest $(\boldsymbol{\beta},\sigma^2)$ further indicates a multiplication of M NIG distributions

$$\begin{aligned} f(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2,\boldsymbol{\Sigma}) \propto &\prod_{m=1}^M(\sigma^2)^{-(a_m+\frac{k}{2}+1)}\exp\{-\frac{1}{\sigma^2}[b_m+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_m)^T\boldsymbol{\Lambda}_m(\boldsymbol{\beta}-\boldsymbol{\mu}_m)]\}\\ &=\prod_{m=1}^M NIG(\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m,a_m,b_m), \end{aligned}$$

where $\boldsymbol{\mu}_m=\hat{\boldsymbol{\beta}}, \boldsymbol{\Lambda}_m=\boldsymbol{X}_m^T\boldsymbol{\Sigma}_m^{-1}\boldsymbol{X}_m$, $a_m=\frac{n_m-k-2}{2}$, $b_m=\frac{1}{2}(\boldsymbol{Y}_m-\boldsymbol{X}_m\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{Y}_m-\boldsymbol{X}_m\hat{\boldsymbol{\beta}})$. Therefore, we have the following **Proposition 5.1**.

Proposition 5.1. *Given regression model (1) and the described data partition rule, the whole data based likelihood and all sub-datasets based likelihood functions follow NIG distributions and satisfy*

$$NIG_k(\boldsymbol{\mu},\boldsymbol{\Lambda},a,b)=\prod_{m=1}^M NIG(\boldsymbol{\mu}_m,\boldsymbol{\Lambda}_m,a_m,b_m), \tag{16}$$

where $\boldsymbol{\mu} = (\sum_{m=1}^M\boldsymbol{\Lambda}_m)^{-1}\sum_{m=1}^M\boldsymbol{\Lambda}_m\boldsymbol{\mu}_m, \boldsymbol{\Lambda} = \sum_{m=1}^M\boldsymbol{\Lambda}_m, a = \sum_{m=1}^M a_m + \frac{(M-1)(k+2)}{2}, b=\sum_{m=1}^M b_m+\frac{1}{2}\sum_{m=1}^M(\boldsymbol{\mu}_m-\boldsymbol{\mu})^T\boldsymbol{\Lambda}_m(\boldsymbol{\mu}_m-\boldsymbol{\mu})$.

Posterior distributions induced by the entire data set can be obtained by combining formulation (16) with specific priors imposed on $\boldsymbol{\beta}$ and σ^2. The following **Theorem 5.1** elaborates the acquisition of the posterior density through the use of the NIG multiplication operator.

Theorem 5.1. *Suppose the posterior distribution, under the prior $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$ and big data observations $\boldsymbol{X}, \boldsymbol{Y}$, be $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$. Partition the entire data into M subsets, then we have the full data posterior distribution*

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) &= NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) \prod_{m=1}^{M} NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m) \\ &= NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}), \end{aligned}$$

where $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} (\boldsymbol{\Lambda}\boldsymbol{\mu} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m), \bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \bar{a} = a + \frac{n}{2}, \bar{b} = b + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m + \boldsymbol{\mu}^T \boldsymbol{\Lambda} \boldsymbol{\mu} - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$.

Corollary 5.1.1. *The full data posterior distribution under the non-informative prior $NIG_k(\mathbf{0}_k, \mathbf{0}_{k \times k}, -\frac{k}{2}, 0)$ can be obtained as $NIG_k(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}, \widetilde{a}, \widetilde{b})$, where $\widetilde{\boldsymbol{\mu}} = (\sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m, \widetilde{\boldsymbol{\Lambda}} = \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \widetilde{a} = \frac{n-k}{2}$, $\widetilde{b} = \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m - \widetilde{\boldsymbol{\mu}}^T \widetilde{\boldsymbol{\Lambda}} \widetilde{\boldsymbol{\mu}}]$.*

5.2. Algorithm for Bayesian scale mixtures of normals regression

The following efficient divide-and-conquer algorithm is provided to facilitate the study of scale mixtures of normals linear regression in big data scenario.

Algorithm 5.1. *Consider the Bayesian scale mixtures of normals linear regression under informative prior $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$ for $(\boldsymbol{\beta}, \sigma^2)$ and with observed $n \times k$ design matrix $\boldsymbol{X}$, $n \times 1$ response vector $\boldsymbol{Y}$ and positive definite $n \times n$ diagonal covariance matrix $\boldsymbol{\Sigma}$, where the data set is too large to be fit into a single computer. By partitioning the entire data set into M subsets and utilizing the aforementioned NIG multiplication operator, we can obtain the full data posterior distribution by the following divide-and-conquer algorithm.*

Step 1 *let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_M \end{bmatrix}$, $\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_M \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_M \end{bmatrix}$, where $\boldsymbol{X}_m$ is an $n_m \times k$ predictor matrix, $\boldsymbol{Y}_m$ is an $n_m \times 1$ response vector, $\boldsymbol{\Sigma}_m$ is an $n_m \times n_m$ diagonal covariance matrix, $m = 1, \ldots, M$ and $\sum_{m=1}^{M} n_m = n$.*

Step 2 *for each subset, the corresponding likelihood has a representation of $NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$ distribution for $(\boldsymbol{\beta}, \sigma^2)$. Calculate the multiplicative distribution $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$, then the full data posterior can be acquired by merging the prior $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0)$ with the distribution $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$:*

$$NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0, a_0, b_0) NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b),$$

where $\bar{\boldsymbol{\mu}} = (\boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} (\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m)$, $\bar{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$, $\bar{a} = a_0 + \frac{n}{2}$, $\bar{b} = b_0 + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m + \boldsymbol{\mu}_0^T \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$.

Remark. *In the high-dimensional setting ($k \gg n$), the induced multicollinearity of $\boldsymbol{X}$ implies the singularity of $\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}$. However, one can always choose proper prior matrix $\boldsymbol{\Lambda}_0$ such that $\boldsymbol{\Lambda}_0 + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$ is non-singular and therefore $\bar{\boldsymbol{\mu}}$ is well-defined.*

5.3. Algorithm for Bayesian quantile regression

Consider the linear QR model for the pth $(0 < p < 1)$ quantile

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_p + \boldsymbol{\epsilon}, \tag{17}$$

where $\boldsymbol{Y}$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times k$ predictor matrix and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of $ALD(0, \sigma, p)$ disturbances. Following the reformulated conditional likelihood (8), model (17) is equivalent to

$$\boldsymbol{Y}_p^* = \boldsymbol{X}^* \boldsymbol{\beta}_p + \sqrt{\sigma} \boldsymbol{\epsilon}^*, \tag{18}$$

where $\boldsymbol{Y}_p^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y} - (1-2p)\boldsymbol{v})$, $\boldsymbol{X}^* = \frac{1}{\sqrt{2}}\boldsymbol{X}$ and $\boldsymbol{\epsilon}^* \sim N_n(\boldsymbol{0}_n, \boldsymbol{V})$ with $n \times n$ diagonal positive definite covariance matrix $\boldsymbol{V}$. We proceed with Bayesian inference for big data quantile regression through the proposed NIG multiplication operator. We consider model (17) under the g-prior (10) and partition the entire data into M subsets $(\boldsymbol{X}_m, \boldsymbol{Y}_m)$ with individual sample size $n_m, m = 1, \ldots, M$. Then the posterior distribution for the whole data can be obtained by merging the given prior with the multiplication of M subset NIG distributions induced from the massive observations. Based on this, an efficient divide-and-conquer algorithm for big data BQR is provided as below.

Algorithm 5.2. *Consider a pth $(0 < p < 1)$ Bayesian quantile regression under g-prior (10) with the observed $n \times k$ design matrix $\boldsymbol{X}$ and $n \times 1$ response vector $\boldsymbol{Y}$, where the large data cannot be fit into a single computer due to the memory constraint. We obtain the full data posterior distribution by the following divide-and-conquer algorithm.*

Step 1 *partition the entire data into M subsets $\boldsymbol{X}_m, \boldsymbol{Y}_m, m = 1, 2, \ldots, M$, where $\boldsymbol{X}_m$ is an $n_m \times k$ matrix, $\boldsymbol{Y}_m$ is an $n_m \times 1$ vector and $\sum_{m=1}^{M} n_m = n$.*

Step 2 *for each $\boldsymbol{X}_m, \boldsymbol{Y}_m$, a Gibbs sampler for sampling $\boldsymbol{\beta}_p, \sigma$ and the $n_m \times 1$ latent vector $\boldsymbol{v}_m$ follows the below sub-steps:*

2.1 *denote j as the iteration count. Then set $j = 0$ and establish $(\boldsymbol{\beta}_p^{(j=0)}, \sigma^{(j=0)}, \boldsymbol{v}_m^{(j=0)})$ to some starting values.*

2.2 *let $\boldsymbol{X}_m^* = \frac{1}{\sqrt{2}}\boldsymbol{X}_m$, $\boldsymbol{Y}_{pm}^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m)$ and $\boldsymbol{V}_m = diag(\boldsymbol{v}_m)$.*

2.3 *follow the full conditional posterior distributions of* $\boldsymbol{\beta}_p, \sigma$ *and* $\boldsymbol{v}_m$ *given in Section 3.2.2,*

(i) sample $\boldsymbol{v}_m^{(j+1)}$ *from its GIG posterior* $f(\boldsymbol{v}_m|\boldsymbol{\beta}_p^{(j)}, \sigma^{(j)})$.

(ii) sample $\sigma^{(j+1)}$ *from its IG posterior* $f(\sigma|\boldsymbol{\beta}_p^{(j)}, \boldsymbol{v}_m^{(j+1)})$.

(iii) sample $\boldsymbol{\beta}_p^{(j+1)}$ *from its multivariate normal posterior* $f(\boldsymbol{\beta}_p|\sigma^{(j+1)}, \boldsymbol{v}_m^{(j+1)})$.

2.4 *set* $j = j+1$ *and return to* ***Step 2.3*** *until* $j = L$*, where* L *is the number of iteration times.*

Step 3 *calculate the empirical estimates of the means* $\bar{\boldsymbol{\beta}}_p$ *and* $\bar{\sigma}$ *separately based on the* $(L-B)$ *realizations of the Gibbs sequence (discarding the first* B *iterations as a burn-in). Then generate an* n_m *i.i.d. sample on* $\bar{v}_i$*, where* $\bar{v}_i \sim GIG(0, \bar{\xi}_i, \bar{\zeta}_i)$ *with* $\bar{\xi}_i^{\,2} = [(y_i - \boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}_p)^2 + \bar{\boldsymbol{\beta}}_p^T\boldsymbol{x}_i\boldsymbol{x}_i^T\bar{\boldsymbol{\beta}}_p/g]/2\bar{\sigma}$ *and* $\bar{\zeta}_i^{\,2} = 1/2\bar{\sigma}, i = 1, 2, \ldots, n_m$*. Let* $\boldsymbol{v}_m^{\dagger} = (\bar{v}_1, \ldots, \bar{v}_{n_m})^T, \boldsymbol{Y}_{pm}^{\dagger} = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m^{\dagger})$ *and* $\boldsymbol{V}_m^{\dagger} = diag(\boldsymbol{v}_m^{\dagger}), m = 1, 2, \ldots, M$.

Step 4 *for each subset, the corresponding likelihood can be represented as a form of* $NIG_k(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$ *distribution for* $(\boldsymbol{\beta}_p, \sigma)$*. Obtain the multiplicative distribution* $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$*, then the full data posterior is given by merging the g-prior* $NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0)$ *and the distribution* $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$*:*

$$NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p) = NIG_k(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_{g0}, a_0, b_0) NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p),$$

where $\bar{\boldsymbol{\mu}}_p = [(1+\frac{1}{g})\sum_{m=1}^{M} \boldsymbol{X}_m^{*T}\boldsymbol{V}_m^{\dagger-1}\boldsymbol{X}_m^{*})]^{-1}\sum_{m=1}^{M}\boldsymbol{X}_m^{*T}\boldsymbol{V}_m^{\dagger-1}\boldsymbol{Y}_{pm}^{\dagger}, \bar{\boldsymbol{\Lambda}} = (1+\frac{1}{g})\sum_{m=1}^{M}\boldsymbol{X}_m^{*T}\boldsymbol{V}_m^{\dagger-1}\boldsymbol{X}_m^{*}, \bar{a} = \frac{3n}{2}, \bar{b}_p = \frac{1}{2}[\sum_{m=1}^{M}\boldsymbol{Y}_{pm}^{\dagger*T}\boldsymbol{V}_m^{\dagger-1}\boldsymbol{Y}_{pm}^{\dagger*} - \bar{\boldsymbol{\mu}}_p^T\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\mu}}_p] + p(1-p)\sum_{m=1}^{M}\|\boldsymbol{v}_m^{\dagger}\|_1$ *and* $\|\cdot\|_1$ *denotes the* ℓ_1 *norm of a vector.*

6. Big data based algorithms for variable selection

6.1. Algorithm for Bayesian LASSO scale mixtures of normals regression

The *LASSO* of Tibshirani [26] was proposed to estimate linear regression coefficients using $L1$-penalized least squares. Consider the linear regression model (1), the *LASSO* shrinkage regression can be formulated as

$$\min_{\boldsymbol{\beta}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{k}|\beta_j|,$$

where λ is a non-negative penalization parameter. According to Tibshirani [26], the *LASSO* estimates can be interpreted as the posterior mode with independent and identical Laplace priors imposed on the regression coefficients.

Following Park and Casella [27], a conditional Laplace prior is given by

$$f(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{k} \frac{\lambda_j}{2\sqrt{\sigma^2}} \exp\{-\lambda_j |\frac{\beta_j}{\sqrt{\sigma^2}}|\},$$

where $\lambda_1, \ldots, \lambda_k$ are non-negative regularization parameters imposed on different regression coefficients. As suggested in Park and Casella [27], any inverse-Gamma prior for σ^2 would maintain conjugacy. Here we consider the marginal prior $f(\sigma^2) = IG(a_0, b_0)$, then the joint prior for $f(\beta, \sigma^2)$ is given by

$$f(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(a_0+\frac{k}{2}+1)} \exp\{-b_0\sigma^{-2} - \sum_{j=1}^{k} \lambda_j |\frac{\beta_j}{\sigma}|\}.$$

Given model (1), we have the posterior distribution

$$f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{\Sigma}) \propto (\sigma^2)^{-(a_0+\frac{n+k}{2}+1)} \exp\{-b_0\sigma^{-2} - \frac{1}{2}\sigma^{-2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) - \sum_{j=1}^{k} \lambda_j |\frac{\beta_j}{\sigma}|\}.$$

Following the equality given by Andrews and Mallows [28]

$$\frac{h}{2}\exp\{-h|z|\} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\{-z^2/(2s)\} \frac{h^2}{2} \exp\{-h^2 s/2\} ds, \; h > 0,$$

and introducing the latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ with prior $f(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2 \gamma_j}{2})$, we have the following Bayesian hierarchical model:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\Sigma} &\sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}), \\
\boldsymbol{\beta}|\sigma^2, \gamma_1, \ldots, \gamma_k &\sim N_k(\mathbf{0}_k, \sigma^2 \boldsymbol{D}_\gamma), \\
\boldsymbol{D}_\gamma &= \mathrm{diag}\,(\gamma_1, \ldots, \gamma_k), \\
\sigma^2, \gamma_1, \ldots, \gamma_k &\sim f(\sigma^2)\, d\sigma^2 \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2 \gamma_j}{2})\, d\gamma_j, \\
\sigma^2, \gamma_1, \ldots, \gamma_k &> 0.
\end{aligned}$$

Then we obtain the conditional prior distribution

$$f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\gamma}) \sim NIG_k(\mathbf{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0), \tag{19}$$

where $\boldsymbol{D}_\gamma^{-1} = \mathrm{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$. For the conditional posterior of $\boldsymbol{\gamma}$, we have $\gamma_j^{-1}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{Y}$ following an inverse-Gaussian distribution with parameters $\sqrt{\frac{\lambda_j^2\sigma^2}{\beta_j^2}}$ and λ_j^2 (see Park and Casella [27]). A corresponding Gibbs sampler algorithm can be provided as below.

Algorithm 6.1. *Consider the Bayesian LASSO scale mixtures of normals re-*

gression model with prior specification (19). Given the big data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we obtain the following Gibbs sampler algorithm.

Step 1 *the same as presented in* **Algorithm 5.1**.

Step 2 *for each subset, the corresponding likelihood has a representation of $NIG_k(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$ distribution for $(\boldsymbol{\beta}, \sigma^2)$. Calculate the multiplicative distribution $NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b) = \prod_{m=1}^{M} NIG(\boldsymbol{\mu}_m, \boldsymbol{\Lambda}_m, a_m, b_m)$, then iterate the following sub-steps until draws $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})$ achieve convergence.*

2.1 *given the current draw of $\boldsymbol{\gamma}$, compute $\boldsymbol{D}_\gamma^{-1} = \mathrm{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$; obtain posterior $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}) = NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0) NIG_k(\boldsymbol{\mu}, \boldsymbol{\Lambda}, a, b)$, where $\bar{\boldsymbol{\mu}} = (\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m)^{-1} \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m, \bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m, \bar{a} = a_0 + \frac{n}{2}, \bar{b} = b_0 + \frac{1}{2}[\sum_{m=1}^{M} \boldsymbol{Y}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{Y}_m - \bar{\boldsymbol{\mu}}^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}]$; then generate a draw of $(\boldsymbol{\beta}, \sigma^2)$ from $NIG_k(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b})$.*

2.2 *given the current draw of $(\boldsymbol{\beta}, \sigma^2)$, generate a draw for each γ_j^{-1} from the inverse-Gaussian distribution with parameters $\sqrt{\frac{\lambda_j^2 \sigma^2}{\beta_j^2}}$ and $\lambda_j^2, j = 1, 2, \ldots, k$.*

Remark. *In the high-dimensional setting ($k \gg n$), one can always choose proper prior matrix $\boldsymbol{D}_\gamma^{-1}$ such that $\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^{M} \boldsymbol{X}_m^T \boldsymbol{\Sigma}_m^{-1} \boldsymbol{X}_m$ is non-singular and therefore $\bar{\boldsymbol{\mu}}$ is well-defined.*

6.2. Algorithm for Bayesian LASSO quantile regression

Following the notations outlined in Section 3.1, the *LASSO* regularized quantile regression (Li and Zhu [29]) can be formulated by

$$\min_{\boldsymbol{\beta}_p} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_p) + \lambda \sum_{j=1}^{k} |\beta_{pj}|,$$

where $\boldsymbol{\beta}_p = (\beta_{p1}, \ldots, \beta_{pk})^T$ and $\lambda \geq 0$ is a penalization parameter. Consider a conditional Laplace prior

$$f(\boldsymbol{\beta}_p|\sigma) = \prod_{j=1}^{k} \frac{\lambda_j}{2\sqrt{\sigma}} \exp\{-\lambda_j |\frac{\beta_{pj}}{\sqrt{\sigma}}|\},$$

where $\lambda_1, \ldots, \lambda_k$ are non-negative penalization parameters and specify the marginal prior $f(\sigma) = IG(a_0, b_0)$, the prior for $f(\boldsymbol{\beta}_p, \sigma)$ is obtained by

$$f(\boldsymbol{\beta}_p, \sigma) \propto \sigma^{-(a_0 + \frac{k}{2} + 1)} \exp\{-b_0 \sigma^{-1} - \sum_{j=1}^{k} \lambda_j |\frac{\beta_{pj}}{\sqrt{\sigma}}|\}.$$

Consider further the reformulated linear QR model (18), we have the posterior distribution

$$f(\boldsymbol{\beta}_p, \sigma | \boldsymbol{v}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(a_0 + \frac{3n+k}{2} + 1)} \exp\{-\sigma^{-1}[b_0 + p\,(1-p) \sum_{i=1}^{n} v_i] \\ - \frac{1}{2}\sigma^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) - \sum_{j=1}^{k} \lambda_j |\frac{\beta_{pj}}{\sqrt{\sigma}}|\}.$$

Again, by introducing the latent variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)^T$ with the prior $f(\boldsymbol{\gamma}) = \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2 \gamma_j}{2})$, we have the following Bayesian hierarchical model:

$$\begin{aligned} \boldsymbol{Y}_p^* | \boldsymbol{\beta}_p, \sigma, \boldsymbol{v}, \boldsymbol{X}^* &\sim N_n(\boldsymbol{X}^*\boldsymbol{\beta}_p, \sigma \boldsymbol{V}), \\ \boldsymbol{\beta}_p | \sigma, \gamma_1, \ldots, \gamma_k &\sim N_k(\boldsymbol{0}_k, \sigma \boldsymbol{D}_\gamma), \\ \boldsymbol{D}_\gamma &= \mathrm{diag}\,(\gamma_1, \ldots, \gamma_k), \\ \sigma, \gamma_1, \ldots, \gamma_k &\sim f(\sigma)\, d\sigma \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp(-\frac{\lambda_j^2 \gamma_j}{2})\, d\gamma_j, \\ \sigma, \gamma_1, \ldots, \gamma_k &> 0. \end{aligned}$$

Then the conditional prior distribution can be denoted as

$$f(\boldsymbol{\beta}_p, \sigma | \boldsymbol{\gamma}) \sim NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0), \tag{20}$$

where $\boldsymbol{D}_\gamma^{-1} = \mathrm{diag}\,(\gamma_1^{-1}, \ldots, \gamma_k^{-1})$. For the conditional posterior of γ_j, we have $\gamma_j^{-1} | \boldsymbol{\beta}_p, \sigma, \boldsymbol{Y}_p^*$ following an inverse-Gaussian with parameters $(\sqrt{\frac{\lambda_j^2 \sigma}{\beta_{pj}^2}}, \lambda_j^2)$, $j = 1, \ldots, k$. The full conditional posterior of $\boldsymbol{\beta}_p$ is obtained by

$$f(\boldsymbol{\beta}_p | \sigma, \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \boldsymbol{D}_\gamma^{-1} \boldsymbol{\beta}_p]\}, \tag{21}$$

which can be expressed as an $N_k(\bar{\boldsymbol{\mu}}_p, \sigma \bar{\boldsymbol{\Lambda}}^{-1})$, where $\bar{\boldsymbol{\mu}}_p = [\boldsymbol{D}_\gamma^{-1} + \boldsymbol{X}^* \boldsymbol{V}^{-1} \boldsymbol{X}^*]^{-1} \boldsymbol{X}^* \boldsymbol{V}^{-1} \boldsymbol{Y}_p^*$ and $\bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \boldsymbol{X}^* \boldsymbol{V}^{-1} \boldsymbol{X}^*$. The full conditional posterior of σ is given by

$$f(\sigma | \boldsymbol{\beta}_p, \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{Y}_p^*, \boldsymbol{X}^*) \propto \sigma^{-(\frac{3n+k+2a_0}{2} + 1)} \exp\{-\frac{1}{2\sigma}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1}(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) \\ + \boldsymbol{\beta}_p^T \boldsymbol{D}_\gamma^{-1} \boldsymbol{\beta}_p + 2p(1-p) \sum_{i=1}^{n} v_i + 2b_0]\}, \tag{22}$$

which is an IG distribution with shape $\frac{3n+k+2a_0}{2}$ and scale $\frac{1}{2}[(\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p)^T \boldsymbol{V}^{-1} (\boldsymbol{Y}_p^* - \boldsymbol{X}^*\boldsymbol{\beta}_p) + \boldsymbol{\beta}_p^T \boldsymbol{D}_\gamma^{-1} \boldsymbol{\beta}_p + 2p(1-p) \sum_{i=1}^{n} v_i + 2b_0]$. The full posterior of each

$v_i, i = 1, 2, \ldots, n$ is also tractable:

$$\begin{aligned} f(v_i|\boldsymbol{\beta}_p, \sigma, y_i, \boldsymbol{x}_i) &\propto v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - (1-2p)v_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2] - \frac{p(1-p)}{\sigma}v_i\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{4\sigma}[v_i^{-1}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2 + v_i]\} \\ &= v_i^{-1/2} \exp\{-\frac{1}{2}(v_i^{-1}\bar{\xi}_i^{\,2} + v_i\bar{\zeta}_i^{\,2})\}, \end{aligned} \tag{23}$$

where $\bar{\xi}_i^{\,2} = (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_p)^2/2\sigma$ and $\bar{\zeta}_i^{\,2} = 1/2\sigma$, which can be recognized as a $GIG(\frac{1}{2}, \bar{\xi}_i, \bar{\zeta}_i)$. A corresponding Gibbs sampling algorithm can be presented as below.

Algorithm 6.2. *Consider a pth $(0 < p < 1)$ Bayesian LASSO regularized QR with prior calibration (20) and the big data $\boldsymbol{X}$ and $\boldsymbol{Y}$, we obtain the following Gibbs sampler algorithm.*

Step 1 *the same as presented in* **Algorithm 5.2**.

Step 2 *for each $\boldsymbol{X}_m, \boldsymbol{Y}_m$, a Gibbs sampler for sampling $\boldsymbol{\beta}_p, \sigma$, the $n_m \times 1$ latent vector $\boldsymbol{v}_m$ and $\boldsymbol{\gamma}$ follows the sub-steps below:*

2.1 *denote r as the iteration count. Then set $r = 0$ and establish $(\boldsymbol{\beta}_p^{(r=0)}, \sigma^{(r=0)}, \boldsymbol{v}_m^{(r=0)}, \boldsymbol{\gamma}^{(r=0)})$ to some starting values.*

2.2 *let $\boldsymbol{X}_m^* = \frac{1}{\sqrt{2}}\boldsymbol{X}_m$, $\boldsymbol{Y}_{pm}^* = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m)$, $\boldsymbol{V}_m = diag(\boldsymbol{v}_m)$ and $\boldsymbol{D}_\gamma = diag(\boldsymbol{\gamma})$.*

2.3 *follow the inverse-Gaussian conditional posterior of γ_j^{-1}, and the full conditional posteriors of $\boldsymbol{\beta}_p, \sigma, \boldsymbol{v}_m$ given in (21) - (23),*

(i) sample $\boldsymbol{v}_m^{(r+1)}$ from its GIG posterior $f(\boldsymbol{v}_m|\boldsymbol{\beta}_p^{(r)}, \sigma^{(r)})$.

(ii) sample $\boldsymbol{\gamma}^{(r+1)} = (\gamma_1^{(r+1)}, \ldots, \gamma_k^{(r+1)})^T$, where $1/\gamma_j^{(r+1)}$ follows an inverse-Gaussian with parameters $(\sqrt{\frac{\lambda_j^2\sigma^{(r)}}{(\beta_{pj}^{(r)})^2}}, \lambda_j^2), j = 1, \ldots, k$.

(iii) sample $\sigma^{(r+1)}$ from its IG posterior $f(\sigma|\boldsymbol{\beta}_p^{(r)}, \boldsymbol{v}_m^{(r+1)}, \boldsymbol{\gamma}^{(r+1)})$.

(iv) sample $\boldsymbol{\beta}_p^{(r+1)}$ from its multivariate normal posterior $f(\boldsymbol{\beta}_p|\sigma^{(r+1)}, \boldsymbol{v}_m^{(r+1)}, \boldsymbol{\gamma}^{(r+1)})$.

2.4 *set $r = r + 1$ and return to **Step 2.3** until $r = L$, where L is the number of iteration times.*

Step 3 *calculate the empirical estimates of the means $\bar{\boldsymbol{\beta}}_p, \bar{\sigma}$ and $\bar{\boldsymbol{\gamma}}$ based on the $(L - B)$ realizations of the Gibbs sequence (discarding the first B iterations as a burn-in). Then generate an n_m i.i.d. sample on $\bar{v}_i$, where $\bar{v}_i \sim$*

$GIG(\frac{1}{2}, \bar{\xi}_i, \bar{\zeta}_i)$ with ${\bar{\xi}_i}^2 = [(y_i - \boldsymbol{x}_i^T \bar{\boldsymbol{\beta}}_p)^2]/2\bar{\sigma}$ and ${\bar{\zeta}_i}^2 = 1/2\bar{\sigma}, i = 1, 2, \ldots, n_m$. Let $\boldsymbol{D}_\gamma^\dagger = diag(\bar{\boldsymbol{\gamma}}), \boldsymbol{v}_m^\dagger = (\bar{v}_1, \ldots, \bar{v}_{n_m})^T, \boldsymbol{Y}_{pm}^\dagger = \frac{1}{\sqrt{2}}(\boldsymbol{Y}_m - (1-2p)\boldsymbol{v}_m^\dagger)$ and $\boldsymbol{V}_m^\dagger = diag(\boldsymbol{v}_m^\dagger), m = 1, 2, \ldots, M$.

Step 4 *for each subset, the corresponding likelihood can be represented as a form of $NIG_k(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$ distribution for $(\boldsymbol{\beta}_p, \sigma)$. Obtain the multiplicative distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p) = \prod_{m=1}^M NIG(\boldsymbol{\mu}_{pm}, \boldsymbol{\Lambda}_m, a_m, b_{pm})$, then the full data posterior is given by merging the prior $NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0)$ and the distribution $NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p)$:*

$$NIG_k(\bar{\boldsymbol{\mu}}_p, \bar{\boldsymbol{\Lambda}}, \bar{a}, \bar{b}_p) = NIG_k(\boldsymbol{0}_k, \boldsymbol{D}_\gamma^{-1}, a_0, b_0) NIG_k(\boldsymbol{\mu}_p, \boldsymbol{\Lambda}, a, b_p),$$

where $\bar{\boldsymbol{\mu}}_p = [\boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{X}_m^*]^{-1} \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{Y}_{pm}^\dagger, \bar{\boldsymbol{\Lambda}} = \boldsymbol{D}_\gamma^{-1} + \sum_{m=1}^M \boldsymbol{X}_m^{*T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{X}_m^*, \bar{a} = \frac{3n+2a_0}{2}, \bar{b}_p = b_0 + \frac{1}{2}[\sum_{m=1}^M \boldsymbol{Y}_{pm}^{\dagger * T} \boldsymbol{V}_m^{\dagger -1} \boldsymbol{Y}_{pm}^{\dagger *} - \bar{\boldsymbol{\mu}}_p^T \bar{\boldsymbol{\Lambda}} \bar{\boldsymbol{\mu}}_p] + p(1-p) \sum_{m=1}^M \|\boldsymbol{v}_m^\dagger\|_1$ and $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector.

7. Numerical demonstrations and real-data analysis

In this section, we assess the performance of the proposed big data based algorithms for posterior distribution calculation through a series of numerical demonstrations and a real-world data analysis. All model runs and analyses were performed using R. The code files are available upon request.

7.1. Numerical demonstrations

7.1.1. Bayesian scale mixtures of normals regression

In the Bayesian scale mixtures of normals linear regression scenario, we generate data from a true model of the form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is a $10^6 \times 1$ response vector, $\boldsymbol{X}$ is a $10^6 \times 10^4$ predictor matrix with the first column assigned as a vector of all 1's and the remaining elements generated from $N(0, 1)$. $\boldsymbol{\beta}$ is a $10^4 \times 1$ vector where only the first 10 coefficients $(\beta_0, \ldots, \beta_9)^T = (10, 9, 8, 7, 6, 5, 4, 3, 2, 1)^T$ are set to be non-zero and σ^2 is set as $\sqrt{1.25}$. $\epsilon_i \stackrel{d}{=} \sqrt{\zeta_i} z_i, i = 1, \ldots, 10^6$ where z_i follows $N(0, 1)$ and ζ_i is an independent random variable generated from the uniform distribution $\mathcal{U}(0.5, \sqrt{5})$. We further specify an informative prior $NIG_{10^4}(\boldsymbol{0}, \boldsymbol{I}, 2, 1)$ for $(\boldsymbol{\beta}, \sigma^2)$ where $\boldsymbol{I}$ denotes the identity matrix. The whole data is partitioned into 100 subsets with each containing 10,000 observations. We implement **Algorithm 5.1** for the specified linear model and Table 1 reports the posterior means, standard deviations and 95% credible intervals for the non-zero coefficients $(\beta_0, \ldots, \beta_9)^T$. The simulation results indicate that our proposed big data based approach for the Bayesian scale mixtures of normals regression behaves well and provides an accurate estimation of the true regression coefficients.

Parameter	True Value	Mean	Std	95% CI	
				P2.5	P97.5
β_0	10	9.9662	0.0450	9.8769	10.0540
β_1	9	8.9525	0.0466	8.8622	9.0443
β_2	8	8.0115	0.0460	7.9206	8.1023
β_3	7	7.0212	0.0456	6.9320	7.1102
β_4	6	6.0759	0.0435	5.9911	6.1608
β_5	5	4.9944	0.0467	4.9030	5.0864
β_6	4	3.9454	0.0441	3.8582	4.0325
β_7	3	2.9999	0.0463	2.9092	3.0899
β_8	2	1.9993	0.0457	1.9106	2.0897
β_9	1	0.9729	0.0458	0.8829	1.0621

Table 1: Estimation results of the first 10 non-zero coefficients for the Bayesian scale mixtures of normals regression model.

7.1.2. Bayesian quantile regression

To investigate the performance of our proposed algorithms for the pth Bayesian quantile regression, we generate data from a true model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is a $10^6 \times 1$ response vector, $\boldsymbol{X}$ is a $10^6 \times 10^4$ design matrix with all elements generated from $N(0,1)$. $\boldsymbol{\beta} = (10, 9, 8, \dots, 1, 0, \dots, 0)^T$ is a $10^4 \times 1$ vector with only the first 10 coefficients set to be non-zero. $\boldsymbol{\epsilon}$ is the disturbance vector where $\epsilon_i \sim ALD(0, \sigma, p), i = 1, \dots, 10^6$ and σ is assigned as 0.1. We implement **Algorithm 5.2** for our big data BQR model at quantiles $p = 0.50$ and $p = 0.95$ respectively. In each scenario, the given full data is partitioned into 100 subsets with equal size of 10,000 and the Gibbs samplers are run for 15,000 iterations with a burn-in of 5000. An informative g-prior with $g = 100$ is specified, as suggested in Smith and Kohn (1996). Table 2 and 3 present the posterior means, standard deviations and 95% credible intervals of the non-zero coefficients for $p = 0.50$ and $p = 0.95$ respectively. The displayed numerical results show that our proposed big data based algorithms for the BQR model give a desirable estimation of the true coefficients.

7.2. Real-data analysis

In this section, we illustrate our divide-and-conquer algorithms for big data Bayesian quantile regression by a real-world data analysis. We use the airline on-time performance data from the 2009 ASA Data Expo, publicly available at `http://stat-computing.org/dataexpo/2009/the-data.html`. The data set has been used for a demonstration of massive data by Wang et al. [30] and Schifano et al. [31]. It consists of flight arrival and departure details for all commercial flights within the United States from October 1987 to April 2008. About 12 million flights were involved with 29 variables. Due to the computing limit, we only consider a complete sub-dataset of the year 2008 with $N = 584,583$ after removing all the missing records. We consider arrival delay (AD) as a continuous variable by modelling $\log(AD - \min(AD) + 1)$ and employ

Parameter	True Value	Mean	Std	95% CI	
				P2.5	P97.5
β_0	10	9.9201	0.0429	9.8355	10.0025
β_1	9	8.9235	0.0396	8.8461	9.0017
β_2	8	7.9363	0.0391	7.8596	8.0129
β_3	7	6.9333	0.0376	6.8590	7.0063
β_4	6	5.9282	0.0331	5.8638	5.9925
β_5	5	4.9604	0.0386	4.8859	5.0361
β_6	4	3.9523	0.0339	3.8860	4.0183
β_7	3	2.9767	0.0354	2.9072	3.0461
β_8	2	1.9761	0.0341	1.9092	2.0426
β_9	1	0.9944	0.0322	0.9311	1.0570

Table 2: Estimation results of the first 10 non-zero coefficients for the Bayesian quantile regression model at $p = 0.50$.

Parameter	True Value	Mean	Std	95% CI	
				P2.5	P97.5
β_0	10	9.6914	0.1711	9.3500	10.0240
β_1	9	8.8140	0.1718	8.4757	9.1524
β_2	8	8.0678	0.1708	7.7326	8.4005
β_3	7	6.9045	0.1611	6.5836	7.2187
β_4	6	5.6809	0.1565	5.3736	5.9869
β_5	5	4.8938	0.1718	4.5582	5.2296
β_6	4	3.7937	0.1547	3.4907	4.0929
β_7	3	3.0477	0.1616	2.7333	3.3663
β_8	2	2.0570	0.1624	1.7381	2.3724
β_9	1	1.0894	0.1606	0.7765	1.4029

Table 3: Estimation results of the first 10 non-zero coefficients for the Bayesian quantile regression model at $p = 0.95$.

a linear model that specifies the pth quantile of AD as follows:

$$Q_p(AD) = \beta_{p0} + \beta_{p1}HD + \beta_{p2}DIS + \beta_{p3}NF + \beta_{p4}WF + \epsilon,$$

where HD is the departure time (continuous, in hours), DIS is the distance (continuous, in thousands of miles), NF is the day/night flight indicator (binary; 1 if departure between 8 p.m. and 5 a.m., 0 otherwise) and WF is the weekend/weekday flight indicator (binary; 1 if departure occurred during the weekend, 0 otherwise). This model was also investigated by Schifano et al. [31].

We fit our big data BQR to the above specified regression model by implementing **Algorithm 5.2** at $p = 0.50$, 0.75 and 0.95 respectively. In each scenario, the whole observations are partitioned into 100 subsets with the size of $n_m = 5845$ for $m = 1, \ldots, 99$ and $n_{100} = 5928$. We assign the informative g-prior by choos-

	$p = 0.50$		$p = 0.75$		$p = 0.95$	
	Coeff	Std	Coeff	Std	Coeff	Std
Intercept	1.9483	3.3380	2.6819	3.3598	4.1028	2.9804
HD	0.0790	0.2038	0.0735	0.2014	0.0403	0.1709
DIS	-0.0577	1.5080	-0.0573	1.5440	-0.0150	1.4152
NF	-0.4222	3.0845	-0.3932	3.0592	-0.1398	2.6500
WF	-0.0545	1.9676	-0.0444	1.9923	-0.0372	1.8048

Table 4: Coefficient estimates and posterior standard deviations ($\times 10^3$) of big data BQR estimator for the airline on-time data.

ing $g = 100$. All results are based on 15,000 draws obtained from the Gibbs samplers with a burn-in of 5000 iterations. Table 4 presents the estimated coefficients and posterior standard deviations at the specified quantile levels. We observe that the departure time bears a positive association with the arrival delay, whereas the distance, night-time and weekend flights have negative effects on the delay across all the three quantiles considered. Nevertheless, the effects of these covariates are mitigated with the increase of quantile. Night-time flight is found to be a non-negligible factor to improve on-time performance of flights facing median and long arrival delays. This empirical study shows that our proposed BQR method facilitates the investigation of the effects of different factors on various levels of flight arrival delays in the big data scenario.

8. Conclusion

The methods of Bayesian scale mixtures of normals linear regression and Bayesian quantile regression for big data analysis, including variable selection and posterior predictive distributions, have been explored. This is achieved by using ALD-based working likelihood functions and conjugate NIG priors. The resulting algorithms are easily implemented and the numerical demonstrations show that the proposed approaches are promising.

References

[1] T. Cole, P. Green, Smoothing reference centile curves: the LMS method and penalized likelihood, Stat. Med. 11 (1992) 1305–1319.

[2] R. Koenker, K. Hallock, Quantile regression: An introduction, J. Econ. Perspect. 15 (2001) 43–56.

[3] K. Yu, Z. Lu, J. Stander, Quantile regression: Applications and current research areas, The Statistician 52 (2003) 331–350.

[4] L. Briollais, G. Durrieu, Application of quantile regression to recent genetic and -omic studies, Hum. Genet. 133 (2014) 951–966.

[5] M. Bernardi, G. Gayraud, L. Petrella, Bayesian tail risk interdependence using quantile regression, Bayesian Anal. 10 (2015) 553–603.

[6] Y. Wang, X.-N. Feng, X.-Y. Song, Bayesian quantile structural equation models, Struct. Equ. Model. 23 (2016) 246–258.

[7] T. Rodrigues, Y. Fan, Regression adjustment for noncrossing Bayesian quantile regression, J. Comput. Graph. Stat. 26 (2017) 275–284.

[8] L. Petrella, V. Raponi, Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress, J. Multivar. Anal. 173 (2019) 70–84.

[9] Y. Wu, G. Yin, Conditional quantile screening in ultrahigh-dimensional heterogeneous data, Biometrika 102 (2015) 65–76.

[10] L. Yu, N. Lin, L. Wang, A parallel algorithm for large-scale nonconvex penalized quantile regression, J. Comput. Graph. Stat. 26 (2017) 935–939.

[11] Y. Gu, J. Fan, L. Kong, S. Ma, H. Zou, Admm for high-dimensional sparse penalized quantile regression, Technometrics 60 (2018) 319–331.

[12] X. Chen, W. Liu, Y. Zhang, Quantile regression under memory constraint, Ann. Stat. 47 (2019) 3244–3273.

[13] R. Koenker, G. Bassett, Regression quantiles, Econometrica (1978) 33–50.

[14] K. Yu, R. A. Moyeed, Bayesian quantile regression, Stat. Probab. 54 (2001) 437–447.

[15] K. Yu, J. Stander, Bayesian analysis of a tobit quantile regression model, J. Econom. 137 (2007) 260–276.

[16] C. Reed, K. Yu, A partially collapsed Gibbs sampler for Bayesian quantile regression, Technical Report, Department of Mathematical Sciences, Brunel University, 2009.

[17] H. Kozumi, G. Kobayashi, Gibbs sampling methods for Bayesian quantile regression, J. Stat. Comput. Simul. 81 (2011) 1565–1578.

[18] O. E. Barndorff-Nielsen, N. Shephard, Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics, J. R. Stat. Soc., B: Stat. Methodol. 63 (2001) 167–241.

[19] R. Alhamzawi, K. Yu, Conjugate priors and variable selection for Bayesian quantile regression, Comput. Stat. Data Anal. 64 (2013) 209–219.

[20] A. Zellner, On assessing prior distributions and Bayesian regression analysis with g-prior distributions, In Bayesian Inference and Decision techniques. Stud. Bayesian Econometrics Statist 6 (1986) 233–243.

[21] M. Smith, R. Kohn, Nonparametric regression using Bayesian variable selection, J. Econom. 75 (1996) 317–343.

[22] K. E. Lee, N. Sha, E. R. Dougherty, M. Vannucci, B. K. Mallick, Gene selection: a Bayesian variable selection approach, Bioinformatics 19 (2003) 90–97.

[23] M. Gupta, P. Qu, J. G. Ibrahim, A temporal hidden Markov regression model for the analysis of gene regulatory networks, Biostatistics 8 (2007) 805–820.

[24] C. G. Bowsher, P. S. Swain, Identifying sources of variation and the flow of information in biochemical networks, Proc. Natl. Acad. Sci. U.S.A. 109 (2012) E1320–E1328.

[25] M. Roth, On the Multivariate t Distribution, Technical Report 3059, Linköping University, Automatic Control, 2012.

[26] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc., B: Stat. Methodol. 58 (1996) 267–288.

[27] T. Park, G. Casella, The Bayesian lasso, J. Am. Stat. Assoc. 103 (2008) 681–686.

[28] D. F. Andrews, C. L. Mallows, Scale mixtures of normal distributions, J. R. Stat. Soc., B: Stat. Methodol. 36 (1974) 99–102.

[29] Y. Li, J. Zhu, $l1$-norm quantile regression, J. Comput. Graph. Stat. 17 (2008) 163–185.

[30] C. Wang, M.-H. Chen, E. Schifano, J. Wu, J. Yan, Statistical methods and computing for big data, Stat. Its Interface 9 (2016) 399–414.

[31] E. D. Schifano, J. Wu, C. Wang, J. Yan, M.-H. Chen, Online updating of statistical inference in the big data setting, Technometrics 58 (2016) 393–403.