Bayesian Analysis (2016) **11**, Number 3, pp. 821–855

Bayesian Quantile Regression Based on the Empirical Likelihood with Spike and Slab Priors

Ruibin Xi*, Yunxiao Li†, and Yiming Hu‡

Abstract. In this paper, we consider nonparametric Bayesian variable selection in quantile regression. The Bayesian model is based on the empirical likelihood, and the prior is chosen as the "spike-and-slab" prior–a mixture of a point mass at zero and a normal distribution. We show that the posterior distribution of the zero coefficients converges to a point mass at zero and that of the nonzero coefficients converges to a normal distribution. To further address the problem of low statistical efficiency in extreme quantile regression, we extend the Bayesian model such that it can integrate information at multiple quantiles to provide more accurate inference of extreme quantiles for homogenous error models. Simulation studies demonstrate that the proposed methods outperform or perform equally well compared with existing methods. We apply this Bayesian method to study the role of microRNAs on regulating gene expression and find that the regulation of microRNA may have a positive effect on the gene expression variation.

Keywords: model selection, Gibbs sampler, oracle property, empirical process, consistency.

1 Introduction

Quantile regression (Koenker, 2005) provides a systematic and robust way of examining the dependence of the response variable on covariates. Unlike mean-based regression, quantile regression examines how covariates influence the conditional quantiles rather than the conditional mean of the response variable. For example, median regression considers the relationship between the median of the response and the covariates and is usually more robust than traditional mean-based regression. Quantile regression also provides a more comprehensive description of the relationship between the response variable and the covariates. In many applications (Okada and Samreth, 2012; Buchinsky, 1994; Fenske et al., 2011; Machado and Mata, 2005; Hulmán et al., 2015), covariates can have different effects on higher or lower quantiles of the response variable than the mean, which can be readily discovered by quantile regression but would be missed by mean regression. For example, Hulmán et al. (2015) showed that the gestational weight gain (GWG) had very heterogenous effects on birth weight. At the 0.05th quantile, a 1-kg difference in GWG corresponded to a 14.2 kg birth weight increase, but to a 29.0 g birth weight increase at the 0.95th quantile.

*School of Mathematical Sciences and Center for Statistical Sciences, Peking University, ruibinxi@math.pku.edu.cn
†Department of Biostatistics and Bioinformatics, Emory University
‡Department of Biostatistics, Yale University

 DOI: 10.1214/15-BA975

Given a sample $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, denote $\boldsymbol{Y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times p}$. The τth linear quantile regression is given by

$$\boldsymbol{y}_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_0 + u_i,$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown parameter and u_i's are independent random variables with their τth quantiles being 0. If $q_{y_i}(\tau|\boldsymbol{x}_i)$ denotes the τth quantile of y_i, we would have $q_{y_i}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_0$. It can be shown that the coefficient $\boldsymbol{\beta}_0$ can be estimated consistently by minimizing

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}), \tag{1}$$

where $\rho_\tau(t) = t(\tau - I(t < 0))$ is the check loss function.

In this paper, we consider Bayesian variable selection in quantile regression. Quantile regression is robust in that the distributions of the random errors u_i are not required to be parametric or to be the same distribution and that they can also be dependent on the covariates. Among many of their advantages, Bayesian methods can handle very complex problems via MCMC, they can easily combine different sources of information (including prior information), statistical inference of Bayesian methods can be naturally performed using the posteriors and sequential analysis is much easier in the Bayesian setting. Bayesian analysis of quantile regression would presumably inherit the merits of both quantile regression and Bayesian methods. However, Bayesian inference of quantile regression is difficult since it is not associated with a parametric error distribution. Recently, there have been active researches on Bayesian inference of quantile regression. If we assume that the errors u_i follow a skewed Laplace distribution, maximizing the likelihood is equivalent to minimizing the objective function of quantile regression. This strategy was first proposed by Yu and Moyeed (2001) and further investigated by many other researchers. Simulation studies show that this strategy can give reasonable results at various scenarios, but this version of Bayesian quantile regression loses its nonparametric merit, and some researches show that its predicted conditional quantiles can be far away from the true quantiles (Lin and Chang, 2012). Another approach for Bayesian inference of quantile regression is by using the Dirichelet Process (DP) (Kottas and Gelfand, 2001; Taddy and Kottas, 2010). The DP-based methods are very flexible in modeling error distributions, but they are computationally very demanding and lack theoretical result that can guarantee the performance of these methods.

We propose a nonparametric Bayesian variable selection method in quantile regression based on the empirical likelihood (Owen, 1988, 1991). Compared with the Bayesian methods of quantile regression based on the skewed-Laplace distribution, the EL-based Bayesian quantile regression does not assume a parametric distribution of the random errors and statistical inference would be more accurate for data whose distribution is far away from the skewed-Laplace distribution. The empirical likelihood (EL), though not a true likelihood, is known to behave like a true likelihood and has many good asymptotic properties. The EL for quantiles was first studied by Chen and Hall (1993) and further studied in Chen and Wong (2009). Bayesian inference based on the EL has been used previously in the literature. For example, Lazar (2003) proposed the Bayesian EL and

discussed its validity; Schennach (2005) considered Bayesian exponentially tilted EL; Kim and Yang (2011) studied Bayesian quantile regression with random effects based on the EL; Yang and He (2012) used Bayesian EL in quantile regression for parameter estimation. Note that the priors in Yang and He (2012) are assumed to have continuous density functions and hence their asymptotic results are not applicable to the model in this paper. Tang and Leng (2010) considered performing variable selection by minimizing the penalized log EL,

$$\log(L(\boldsymbol{\theta})) - n\sum_{j=1}^{p} p_\lambda(|\theta_j|), \tag{2}$$

where $L(\boldsymbol{\theta})$ is the EL and $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. The penalty can be LASSO (Tibshirani, 1996), Elastic Net (Zou and Hastie, 2005) or SCAD (Fan and Li, 2001) type of penalties. Though Tang and Leng (2010) showed that estimates by maximizing the penalized EL enjoys good asymptotic properties, it is computationally very expensive to minimize (2) for a given λ. In practice, cross-validation is often used to select the best λ, and one often needs to solve the minimization problem (2) multiple times, which makes it computationally even more expensive. In addition, the EL for quantile regression is not a differentiable function. Newton-type algorithms would generally not work for such a non-differentiable function, and to our best knowledge, there is no available efficient algorithm to solve the penalized EL of quantile regression.

To perform Bayesian variable selection, we put the spike-and-slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993) on the coefficient parameters in quantile regression and propose a hierarchical Bayesian model. The spike-and-slab prior was first proposed for performing Bayesian variable selection by Mitchell and Beauchamp (1988), where the prior consists of two components–the spike component and the slab component. The spike component is a point mass at 0 and the slab component is a uniform distribution on a finite interval. George and McCulloch (1993) instead used the spike-and-slab prior as a mixture of a small variance normal distribution and a large variance normal distribution. Similar spike-and-slab prior was also studied by Ishwaran and Rao (2005) and Ishwaran and Rao (2011). More recently, Narisetty and He (2014) developed a new Bayesian model with Gaussian-like spike-and-slab prior and showed that their method can guarantee model selection consistency even for normal models. In this paper, similar to Mitchell and Beauchamp (1988), we use the spike-and-lab prior with the spike as the point mass at 0 and the slab as a normal distribution. Such a choice of the spike-and-slab prior gives nonzero posterior probability of $\beta_i = 0$, and models can be directly compared based on their posterior probabilities. We show that under certain regularity conditions, the estimators given by this method have the property similar to the oracle property. More specifically, we show that the posterior distribution of zero parameters converges to the point mass at zero and the posterior distribution of the nonzero parameters converges to a normal distribution. The technique used in the proof may also be useful for proving asymptotic properties of Bayesian methods with the spike-and-slab prior as used in this paper. To avoid the daunting task of maximizing the posterior distribution, we develop a Gibbs sampler for parameter estimation. In each

iteration of the Gibbs sampler, we only need to calculate an EL for a given parameter $\boldsymbol{\beta}$ and thus we totally avoid the expensive step of maximizing the EL.

In quantile regression, the asymptotic variance of the parameter estimate is generally inversely proportional to the density of the error distribution at the quantile point. More specifically, if the error distributions have a density function f and a cumulative distribution function F, the asymptotic distribution of the parameter estimate $\hat{\boldsymbol{\beta}}_\tau$ in quantile regression is largely $N(\boldsymbol{\beta}_0, n^{-1}\tau(1-\tau)f^{-2}(\xi_\tau)H)$, where H is a matrix depending on the covariates and $\xi_\tau = F^{-1}(\tau)$. For extreme quantiles (τ being close to 0 or 1), the density function f at ξ_τ is usually small and the asymptotic variance of $\hat{\boldsymbol{\beta}}_\tau$ would be large when n is relatively small. In Bayesian quantile regression, this means that the variance of the posterior would be large for extreme quantiles when n is relatively small, making the Bayesian estimates less accurate. Under the homogeneous error assumption (the random errors follow a common distribution), the regression coefficients other than the intercept will be invariant at different quantiles. Thus, these parameters may be estimated at many different quantile points and the estimates given at different quantile points can have significantly different statistical efficiency. For example, consider the model $Y_i = \mu_1 + x_i\beta + u_i$, where u_i are i.i.d. random variables. The parameter β can be estimated by quantile regression at any quantiles since the random errors follows the same distribution. For bell shaped distributions, median regression would generally give estimates with smaller posterior variance than extreme quantile regression, although both estimates are estimators of β. If we can take full advantage of the information implied in multiple quantile points, we may derive a better estimator for the extreme quantiles. Based on this observation, we further develop a new EL-based Bayesian method that can accurately perform variable selection for extreme quantile regression under the homogeneous error assumption. This new Bayesian method is based on a new loss function that is a weighted sum of check loss functions at multiple quantiles. Simulation studies demonstrate that this method can give superior estimates for extreme quantiles than other methods under the homogeneous error assumption. We also prove an asymptotic result analogue to the single quantile regression scenario.

This paper is organized as follows. In Section 2, we introduce the Bayesian EL-based quantile regression model with the spike-and-slab prior, present its asymptotic properties and describe an efficient Gibbs sampler for estimating the parameters. Section 3 discusses the Bayesian method for accurate statistical inference of extreme quantile regression and its asymptotic properties. Simulation studies are performed in Section 4. In Section 5, we apply the method developed here to study the role of microRNA on regulating gene expression. We conclude the paper in Section 6. The proofs are all given in Appendix.

2 Bayesian hierarchical model for quantile regression

2.1 Bayesian hierarchical model based on the EL

EL was first introduced by Owen in his seminal work (Owen, 1988) for constructing confidence intervals for the mean, and later was extended to linear models (Owen,

1991) or general estimating equations (Qin and Lawless, 1994). A more comprehensive review of the EL can be found in Owen (2001) and Chen and Van Keilegom (2009). In general, given an estimating equation

$$\sum_{i=1}^{n} g(z_i, \boldsymbol{\beta}) = 0, \tag{3}$$

the EL is defined as

$$L(\boldsymbol{\beta}) = \sup\{\prod_{i=1}^{n} p_i \mid \sum_{i=1}^{n} p_i = 1, \ \sum_{i=1}^{n} p_i g(z_i, \boldsymbol{\beta}) = 0, \text{ and } 0 \le p_i \le 1\}. \tag{4}$$

Parameter estimates in quantile regression are defined by minimizing the objective function (1). However, if we take directional derivatives about $\boldsymbol{\beta}$, the estimates may also be defined as the solution to the following equation

$$\sum_{i=1}^{n} \phi_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})\boldsymbol{x}_i \approx 0,$$

where $\phi_\tau(t) = \tau - I_{[t<0]}$. Thus, we may define the EL for quantile regression as

$$L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) = \sup\{\prod_{i=1}^{n} p_i | \sum_{i=1}^{n} p_i \phi_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})\boldsymbol{x}_i = 0, \sum_{i=1}^{n} p_i = 1, 0 \le p_i \le 1\}. \tag{5}$$

We put a spike-and-slab prior $\pi(\beta_i|\theta_i, \sigma^2)$ on the regression coefficient β_i, where θ_i and σ^2 are parameters in the prior. Denote $\eta = \sigma^{-2}$. The hyper priors $\pi(\theta_i)$ and $\pi(\eta) = \pi(\sigma^{-2})$ for θ_i and and η are chosen as the uniform distribution and the Gamma distribution, respectively. More specifically, we assume the following Bayesian hierarchical model

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta} &\sim L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \\
&= \sup\{\prod_{i=1}^{n} p_i \mid \sum_{i=1}^{n} p_i \phi_\tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})\boldsymbol{x}_i = 0, \sum_{i=1}^{n} p_i = 1, 0 \le p_i \le 1\}, \\
\beta_i|\theta_i, \sigma^2 &\sim \theta_i I_{\{\beta_i = 0\}} + (1 - \theta_i) I_{\{\beta_i \ne 0\}} N(0, \sigma^2), \quad i = 1, \ldots, p, \\
\theta_i &\sim U(0, 1), \quad i = 1, \ldots, p, \\
\eta = \sigma^{-2} &\sim \Gamma(a, b), \quad a > 0, b > 0.
\end{aligned} \tag{6}$$

Thus, we get the quasi-posterior

$$f(\boldsymbol{\beta}, \boldsymbol{\theta}, \eta|\boldsymbol{X}, \boldsymbol{Y}; \alpha, \eta) \propto L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) \prod_{i=1}^{p} \pi(\beta_i|\theta_i, \eta) I_{(0,1)}(\theta_i) \Gamma(\eta; a, b), \tag{7}$$

where $\Gamma(\eta; a, b)$ is the density of $\Gamma(a, b)$ evaluated at η. In simulation and real-data analysis, we always set $a = 0.1$ and $b = 0.0005$, which gives roughly equal probabilities for $\eta < 1$ and $\eta > 1$. Note that as a and b tend to zero, the prior $\Gamma(a, b)$ will approximate

the noninformative prior $\pi(\eta) \propto \eta^{-1}$. However, we cannot directly use the noninformative prior $\pi(\eta) \propto \eta^{-1}$ since it will lead to an improper posterior. This is viewed as an unattractive characteristic by some statisticians including Berger (2006). However, our theoretical results (Section 2.2) guarantee that the proper prior $\Gamma(a, b)$ $(a, b > 0)$ can give consistent estimates of β_i's and our simulation studies also show the effectiveness of this prior. Therefore, we think this is a reasonable prior. Further arguments supporting that $\Gamma(a, b)$ is a reasonable prior can be found in Appendix.

2.2 Asymptotic property

We denote the estimating function as $m(\boldsymbol{x}, y, \boldsymbol{\beta}), \boldsymbol{x}, \boldsymbol{\beta} \in \mathbb{R}^p$, where

$$m(\boldsymbol{x}, y, \boldsymbol{\beta}) = \phi_\tau(y - \boldsymbol{x}^T\boldsymbol{\beta})\boldsymbol{x}.$$

Let $\bar{\boldsymbol{\beta}} = \text{argmax}\, L(\boldsymbol{\beta})$ be the MELE. The estimating function $m(\boldsymbol{x}, y, \boldsymbol{\beta})$ are not smooth in $\boldsymbol{\beta}$, but the expectations of $m(\boldsymbol{x}, y, \boldsymbol{\beta})$ and the EL function are sufficiently smooth under the following assumptions:

(Assumption 1) There exists a neighborhood Θ of $\boldsymbol{\beta}_0$ such that $P(L(\boldsymbol{\beta}) > 0) \to 1$ for any $\boldsymbol{\beta} \in \Theta$, as $n \to \infty$.

(Assumption 2) The distribution function of the p covariates, G_X has a bounded support $\mathcal{X}$.

(Assumption 3) The conditional distribution $F_X(t)$ of Y given X is twice continuously differentiable in t for all $X \in \mathcal{X}$.

(Assumption 4) At any $X \in \mathcal{X}$, the conditional density function $F'_X(t) = f_X(t) > 0$ for t in a neighborhood of $F_X^{-1}(\tau)$.

(Assumption 5) $E\{m(X, Y, \boldsymbol{\beta}_0)m(X, Y, \boldsymbol{\beta}_0)^T\}$ is positive definite.

We first introduce some notations before stating the main theorem. Given $\boldsymbol{\kappa} \in \{0, 1\}^p$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$, we denote $\boldsymbol{\beta}_{\boldsymbol{\kappa}}$ and $\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}$ to be the sub-vector of $\boldsymbol{\beta}$ corresponding to nonzero and zero elements of $\boldsymbol{\kappa}$, respectively. More specifically, if $i_1 < i_2 < \cdots < i_l$ are the indices of nonzero elements of $\boldsymbol{\kappa}$ and $j_1 < j_2 < \cdots < j_m$ are the indices of zero elements of $\boldsymbol{\kappa}$ $(l + m = p)$, we define $\boldsymbol{\beta}_{\boldsymbol{\kappa}} = (\beta_{i_1}, \ldots, \beta_{i_l})^T$ and $\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = (\beta_{j_1}, \ldots, \beta_{j_m})^T$. Given a $p \times p$ matrix A, $A_{\boldsymbol{\kappa}\boldsymbol{\kappa}}$ is denoted as the submatrix $(l \times l)$ of A such that its (s, t)th element $A_{\boldsymbol{\kappa}\boldsymbol{\kappa}}(s, t) = A(i_s, i_t)$. We can also define similar notations $A_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}$, $A_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}$ and $A_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}$.

Theorem 1. *Assume* $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0q}, 0, \ldots, 0)^T$ $(q \le p)$, *where* $\beta_{0j} \neq 0$ *for* $j = 1, \ldots, q$, *and* $\boldsymbol{\kappa}_1 = (1, \ldots, 1, 0, \ldots, 0)^T \in \mathbb{R}^p$ *with its first* q *elements being 1. Let* $\Omega = V_{12}^T V_{11}^{-1} V_{12}$ *with* $V_{11} = \tau(1-\tau)E(XX^T)$ *and* $V_{12} = -\partial E\{m(X, Y, \boldsymbol{\beta})\}/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$. *Under Assumptions 1–5, for any fixed* $a > 0, b > 0$, *if* $\boldsymbol{\beta}$ *follows the posterior distribution (7), we have that the posterior probability* $P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} = \mathbf{0} |\ \boldsymbol{Y}, \boldsymbol{X}) \to 1$ *in probability as* $n \to \infty$. *Furthermore, for any* $\boldsymbol{t} \in \mathbb{R}^q$, $P(\sqrt{n}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t} |\ \boldsymbol{Y}, \boldsymbol{X}) = F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}(\boldsymbol{t}) + o_p(1)$, *where*

$F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}$ is the cumulative distribution function of the normal distribution $\mathcal{N}(\mathbf{0}, \Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1})$ and $\bar{\boldsymbol{\beta}}$ is the MELE.

The proof of this theorem will be given in Appendix. Roughly speaking, Theorem 1 says that the variables selected by the Bayesian hierarchical model (6) are asymptotically correct and the posterior distribution for the nonzero parameters is approximately a normal distribution. Note that the asymptotic variance of the posteriors of the nonzero β is $\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}$. If we know which βs are nonzero in priori, the asymptotic variance of these βs is also $\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}$. Therefore, Theorem 1 is similar to the Oracle property. Analogous results for simultaneously estimating coefficients corresponding to a variety of τ's (e.g., the scenario considered in Yang and He (2012)) can be proved using similar techniques. Using Theorem 1, we have

Corollary 1. *Under the assumptions in Theorem 1, we have that the posterior probability of the true model $\mathcal{M}_0 = \{\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \neq 0, \boldsymbol{\beta}_{\bar{\kappa}_1} = 0\}$ converges to 1 in probability as $n \to \infty$.*

2.3 Gibbs sampler

Potentially, we may integrate (7) about η and θ_i to get the marginal posterior of $\boldsymbol{\beta}$ and get an estimate of $\boldsymbol{\beta}$ by maximizing this marginal posterior. However, it is computationally very expensive to maximize the posterior distribution, we instead use a Gibbs sampler to perform statistical inference based on the Bayesian model. The full conditional distribution of $\eta = \sigma^{-2}$ is

$$\begin{aligned} f(\eta|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Y}) &\propto \pi(\eta) \prod_{j=1}^{p} \pi(\beta_j|\theta_j, \eta) \\ &\propto (\eta)^{a-1} \exp(-b\eta) \prod_{j=1}^{p} \Big[\theta_j I_{\{\beta_j=0\}} \\ &\quad + (1-\theta_j) I_{\{\beta_j\neq 0\}} \frac{1}{\sqrt{2\pi\eta^{-1}}} \exp(-\frac{\beta_j^2}{2}\eta)\Big] \\ &\propto (\eta)^{a+h/2-1} \exp(-(b + \frac{1}{2}\sum_{j\in H} \beta_j^2)\eta), \end{aligned}$$

where $H = \{j : \beta_j \neq 0\}$ and $h = \#H$. Thus, the full conditional distribution of η is $\Gamma(a+h/2, b+\frac{1}{2}\sum_{j\in H}\beta_j^2)$. Denote $\boldsymbol{\theta}_{-j}$ and $\boldsymbol{\beta}_{-j}$ the sub-vectors of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ by removing their jth element, respectively. The full conditional distribution of θ_j is

$$\begin{aligned} f(\theta_j|\boldsymbol{\beta}, \boldsymbol{\theta}_{-j}, \eta, \boldsymbol{X}, \boldsymbol{Y}) &\propto \pi(\beta_j|\theta_j, \eta)\pi(\theta_j) \\ &\propto \big[\theta_j I_{\{\beta_j=0\}} + (1-\theta_j) I_{\{\beta_j\neq 0\}} \frac{1}{\sqrt{2\pi\eta^{-1}}} \exp(-\frac{\beta_j^2}{2}\eta)\big] I_{(0,1)}(\theta_j). \end{aligned}$$

Thus, the full conditional distribution of θ_j is $\text{Beta}(1+I(\beta_j=0), 1+I(\beta_j\neq 0))$, where $I(\cdot)$ is the indicator function. The full conditional distribution of β_j is

$$f(\beta_j|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}, \eta, \boldsymbol{\beta}_{-j}) \propto L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y})\pi(\beta_j|\theta_j, \eta).$$

Clearly, $f(\beta_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta},\eta,\boldsymbol{\beta}_{-j})$ is also a mixture of a continuous distribution and a point mass distribution at 0, but we cannot directly sample from this conditional distribution since $L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$ has no explicit form. To avoid this difficulty, we employ a Metropolis–Hastings (M–H) step in the Gibbs sampler for sampling from $f(\beta_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta},\eta,\boldsymbol{\beta}_{-j})$. The choice of the proposal distribution in the M–H step has an important impact on the efficiency of the Gibbs sampler.

Kim and Yang (2011) proposed to choose a pre-specified normal distribution such that the acceptance rates in the range of $0.1 \sim 0.4$ as the proposal distribution in their Gibbs sampler. In our setting, we may similarly use a pre-specified distribution–a mixture of a point mass distribution at zero and a continuous distribution–as the proposal distribution. However, we find that it is generally very hard to select a good pre-specified distribution and a poor selection of the proposal distribution can make the Gibbs sampler converge slowly (see Xi et al. (2015) Supplementary Figure S1). Another common practice for choosing the proposal distribution is to use the random walk Metropolis-algorithm (Tierney, 1994; Roberts et al., 1997). Specifically, given $\beta_i^{(t)}$ at the tth step, the proposal distribution can be set as $q(\beta - \beta_i^{(t)})$ with q being a simple symmetric random walk. For example, q may be taken as $N(0, s^2)$, where s^2 is usually chosen such that the acceptance rate is optimal (Roberts et al., 1997; Roberts and Rosenthal, 2001). However, the target distribution here is a mixture of a discrete distribution and a continuous distribution, the "random walk" in this scenario should have nonzero probability jumping back to zero. A reasonable choice would be a mixture of the point mass at zero and a true random walk such as $pI(\beta = 0) + (1-p)I(\beta \neq 0)N(\beta_i^{(t)}, s^2)$, but we have to tune two parameters to make acceptance rate optimal, which is quite difficult, in general.

If the EL $L(\boldsymbol{\beta})$ were smooth enough, given $\boldsymbol{\beta}_{-j}$, the Laplace approximation of the likelihood function $l(\beta_j) = \log(L(\beta_j, \boldsymbol{\beta}_{-j}))$ would be

$$l(\beta_j) \approx l(\bar{\beta}_j) + \frac{1}{2} l''(\bar{\beta}_j)(\beta_j - \bar{\beta}_j)^2$$

where $l(\cdot)$ is maximized at $\bar{\beta}_j$. Since the likelihood function is usually concave, we have $v_j^{-2} = -l''(\bar{\beta}_j) > 0$. Then, largely we would have

$$f(\beta_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta},\eta,\boldsymbol{\beta}_{-j}) \quad \propto \quad \exp\{-\frac{1}{2} v_j^{-2}(\beta_j - \bar{\beta}_j)^2\}\pi(\beta_j|\theta_j,\eta).$$

Thus, the posterior distribution $f(\beta_j|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta},\eta,\boldsymbol{\beta}_{-j})$ is also a mixture of the point mass at 0 and a normal distribution. Unfortunately, the likelihood $L(\boldsymbol{\beta})$ is not differentiable, and we cannot have the Laplace approximation by differentiating the likelihood, but this inspired us to choose the proposal distribution in the M–H step as $\phi(\beta_j|\bar{\beta}_j, v_j^2)\pi(\beta_j|\theta_j,\eta)$ with a proper $\bar{\beta}_j$ and v_j^2, where $\phi(\beta_j|\bar{\beta}_j, v_j^2)$ is the density of a normal distribution with mean $\bar{\beta}_j$ and variance v_j^2. Given $\boldsymbol{\beta}_{-j}$, let $\tilde{y}_i = y_i - \sum_{l\neq j} x_{il}\beta_l$. The best estimate of β_j given $\boldsymbol{\beta}_{-j}$ should minimize

$$\sum_{i=1}^{n} \rho_\tau(\tilde{y}_i - x_{ij}\beta_j). \tag{8}$$

In this paper, we propose to use $\bar{\beta}_j$ as the value that minimizes (8) and use v_j^2 as the variance estimate $\tilde{\sigma}^2$ of $\bar{\beta}_j$ by bootstrapping $(\tilde{y}_i, x_{ij})$. Note that if we consider the EL

$$L_{j1}(\beta_j) = \sup\{\prod_{i=1}^{n} p_i | \sum_{i=1}^{n} p_i \phi_\tau(\tilde{y}_i - x_{ij}\beta_j)x_{ij} = 0, \sum_{i=1}^{n} p_i = 1, 0 \le p_i \le 1\},$$

then, $\bar{\beta}_j$ maximizes $L_{j1}(\beta_j)$. If $\boldsymbol{\beta}_{-j} = \boldsymbol{\beta}_{0,-j}$, we have $L_{j1}(\beta_j) \propto \exp\{-\frac{1}{2}v_1^{-2}(\beta_j - \bar{\beta}_j)^2 + o_p(1)\}$ for some $v_1^2 > 0$, according to Theorem 3.2 in Yang and He (2012). Thus, our proposal distribution can be largely viewed as sampling from the distribution $L_{j1}(\beta_j)\pi(\beta_j|\theta_j, \eta)$. Clearly, $L_{j1}(\beta_j)$ is different from the likelihood $L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y})$ (here, we write $L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y}) = L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y})$ to emphasize that $\boldsymbol{\beta}_{-j}$ is given). However, if $\boldsymbol{\beta}_{-j} = \boldsymbol{\beta}_{0,-j}$, $\bar{\beta}_j$ should also largely maximize $L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y})$ since both $\bar{\beta}_j$ and the maximizer of $L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y})$ are consistent estimators of β_{0j} (Molanes Lopez et al., 2009). According to Lemma 6 of Molanes Lopez et al. (2009), we approximately have $L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y}) = \exp\{-\frac{1}{2}v^{-2}(\beta_j - \bar{\beta}_j)^2 + o_p(1)\}$ for some $v^2 > 0$. Thus, when n is large enough, both $L_{j1}(\beta_j)$ and $L(\beta_j, \boldsymbol{\beta}_{-j}|\boldsymbol{X}, \boldsymbol{Y})$ can be approximated by normal distributions with the same means. Our proposal distribution can be made to have similar shape as the full conditional distribution $f(\beta_j|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\theta}, \eta, \boldsymbol{\beta}_{-j})$ by replacing the bootstrapped variance $\tilde{\sigma}^2$ with $s\tilde{\sigma}^2$, where $s > 0$ is a fixed constant. In the simulation and the real data study below, we always use $s = 1$.

In each step of the Gibbs sampler, we have to solve the minimization problem (8) $B + 1$ times, where B is the number of bootstrap resamplings. Thus, it could be computational expensive if minimizing (8) is slow. Fortunately, we can use the following procedure to efficiently minimize (8). Let $b_i = \tilde{y}_i/x_{ij}$ (if $x_{ij} = 0$, we can just drop this observation). Assume that $b_1 \le \cdots \le b_n$ and let $i_0 = \max\{j : \sum_{i=j}^{n} |x_{ij}| \ge -(\tau - \frac{1}{2})\sum x_{ij} + \frac{1}{2}\sum_{i=1}^{n} |x_{ij}|\}$. Then, we have that $\bar{\beta}_j = b_{i_0}$ minimizes (8). Therefore, the computational complexity of minimizing (8) is $O(Bn\log(n))$. After sampling from the proposal distribution, we need to calculate the EL (5) to determine the acceptance probability. It is known that there may not be any $\{p_i \mid i = 1, \ldots, n\}$ satisfying the constraints. In the R-package we developed for this method, we instead use the adjusted EL as proposed in Chen et al. (2008) to avoid this problem.

2.4 Comparison with linear regression

The median regression can be viewed as a more robust alternative to the mean-based regression. For mean-based linear regression, we may also construct a similar Bayesian hierarchical model based on the EL to perform Bayesian variable selection. It would thus be interesting to compare the asymptotic variances of the EL-based Bayesian estimates for the linear regression and for the quantile regression. We first introduce the EL-based Bayesian hierarchical model for general estimating equations and present an asymptotic result for such hierarchical models. Considering the estimating equation (3) and the EL (4), we can have the following Bayesian hierarchical model

$$\boldsymbol{Z}|\boldsymbol{\beta} \sim L(\boldsymbol{\beta}|\boldsymbol{Z}) = \sup\{\prod_{i=1}^{n} p_i \mid \sum_{i=1}^{n} p_i g(\boldsymbol{z}_i, \boldsymbol{\beta})\boldsymbol{z}_i = 0, \sum_{i=1}^{n} p_i = 1, 0 \le p_i \le 1\},$$

$$\begin{aligned} \beta_i|\theta_i,\sigma^2 &\sim \theta_i I_{\{\beta_i=0\}} + (1-\theta_i) I_{\{\beta_i\neq 0\}} N(0,\sigma^2), \quad i=1,\ldots,p, \\ \theta_i &\sim U(0,1), \quad i=1,\ldots,p, \\ \eta=\sigma^{-2} &\sim \Gamma(a,b), \quad a>0, b>0. \end{aligned} \tag{9}$$

Analogously to the quantile regression, we have the following asymptotic result for the Bayesian hierarchical model (9).

Theorem 2. *Assume the true* $\boldsymbol{\beta}_0 = (\beta_{01},\ldots,\beta_{0q},0,\ldots,0)^T$ *(*$q \le p$*), where* $\beta_{0j}\neq 0$ *for* $j=1,\ldots,q$*, and* $\boldsymbol{\kappa}_1 = (1,\ldots,1,0,\ldots,0)^T \in \mathbb{R}^p$ *with its first* q *elements being 1. Suppose that* $E[g(Z,\boldsymbol{\beta}_0)g^T(Z,\boldsymbol{\beta}_0)]$ *is positive definite,* $\partial g(z,\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ *and* $\partial^2 g(x,\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T$ *are continuous in a neighborhood of* $\boldsymbol{\beta}_0$*,* $\|\partial g(z,\boldsymbol{\beta})/\partial\boldsymbol{\beta}\|$*,* $\|\partial^2 g(z,\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T\|$ *and* $\|g(z,\boldsymbol{\beta})\|^3$ *are bounded by some integrable function* $G(z)$ *in this neighborhood, and the rank of* $E[\partial g(Z,\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}]$ *is p. Let* $V_{11}=E[g(Z,\boldsymbol{\beta}_0)g^T(Z,\boldsymbol{\beta}_0)]$*,* $V_{12}=E[\partial g(Z,\boldsymbol{\beta})/\partial\boldsymbol{\beta}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}]$ *and* $\Omega = V_{12}^T V_{11}^{-1} V_{12}$*. Then, for any fixed* $a>0, b>0$*, if* $\boldsymbol{\beta}$ *follows the posterior distribution of the hierarchical model (9), we have the posterior probability* $P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} = \mathbf{0}|\ Z) \to 1$ *in probability as* $n\to\infty$*. Furthermore, for any* $\boldsymbol{t}\in\mathbb{R}^q$*,* $P(\sqrt{n}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t}|\ Z) = F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}(\boldsymbol{t}) + o_p(1)$*, where* $F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}$ *is the cumulative distribution function of the normal distribution* $\mathcal{N}(\mathbf{0},\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1})$ *and* $\bar{\boldsymbol{\beta}}$ *is the MELE that maximizes the EL (4).*

We now can compare the asymptotic variance of the EL-based Bayesian estimates of the quantile regression and the linear regression. For linear regression, we have $Z=(Y,X)$ and $g(\boldsymbol{z}_i,\boldsymbol{\beta}) = (y_i - \boldsymbol{x}_i^T\beta)\boldsymbol{x}_i$. Suppose that the random error $e_i = y_i - \boldsymbol{x}_i^T\beta$ has a density function $f(\cdot)$ and a variance $\sigma_e^2>0$. Assuming that X has a compact support, $E(XX^T)$ is positive definite. Also suppose that f is such that conditions in Theorem 1 or Theorem 2 are satisfied. For linear regression, asymptotically, the variance of the posterior distribution of the nonzero β_is is $\sigma_e^2[E(X_\kappa X_\kappa^T)]^{-1}$. For the quantile regression, the asymptotic variance of the posterior distribution of the nonzero β_is is $\tau(1-\tau)[E(X_\kappa X_\kappa^T)]^{-1}f^{-2}(0)$. Particularly, for the median regression, the asymptotic variance is $f^{-2}(0)[E(X_\kappa X_\kappa^T)]^{-1}/4$. If f is $N(0,\sigma_e^2)$, we have $f^{-2}(0)/4 = \pi\sigma_e^2/2$ and hence the mean-based linear regression has a smaller variance. If f is a t-distribution t_ν $(\nu>3)$, $\sigma_e^2 = \nu/(\nu-2)$ and $f^{-2}(0)/4 = \pi\nu\Gamma^2(\nu/2)\Gamma^{-2}((\nu+1)/2)/4$. When ν is large, $\sigma_e^2 = \nu/(\nu-2)$ would be smaller, but when μ is close to 3, $f^{-2}(0)/4$ would be smaller. For example, when $\nu=4$, $\sigma_e^2=2$ and $f^{-2}(0)/4 \simeq 1.78$, implying that median regression has a smaller asymptotic variance. Note that when $\nu\le 3$, the conditions in Theorem 2 do not hold, and we cannot guarantee the convergence of the EL-based Bayesian estimate of the linear regression; However, the conditions in Theorem 1 still hold, and we still have the asymptotic results for quantile regression. Similarly, if f is a Cauchy distribution, the conclusion in Theorem 2 does not hold but the conclusion in Theorem 1 is still valid. In this respect, the EL-based Bayesian quantile regression is more robust than the EL-based Bayesian linear regression.

3 Baeysian weighted mutiple-quantile regression

In this section, we focus on the model that satisfies

$$q_{y_i}(\tau_k|\boldsymbol{x}_i) = \mu_k + \boldsymbol{x}_i^T\boldsymbol{\beta}_0, \quad i=1,\ldots,n, \tag{10}$$

for $k = 1, \ldots, m$ (i.e., the parameter $\boldsymbol{\beta}_0$ is constant across different τ_k). Here, the notation $\boldsymbol{x}_i$ denotes the explanatory variables excluding the intercept term. Note that (10) holds for any τ if the errors u_i are homogeneous. Theorem 1 shows that the posterior distribution of the nonzero coefficients are approximately normally distributed. Consider the model $Y_i = \mu_1 + x_i\beta + u_i$ $(0 \neq \beta \in \mathbb{R})$, where u_i are assumed to be i.i.d. random variables with their τ_1th quantile ξ_{τ_1} being zero. Assume that the distribution F of u_i has a Lebesgue p.d.f. f. The asymptotic variance $v_{\tau_1}^2$ of the posterior distribution of β is inversely proportional to $f(\xi_{\tau_1})$ $(\xi_{\tau_1} = F^{-1}(\tau_1))$. Thus, if $f(\xi_{\tau_1})$ is small, $v_{\tau_1}^2$ would be large, making the estimate of β and the τ_1th quantile inaccurate when the sample size is small. Suppose that $f(\xi_{\tau_2}) > f(\xi_{\tau_1})$ for the τ_2th quantile $\xi_{\tau_2} = F^{-1}(\tau_2)$. Since u_i are i.i.d., we may consider the model $Y_i = \mu_2 + x_i^T\beta + w_i$, where $\mu_2 = \mu_1 + \xi_{\tau_2} - \xi_{\tau_1}$ and $w_i = u_i - (\xi_{\tau_2} - \xi_{\tau_1})$. This model can also give consistent estimate for β, but the asymptotic variance $v_{\tau_2}^2$ of this posterior distribution is inversely proportional to $f(\xi_{\tau_2})$ and hence smaller than $v_{\tau_1}^2$. The latter model can thus give a better estimate of β, but it cannot give estimate of the τ_1th quantile, which can be particularly interesting in certain applications. However, we can get estimates of the τ_1th and τ_2th quantile by minimizing $\sum_{i=1}^n [\rho_{\tau_1}(y_i - \mu_1 - x_i\beta) + \rho_{\tau_2}(y_i - \mu_2 - x_i\beta)]$. Since this new objective function uses information from both quantile points, the resulting estimate of β and the τ_1th quantile should be at least not worse than that given by minimizing $\sum_{i=1}^n \rho_{\tau_1}(y_i - \mu_1 - x_i\beta)$.

In general, suppose that $(y_i, \boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ are n independent observations. Assume that the conditional τ_kth quantile $q_{y_i}(\tau_k|\boldsymbol{x}_i)$ of y_i given $\boldsymbol{x}_i$ satisfy $q_{y_i}(\tau_k|\boldsymbol{x}_i) = \mu_{0k} + \boldsymbol{x}_i^T\boldsymbol{\beta}_0$, where $\tau_k \in (0,1)$ $(k = 1, \ldots, m)$. The weighted quantile objective function at τ_k $(k = 1, \ldots, m)$ is

$$\sum_{k=1}^m a_k \sum_{i=1}^n \rho_{\tau_k}(y_i - \mu_k - \boldsymbol{x_i}^T\boldsymbol{\beta}), \tag{11}$$

where $a_k \in (0,1)$ are the fixed weights for the τ_kth quantile. Similar to the usual quantile regression, we can define the corresponding EL as

$$L(\boldsymbol{\beta}, \boldsymbol{\mu}) = \sup\{\prod_{i=1}^n p_i | \sum_{k=1}^m a_k \sum_{i=1}^n p_i \phi_{\tau_k}(y_i - \mu_k - \boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i = \mathbf{0},$$
$$\sum_{i=1}^n p_i \phi_{\tau_k}(y_i - \mu_k - \boldsymbol{x}_i^T\boldsymbol{\beta}) = 0, \forall 1 \leq k \leq m, \sum_{i=1}^n p_i = 1, 0 \leq p_i \leq 1\}. \tag{12}$$

We have the following Bayesian hierarchal model:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta} &\sim L(\boldsymbol{\beta}, \boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{Y}), \\
\mu_k &\sim N(0, \sigma^2), \quad k = 1, \ldots, m, \\
\beta_i|\theta_i, \sigma^2 &\sim \theta_i I_{\{\beta_i = 0\}} + (1 - \theta_i) I_{\{\beta_i \neq 0\}} N(0, \sigma^2), \quad i = 1, \ldots, p, \\
\theta_i &\sim U(0,1), \quad i = 1, \ldots, p, \\
\eta = \sigma^{-2} &\sim \Gamma(a, b), \quad a > 0, b > 0.
\end{aligned} \tag{13}$$

For the Bayesian model (13), we can also have the asymptotic result similar to Theorem 1. We introduce some notations before stating the theorem. Let $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\mu}^T)^T$,

$\boldsymbol{\mu_0} = (\mu_{01}, \ldots, \mu_{0m})$ and $\boldsymbol{\zeta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\mu}_0^T)^T$. Denote $Q_{\boldsymbol{\beta}}(\boldsymbol{x}, y, \boldsymbol{\zeta}) = \sum_{k=1}^m a_k \phi_{\tau_k}(y - \mu_k - \boldsymbol{x}^T\boldsymbol{\beta})\boldsymbol{x}$, $Q_{\boldsymbol{\mu}}(\boldsymbol{x}, y, \boldsymbol{\zeta}) = (\phi_{\tau_1}(y - \mu_1 - \boldsymbol{x}^T\boldsymbol{\beta}), \ldots, \phi_{\tau_m}(y - \mu_m - \boldsymbol{x}^T\boldsymbol{\beta}))^T$ and $Q(\boldsymbol{x}, y, \boldsymbol{\zeta}) = (Q_{\boldsymbol{\beta}}(\boldsymbol{x}, y, \boldsymbol{\zeta})^T, Q_{\boldsymbol{\mu}}(\boldsymbol{x}, y, \boldsymbol{\zeta})^T)^T$. Thus, $EQ(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0) = \mathbf{0}$. We need the following regularity conditions for the asymptotic result:

(Assumption 1′) There exists a neighborhood Θ of $\boldsymbol{\zeta}_0$ such that equation $EQ(\boldsymbol{X}, Y, \boldsymbol{\zeta}) = 0$ has a unique solution and $P(L(\boldsymbol{\beta}, \boldsymbol{\mu}) > 0) \to 1$ for any $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \boldsymbol{\mu}) \in \Theta$, as $n \to \infty$.

(Assumption 4′) At any $X \in \mathcal{X}$, the conditional density function $F_X'(t) = f_X(t) > 0$ for t in a neighborhood of $F_X^{-1}(\tau_k)$ $(k = 1, \ldots, m)$.

(Assumption 5′) $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0)Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0)^T\}$ is positive definite.

Theorem 3. *Assume $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0q}, 0, \ldots, 0)^T$ $(q \le p)$, where $\beta_{0j} \neq 0$ for $j = 1, \ldots, q$, and $\boldsymbol{\kappa}_1 = (1, \ldots, 1, 0, \ldots, 0, 1, \ldots, 1)^T \in \mathbb{R}^{p+m}$ with its first q and last m elements being 1. Let $\Omega = V_{12}^T V_{11}^{-1} V_{12}$ with $V_{11} = E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0)Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0)^T\}$ and $V_{12} = -\partial E\{Q(X, Y, \boldsymbol{\zeta})\}/\partial\boldsymbol{\zeta}|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}$. Under Assumptions 1′,2,3,4′,5′, for any fixed $a > 0, b > 0$, if $\boldsymbol{\zeta} = (\boldsymbol{\beta}^T, \boldsymbol{\mu}^T)^T$ follows the posterior distribution of the Bayesian model (13), we have the posterior probability $P(\boldsymbol{\zeta}_{\bar{\boldsymbol{\kappa}}_1} = \mathbf{0} |\ \boldsymbol{Y}, \boldsymbol{X}) \to 1$ in probability as $n \to \infty$. Furthermore, for any $\boldsymbol{t} \in \mathbb{R}^{q+m}$, $P(\sqrt{n}(\boldsymbol{\zeta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\zeta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t} |\ \boldsymbol{Y}, \boldsymbol{X}) = F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}(\boldsymbol{t}) + o_p(1)$, where $F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}$ is the cumulative distribution function of the normal distribution $\mathcal{N}(\mathbf{0}, \Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1})$ and $\bar{\boldsymbol{\zeta}} = (\bar{\boldsymbol{\beta}}^T, \bar{\boldsymbol{\mu}}^T)$ is the MELE of (12).*

Similarly, we can also have the following corollary

Corollary 2. *Under the assumptions in Theorem 3, we have that the posterior probability of the true model $\mathcal{M}_0 = \{\boldsymbol{\zeta}_{\boldsymbol{\kappa}_1} \neq 0, \boldsymbol{\zeta}_{\bar{\kappa}_1} = 0\}$ converges to 1 in probability as $n \to \infty$.*

4 Simulation study

In this section, we use Monte Carlo simulations to study the performance of the EL-based Bayesian quantile regression with the spike-and-slab prior (BEQR) and the EL-based Bayesian weighted multiple-quantile regression (BEQR.W). We also compared our method with existing methods including linear regression with the lasso penalty (Lasso) (Tibshirani, 1996), quantile regression with the lasso penalty (qrLasso) (Li and Zhu, 2008) and Bayesian regularized quantile regression with the lasso penalty (bqrLasso) (Li et al., 2010). The data in the simulation studies are generated by

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_0 + u_i,\ \ i = 1, \ldots, n,$$

where $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)$ and u_i's τth quantile is equal to 0. We set the number of observations n as 200 and 500 in the following simulations.

4.1 Independent and identically distributed random errors

In this section, we consider the random errors u_i to be independently and identically distributed. The rows of $\boldsymbol{X}$ are generated independently from $N(\mathbf{0}, \boldsymbol{\Sigma})$, where the (i, j)th element of $\boldsymbol{\Sigma}$ is $0.5^{|i-j|}$. Within each simulation study, we consider five different choices for the distribution of u_i's.

- The first choice is a normal distribution $\mathcal{N}(\mu, \sigma^2)$, with $\mu = 0, \sigma^2 = 9$.
- The second choice is a Laplace distribution Laplace(μ, b), with $\mu = 0, b = 3$.
- The third choice is a mixture of two normal distributions, $0.6\mathcal{N}(\mu_1 - a, \sigma^2) + 0.4\mathcal{N}(\mu_2 - a, \sigma^2)$ where $\mu_1 = 2, \mu_2 = -2, \sigma^2 = 9$, and $a = 0.4$ is such that its mean equal to 0.
- The forth choice is a mixture of two Laplace distributions, $0.6\text{Laplace}(\mu_1 - a, b) + 0.4\text{Laplace}(\mu_2 - a, b)$ with μ_1, μ_2 and a are chosen as above.
- The fifth choice is a Cauchy distribution with the scale parameter being 1/5.

While performing the regression analysis, an intercept term is always added to the model. For each choice of the error distribution, we run 100 simulations. Priors for the Bayesian methods are taken to be almost noninformative. The quantiles for BEQR.W are $(0.1, 0.5, 0.9)$ and the weights are all set as 1. In the following, we use two criterions to evaluate the performance of each algorithm. The first criterion is the mean distance between the estimated conditional quantile and the true conditional quantile. Specifically, for the τth quantile, suppose that $\hat{\mu}_\tau$ and $\hat{\boldsymbol{\beta}}_\tau$ are the estimates of the intercept and the coefficient $\boldsymbol{\beta}$ by one of quantile methods, the mean of mean absolute deviation (MMAD) is defined as $\text{mean}(1/n \sum_{i=1}^{n} |\hat{\mu}_\tau + \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}_\tau - q_\tau - \boldsymbol{x_i}^T \boldsymbol{\beta}|)$, where q_τ is the τth quantile of the random errors u_i and the mean is taken over the 100 simulations. In terms of Lasso, suppose that $\hat{\mu}_0$ and $\hat{\boldsymbol{\beta}}$ are the estimates of intercept and the coefficient $\boldsymbol{\beta}$, and $\hat{\sigma}^2$ is the sample variance of the residuals, the estimated conditional quantile is taken as $\hat{\mu}_0 + \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}} + \hat{q}_\tau$, where $\hat{q}_\tau$ is the τth quantile of the normal distribution $\mathcal{N}(0, \hat{\sigma}^2)$. The MMAD for Lasso can then be similarly defined. The second criterion is the mean of true positives (TP) and False positives (FP) selected by each algorithm.

Table 1 shows the MMAD of each algorithm for simulations with homogeneous errors. BEQR and BEQR.W outperforms other methods in most cases. Especially, BEQR.W often has much smaller MMADs than other methods in estimating extreme quantiles. This is expected because BEQR.W fully utilizes the assumption that errors are unrelated with the covariates (which is true in this simulation). The asymptotic variance of the other quantile-based methods is inversely proportional to the density of the errors at the quantile point. When n is small, the corresponding asymptotic variance would be large and hence the estimates would be less accurate. On the other hand, since BEWQ.W borrows information from the quantiles with larger density, its asymptotic variance would be generally smaller for extreme quantiles and hence would give more accurate estimates. Table 2 shows the true positive rates and false positive rates for different algorithms. Note that the results for Lasso and BEQR.W are all the

n	quantile	Method	Error Distribution				
			normal	Laplace	normal mixture	Laplace mixture	Cauchy
200	$\tau = 0.9$	Lasso	0.43 (0.13)	0.79 (0.27)	0.61 (0.13)	0.80 (0.24)	12.78 (27.84)
		QR	0.84 (0.21)	1.46 (0.35)	0.35 (0.08)	0.52 (0.14)	0.31(0.15)
		qrLasso	0.72 (0.14)	1.16 (0.42)	0.29 (0.10)	0.44 (0.13)	0.30(0.18)
		bqrLasso	0.73 (0.17)	1.21 (0.36)	0.29 (0.07)	0.44 (0.13)	0.34(0.26)
		BEQR	0.53 (0.24)	1.17 (0.50)	**0.20** (0.08)	0.35 (0.21)	0.62 (0.80)
		BEQR.W	**0.41** (0.19)	**0.65** (0.31)	0.25 (0.10)	**0.29** (0.11)	**0.17**(0.14)
	$\tau = 0.5$	Lasso	0.40 (0.17)	0.57 (0.21)	0.80 (0.04)	0.85 (0.04)	0.87(1.00)
		QR	0.60 (0.16)	0.58 (0.15)	0.47 (0.14)	0.50 (0.16)	0.06(0.02)
		qrLasso	0.53 (0.17)	0.54 (0.17)	0.41 (0.14)	0.46 (0.17)	0.19(0.16)
		bqrLasso	0.51 (0.14)	0.51 (0.14)	0.44 (0.11)	0.48 (0.14)	0.06(0.02)
		BEQR	**0.37** (0.15)	**0.37** (0.19)	0.65 (0.26)	0.67 (0.24)	**0.03**(0.02)
		BEQR.W	0.39 (0.17)	0.42 (0.17)	**0.29** (0.12)	**0.33** (0.17)	0.06(0.02)
	$\tau = 0.1$	Lasso	0.43 (0.13)	0.78 (0.28)	0.66 (0.12)	0.45 (0.15)	12.75(27.71)
		QR	0.85 (0.19)	1.44 (0.39)	0.85 (0.21)	1.04 (0.25)	0.31(0.15)
		qrLasso	0.71 (0.24)	1.14 (0.32)	0.66 (0.24)	0.80 (0.30)	0.28(0.13)
		bqrLasso	0.73 (0.19)	1.18 (0.27)	0.71 (0.20)	0.84 (0.25)	0.34(0.22)
		BEQR	0.56 (0.23)	1.03 (0.50)	0.63 (0.25)	0.63 (0.43)	0.79(0.90)
		BEQR.W	**0.41** (0.18)	**0.56** (0.26)	**0.36** (0.18)	**0.41** (0.19)	**0.17**(0.12)
500	$\tau = 0.9$	Lasso	0.28 (0.09)	0.65 (0.22)	0.56 (0.09)	0.80 (0.17)	12.37(18.24)
		QR	0.53 (0.13)	0.90 (0.21)	0.22 (0.05)	0.33 (0.08)	0.19(0.06)
		qrLasso	0.44 (0.14)	0.79 (0.23)	0.18 (0.05)	0.27 (0.08)	0.20(0.12)
		bqrLasso	0.48 (0.12)	0.83 (0.22)	0.18 (0.04)	0.29 (0.08)	0.18(0.09)
		BEQR	0.33 (0.12)	0.72 (0.32)	**0.13** (0.05)	0.21 (0.08)	0.21(0.15)
		BEQR.W	**0.26** (0.11)	**0.42** (0.24)	0.15 (0.06)	**0.18** (0.08)	**0.08**(0.06)
	$\tau = 0.5$	Lasso	0.26 (0.09)	0.36 (0.12)	0.77 (0.01)	0.84 (0.01)	0.73(0.74)
		QR	0.40 (0.08)	0.34 (0.08)	0.28 (0.08)	0.32 (0.09)	0.04(0.01)
		qrLasso	0.33 (0.10)	0.31 (0.11)	0.25 (0.09)	0.33 (0.11)	0.20(0.18)
		bqrLasso	0.36 (0.08)	0.31 (0.08)	0.26 (0.07)	0.31 (0.10)	0.03(0.01)
		BEQR	**0.21** (0.09)	**0.20** (0.08)	0.21 (0.12)	0.30 (0.19)	**0.02**(0.01)
		BEQR.W	0.23 (0.08)	0.25 (0.09)	**0.15** (0.07)	**0.20** (0.09)	0.03(0.01)
	$\tau = 0.1$	Lasso	0.28 (0.09)	0.65 (0.23)	0.62 (0.09)	0.32 (0.09)	12.38(18.24)
		QR	0.54 (0.12)	0.91 (0.21)	0.49 (0.13)	0.60 (0.14)	0.21(0.06)
		qrLasso	0.45 (0.13)	0.72 (0.23)	0.42 (0.13)	0.50 (0.14)	0.21(0.12)
		bqrLasso	0.51 (0.12)	0.81 (0.19)	0.45 (0.12)	0.54 (0.13)	0.20(0.07)
		BEQR	0.34 (0.13)	0.66 (0.28)	0.32 (0.11)	0.38 (0.13)	0.22(0.18)
		BEQR.W	**0.26** (0.12)	**0.34** (0.16)	**0.22** (0.10)	**0.24** (0.11)	**0.08**(0.06)

Table 1: The MMAD of Lasso, qrLasso, bqrLasso and BEQR, BEQR.W for simulations with homogeneous errors. The numbers in the parentheses are the standard errors of the 100 MMADs in the corresponding scenarios.

same for different τ. So we only list these results once in Table 2. Here, a predictor is said to be selected by QR/bqrLasso if its 95% confidence/credible interval given by QR/bqrLasso does not cover 0; otherwise, it is unselected. We again see that generally BEQR.W performs the best in all cases.

n	τ/mean	Method	Error Distribution				
			normal TP/FP	Laplace TP/FP	normal mixture TP/FP	Laplace mixture TP/FP	Cauchy TP/FP
200	mean	Lasso	3.00/2.30	3.00/2.16	3.00/2.20	3.00/2.12	2.65/2.41
	$\tau = (0.9, 0.5, 0.1)$	BEQR.W	3.00/0.18	2.99/0.10	3.00/0.13	3.00/0.10	3.00/0.03
	$\tau = 0.9$	QR	2.82/0.22	2.17/0.23	3.00/0.15	2.98/0.16	2.99/0.10
		qrLasso	3.00/2.78	2.94/2.83	3.00/2.66	3.00/2.74	3.00/2.16
		bqrLasso	3.00/0.86	2.80/0.68	3.00/0.42	3.00/0.42	2.98/0.28
		BEQR	2.96/0.19	2.63/0.15	3.00/0.08	2.99/0.12	2.95/0.22
	$\tau = 0.5$	QR	2.98/0.16	2.98/0.13	3.00/0.16	3.00/0.13	3.00/0.10
		qrLasso	3.00/2.01	3.00/1.80	3.00/1.64	3.00/1.57	3.00/0.44
		bqrLasso	3.00/0.20	3.00/0.06	3.00/0.19	3.00/0.21	3.00/0.02
		BEQR	2.98/0.12	2.99/0.08	3.00/0.16	3.00/0.11	3.00/0.01
	$\tau = 0.1$	QR	2.77/0.15	2.05/0.24	2.78/0.23	2.52/0.18	3.00/0.10
		qrLasso	3.00/2.78	2.92/2.83	3.00/2.66	2.99/2.74	3.00/2.15
		bqrLasso	2.99/0.86	2.80/0.68	3.00/0.42	2.98/0.42	3.00/0.23
		BEQR	2.97/0.19	2.68/0.15	2.97/0.08	2.95/0.12	3.00/0.21
500	mean	Lasso	3.00/2.37	3.00/2.43	3.00/2.21	3.00/2.28	2.86/2.41
	$\tau = (0.9, 0.5, 0.1)$	BEQR.W	3.00/0.07	3.00/0.08	3.00/0.04	3.00/0.06	3.00/0.00
	$\tau = 0.9$	QR	3.00/0.26	2.77/0.21	3.00/0.22	3.00/0.21	3.00/0.11
		qrLasso	3.00/2.56	2.99/2.62	3.00/2.42	3.00/2.89	3.00/1.93
		bqrLasso	3.00/0.86	2.98/0.85	3.00/0.46	3.00/0.71	3.00/0.13
		BEQR	3.00/0.03	2.86/0.15	3.00/0.01	3.00/0.05	3.00/0.02
	$\tau = 0.5$	QR	3.00/0.20	3.00/0.17	3.00/0.22	3.00/0.17	3.00/0.14
		qrLasso	3.00/2.03	3.00/1.73	3.00/1.55	3.00/1.48	3.00/0.21
		bqrLasso	3.00/0.31	3.00/0.11	3.00/0.17	3.00/0.28	3.00/0.00
		BEQR	3.00/0.04	3.00/0.03	3.00/0.04	3.00/0.07	3.00/0.00
	$\tau = 0.1$	QR	2.99/0.31	2.80/0.16	2.99/0.11	2.97/0.14	3.00/0.29
		qrLasso	3.00/2.72	3.00/2.73	3.00/2.36	3.00/2.59	3.00/2.01
		bqrLasso	3.00/0.90	2.98/1.03	3.00/0.73	3.00/0.74	3.00/0.30
		BEQR	3.00/0.04	2.89/0.17	3.00/0.06	3.00/0.06	3.00/0.04

Table 2: Mean True positives (TP) and False positives (FP) given by Lasso, qrLasso, bqrLasso, BEQR and BEQR.W for simulations with homogeneous error.

4.2 Heterogenous random errors

Now we consider simulations with heterogenous random errors. The data are generated from the model

$$y_i = \beta_{10} x_{i1} + \sum_{j=2}^{8} \beta_{j0} x_{ij} + x_{i1}\epsilon_i \quad (i = 1, \ldots, n)$$

n	quantile	Method	Error Distribution				
			normal	Laplace	normal mixture	Laplace mixture	Cauchy
200	$\tau = 0.90$	Lasso	2.11 (0.24)	2.83 (0.30)	1.56 (0.14)	1.80 (0.20)	12.30 (51.51)
		QR	1.85 (0.35)	2.30 (0.48)	1.26 (0.12)	1.34 (0.17)	0.40 (0.11)
		qrLasso	1.08 (0.31)	1.57 (0.44)	0.48 (0.16)	0.64 (0.16)	**0.30**(0.12)
		bqrLasso	0.65 (0.21)	**1.08** (0.29)	0.25 (0.09)	0.38 (0.13)	**0.30**(0.18)
		BEQR	**0.49** (0.33)	1.16 (0.92)	**0.16** (0.10)	**0.33** (0.18)	0.40(0.42)
	$\tau = 0.50$	Lasso	0.65 (0.28)	0.92 (0.39)	1.05 (0.36)	1.17 (0.47)	0.87(0.90)
		QR	0.46 (0.18)	0.42 (0.16)	0.52 (0.14)	0.53 (0.13)	0.05(0.02)
		qrLasso	0.52 (0.24)	0.51 (0.22)	0.62 (0.25)	0.73 (0.28)	0.17(0.13)
		bqrLasso	0.44 (0.14)	0.41 (0.14)	0.32 (0.14)	**0.39** (0.20)	0.05(0.04)
		BEQR	**0.33** (0.16)	**0.33** (0.15)	**0.31** (0.23)	0.42 (0.33)	**0.03**(0.01)
	$\tau = 0.10$	Lasso	1.99 (0.20)	2.81 (0.41)	1.84 (0.20)	1.94 (0.25)	13.15(52.2)
		QR	1.80 (0.27)	2.40 (0.51)	1.77 (0.25)	1.82 (0.29)	0.43(0.15)
		qrLasso	0.65 (0.16)	1.25 (0.24)	0.61 (0.18)	0.70 (0.22)	**0.26**(0.18)
		bqrLasso	**0.57** (0.19)	**1.01** (0.28)	**0.55** (0.20)	**0.65** (0.26)	0.37(0.40)
		BEQR	0.62 (0.21)	1.12 (0.49)	0.61 (0.22)	0.69 (0.24)	0.60(0.50)
500	$\tau = 0.90$	Lasso	2.05 (0.13)	2.70 (0.15)	1.53 (0.07)	1.81 (0.11)	22.47(90.65)
		QR	1.82 (0.18)	2.32 (0.29)	1.26 (0.08)	1.37 (0.11)	0.35(0.08)
		qrLasso	0.95 (0.22)	1.32 (0.32)	0.45 (0.10)	0.54 (0.13)	0.24(0.09)
		bqrLasso	0.37 (0.13)	0.62 (0.26)	0.15 (0.05)	0.24 (0.08)	0.18(0.10)
		BEQR	**0.28** (0.15)	**0.46** (0.29)	**0.10** (0.06)	**0.16** (0.10)	**0.14**(0.11)
	$\tau = 0.50$	Lasso	0.41 (0.18)	0.55 (0.23)	0.94 (0.22)	1.04 (0.30)	0.89(0.82)
		QR	0.28 (0.11)	0.22 (0.07)	0.42 (0.07)	0.48 (0.09)	0.03(0.01)
		qrLasso	0.31 (0.14)	0.32 (0.15)	0.58 (0.14)	0.67 (0.19)	0.18(0.13)
		bqrLasso	0.26 (0.09)	0.23 (0.10)	0.18 (0.07)	0.23 (0.10)	0.03(0.03)
		BEQR	**0.20** (0.11)	**0.16** (0.10)	**0.14** (0.09)	**0.19** (0.14)	**0.01**(0.01)
	$\tau = 0.10$	Lasso	1.95 (0.10)	2.65 (0.15)	1.83 (0.12)	1.92 (0.11)	23.32(91.29)
		QR	1.84 (0.16)	2.27 (0.27)	1.78 (0.16)	1.85 (0.20)	0.37(0.08)
		qrLasso	0.55 (0.08)	1.03 (0.17)	0.50 (0.09)	0.54 (0.10)	**0.20**(0.16)
		bqrLasso	**0.35** (0.13)	**0.62** (0.22)	0.32 (0.11)	**0.40** (0.15)	0.23(0.30)
		BEQR	**0.35** (0.18)	0.64 (0.34)	**0.30** (0.18)	**0.40** (0.23)	0.21(0.15)

Table 3: The MMAD of Lasso, qrLasso, bqrLasso and BEQR for simulations with heterogeneous errors.

where ϵ_i are generated as in the i.i.d. random error case. The covariates x_{i1} are generated from uniform distribution on the interval $(0, 2)$ and $\boldsymbol{x}_i = (x_{i2}, \ldots, x_{i8})$ are generated from $N(0, \boldsymbol{\Sigma})$ with the (i, j)th element of $\boldsymbol{\Sigma}$ is $0.5^{|i-j|}$.

Clearly, the BEQR.W is not suitable for models with heterogenous errors, and so we did not include BEQR.W in this simulation study. Table 3 shows the MMAD of each algorithm under different simulation setups. The two Bayesian methods bqrLasso and BEQR generally perform better than the non-Bayesian methods. For example, for $\tau = 0.9$ and the normal mixture error model, the MMADs of Lasso, QR and qrLasso are 1.56, 1.26 and 0.48 which are all significantly larger than the MMAD of BEQR (0.16).

5 Real data analysis

In this section, we apply our BEQR method to study the role of microRNA (miRNA) regulation on the gene expression variation within a population. miRNAs are a class of small noncoding RNA that usually bind to the 3′ untranslated region (UTR) of their target messenger RNA (mRNA) transcripts to post-transcriptionally repress protein translation. In the literature, there have been contradicting opinions about the role of miRNA regulation on the gene expression variation. Earlier studies suggested that the miRNAs may have canalization effect on the gene expression, meaning that they can reduce the gene expression variance around a preset mean (Hornstein and Shomron, 2006; Wu et al., 2009). Recent study instead showed that genes targeted by miRNAs unexpectedly had increased expression variation compared with non-target genes (Lu and Clark, 2012). A natural measure for gene expression variation is the standard deviation of the gene expression in a population. However, the standard deviation of the gene expression is highly correlated with the mean gene expression; Genes with large gene expression mean often have large gene expression variance (Figure 1). Therefore, to account for this dependence, researchers often use the coefficient of variance – the ratio between the standard deviation and the mean of the gene expression – as a measure for the gene expression variance. Here, we instead use quantile regression to study this problem.

We obtained gene expression data from Lu and Clark (2012). This data set contains expression data of 22834 genes measured from 70 individuals. The genes with low expression level (mean expression less than 0.1) were removed since these genes are probably not expressed and the expression mean and variance of these genes largely only reflect the background noise. This filtering step gives us 14966 genes. The response variable of the regression is the standard deviation of gene expression in 70 individuals. The covariates include the mean gene expression (GeneExp), the gene length (GeneLen), the length of the 3′-UTR (Len3UTR), the number of miRNA targets in the 3′-UTR (Target) as predicted by TargetScan (Lewis et al., 2005), the mean target score of the miRNA targets (TargetScore), the number of common Single Nucleotide Polymorphism (SNP) in 3′-UTRs (NSNP) and the mean of minor allele frequencies of common SNPs in 3′-UTRs (MAF). We downloaded the miRNA targets sites and their associated target score from the UCSC genome browser (`http://genome.ucsc.edu/`). Common SNPs and their minor allele frequencies were also obtained from the UCSC genome browser. We included the SNPs in 3′-UTRs into the model because these SNPs may influence miRNA binding at 3′-UTRs and eventually affect the gene expression variance. Supplementary Table S1 shows the pairwise Spearman's correlation between these variables.

We first compare the conditional quantile predictions of the 5 methods, QR, Lasso, qrLasso, bqrLasso and BEQR. All variables are standardized so that they have mean 0 and standard deviation 1. We randomly partition the entire data sets into 5 subsets of roughly equal sizes. Each time, we use one subset as the testing set and the other 4 subsets as the training set. The penalty parameters for Lasso and qrLasso were estimated by cross-validation only based on the training data set. The hyper-parameters for bqrLasso and BEQR were chosen as before. After training, we calculate the mean check loss of the observations in the testing data. Table 5 shows the mean of the mean testing

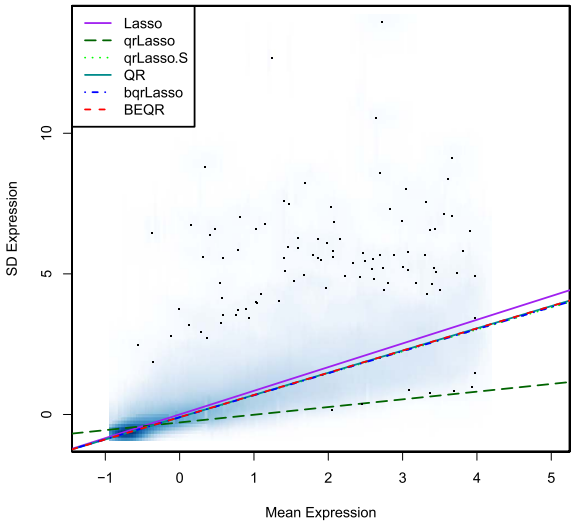

Figure 1: The smoothed scatter plot of the gene expression standard deviation versus the gene expression mean. The lines correspond to median regression or mean regression lines given by the methods discussed in the text. Note that the regression lines are very close to each other except Lasso and qrLasso.

errors of the 5 testing data sets. We can clearly see that the two Bayesian methods and the quantile regression perform very similarly and their testing errors are always smaller (often essentially smaller) than the two non-Bayesian regularization methods (Lasso and qrLasso). We also observe that the performance of qrLasso for this data set is not good even compared with the Lasso, but when we only use a subset of the observations (1000 observations), its performance is improved significantly. Therefore, to make the comparison fair, we also calculate the testing error of qrLasso trained on subsets of the training data set (qrLasso.S). More specifically, we randomly partition the training data set into 10 subsets and estimate the parameters by qrLasso (the tuning parameter λ is chosen by cross-validation in each subset). The parameter estimate of qrLasso.S is taken as the mean of the estimates based on each subset. With this modification, the performance of qrLasso is improved significantly and is similar to that of QR, bqrLasso and BEQR (Table 5 and Figure 1).

In the following, we will focus on the results of the median regression since we are most interested in the general role of miRNAs on the gene expression. If miRNAs have a canalization effect, we would expect the coefficient estimates of Target and/or TargetScore to be negative; If, instead, the regularization of miRNAs tends to increase the variation of gene expression, one of these two estimates would be positive. Figure 2 shows the parameter estimates and their 95% confidence/credible intervals (if any) of the 6 methods considered above based on the entire data set. The confidence interval of QR is given by the R-package *quantreg* with the *rank* option under the non-i.i.d. assumption. The parameter estimates given by Lasso and qrLasso generally lie far away

n	quantile	Method	Error Distribution				
			normal TP/FP	Laplace TP/FP	normal mixture TP/FP	Laplace mixture TP/FP	Cauchy TP/FP
200	mean	Lasso	3.00/2.31	3.00/2.37	3.00/2.33	3.00/2.44	2.79/2.46
	$\tau = 0.90$	QR	2.98/0.07	2.64/0.08	3.00/0.05	3.00/0.04	3.00/0.08
		qrLasso	3.00/1.42	2.89/1.48	3.00/1.05	3.00/1.41	3.00/1.82
		bqrLasso	3.00/0.45	2.95/0.46	3.00/0.19	3.00/0.22	3.00/0.16
		BEQR	2.99/0.15	2.71/0.21	3.00/0.06	2.98/0.05	3.00/0.12
	$\tau = 0.50$	QR	3.00/0.10	3.00/0.06	3.00/0.07	3.00/0.06	3.00/0.08
		qrLasso	3.00/1.79	3.00/1.37	3.00/0.88	3.00/0.95	3.00/0.46
		bqrLasso	3.00/0.13	3.00/0.06	3.00/0.06	3.00/0.02	3.00/0.00
		BEQR	3.00/0.11	3.00/0.11	3.00/0.05	3.00/0.06	3.00/0.01
	$\tau = 0.10$	QR	2.04/0.15	1.87/0.03	2.07/0.11	1.95/0.13	2.99/0.07
		qrLasso	2.38/2.01	2.50/2.24	2.50/2.14	2.46/2.51	3.00/0.49
		bqrLasso	2.29/0.26	2.42/0.33	2.28/0.27	2.20/0.29	2.93/0.10
		BEQR	2.38/0.20	2.41/0.40	2.38/0.27	2.33/0.42	3.00/0.10
500	Mean	Lasso	3.00/2.35	3.00/2.50	3.00/2.35	3.00/2.45	2.79/2.63
	$\tau = 0.90$	QR	3.00/0.11	3.00/0.08	3.00/0.05	3.00/0.06	3.00/0.15
		qrLasso	3.00/0.80	3.00/1.04	3.00/0.66	3.00/0.94	3.00/1.23
		bqrLasso	3.00/0.41	3.00/0.60	3.00/0.19	3.00/0.25	3.00/0.14
		BEQR	3.00/0.05	2.99/0.04	3.00/0.02	3.00/0.00	3.00/0.00
	$\tau = 0.50$	QR	3.00/0.10	3.00/0.04	3.00/0.11	3.00/0.08	3.00/0.17
		qrLasso	3.00/1.63	3.00/1.40	3.00/0.33	3.00/0.45	3.00/0.20
		bqrLasso	3.00/0.07	3.00/0.04	3.00/0.06	3.00/0.06	3.00/0.00
		BEQR	3.00/0.07	3.00/0.06	3.00/0.03	3.00/0.03	3.00/0.00
	$\tau = 0.10$	QR	2.32/0.06	2.65/0.03	2.34/0.10	2.31/0.04	3.00/0.22
		qrLasso	2.51/1.34	2.77/1.26	2.47/1.19	2.58/1.46	3.00/2.64
		bqrLasso	2.71/0.26	2.87/0.39	2.72/0.25	2.63/0.33	2.97/0.16
		BEQR	2.77/0.11	2.84/0.15	2.82/0.08	2.70/0.12	3.00/0.04

Table 4: Mean True positives (TP) and False positives (FP) given by Lasso, qrLasso, bqrLasso and BEQR for simulations with hetrogeneous errors.

from the estimates given by the other 4 methods. This is consistent with the above observation that Lasso and qrLasso do not provide good prediction and we will not discuss these results in the following.

The most significant factor influencing the variation of the gene expression is the mean expression. All methods give very significant (and similar) estimates (the BEQR estimate is 0.79012). The parameter estimates of Target given by Lasso.S, QR, bqrLasso and BEQR are all positive and their associated confidence/credible intervals (if any) do not contain zero. In fact, BEQR estimates that the posterior probability of Target being 0 is around 1.44%. These imply that the regularization of miRNAs may increase the gene expression variation, consistent with the analysis of Lu and Clark

Method	Error Distribution				
	$\tau = 0.1$	$\tau = 0.3$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 0.9$
Lasso	0.08109	0.12607	0.13156	0.16285	0.11509
qrLasso	0.03792	0.12233	0.20715	0.22757	0.13412
qrLasso.S	0.03750	0.09058	0.12150	0.13073	0.10171
QR	0.03749	0.09056	0.12144	0.12743	0.09053
bqrLasso	0.03750	0.09057	0.12143	0.12742	0.09053
BEQR	0.03750	0.09059	0.12146	0.12743	0.09062

Table 5: Mean testing errors as measured by the check loss function on the gene expression variance data.

(2012). However, the parameter estimate of Target is small in magnitude (the BEQR estimate is only 0.005231 and similar for the other methods). We must be cautious while interpreting these results. There are many possible reasons that can lead to this small magnitude effect. For example, it is possible that unknown sources of biases in measuring the RNA expression can lead to such a small magnitude estimate. Another possible reason is that the expression of miRNAs may vary significantly in the human population. Even though the effect of miRNAs is to stabilize the expression level of the genes that they regulate, the marginal gene expression variance can still appear to be increasing with more miRNA regularization. Clarification of this problem requires incorporating more data (e.g., miRNA expression) and precise biological experiments. For the effect of TargetScore, BEQR's estimate is 0 and its posterior probability of being zero is 99.7%; the confidence/credible intervals given by QR and bqrLasso cover 0; The qrLasso.S estimate is very close to 0 (4 of the 10 subsets gave 0 estimate).

Lu and Clark (2012) observed that genes with mutations in 3′-UTRs tend to have increased gene expression variation, which implies that mutations in 3′-UTRs might lead to greater gene expression variation. We also observe that number of SNPs in 3′-UTRs and the mean MAF of these SNPs positively correlate with the gene expression variation (Supplementary Table S1). However, in the median regression model, BEQR estimates the effects of NSNP and MAF to be 0 and the confidence/credible intervals given by bqrLasso and QR cover 0. This means that there is not enough evidence supporting the claim that mutations in 3′-UTRs have a direct effect on gene expression variation. In fact, the variables NSNP and MAF have a relatively strong correlation with the variable GeneLen (Supplementary Table S1) and GeneLen has a significant effect on the gene expression variation (Figure 2). Thus, marginally, we can observe positive correlation between NSNP/MAF and the gene expression variation, but this correlation disappears after we control for GeneLen.

6 Conclusion and discussion

In this paper, we propose a nonparametric Bayesian variable selection model in quantile regression based on the EL. The prior of the Bayesian model is chosen as the spike-and-slab prior. We theoretically show that the posterior distribution of the regres-

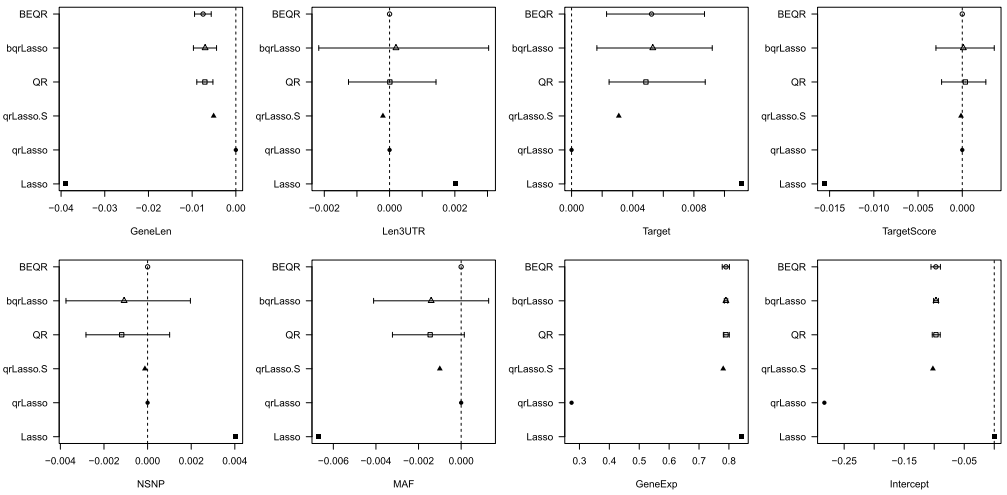

Figure 2: The parameter estimates of the gene expression variation data given by the 6 methods. The 95% confidence intervals (if any) are also shown. Note that for BEQR, the confidence interval may only contain 1 point (zero). In such cases, no confidence bar is shown in the plots.The dashed lines in the plots correspond to $\beta_i = 0$.

sion coefficients has an asymptotic property similar to the oracle property. An efficient MCMC algorithm is developed for the Bayesian quantile regression model. The most difficult part in the MCMC algorithm is to sample from the full conditional distribution of the regression coefficient β_i. Inspired by the idea of the Laplace approximation, we propose to apply an M–H step and use a mixture of a normal distribution and the point mass at zero as the proposal distribution in the M–H step to overcome this sampling difficulty. Simulation and real data analyses show that this method works equally well or better than current available methods. In addition, when the random errors u_i are homogenous, we propose a Bayesian weighted multiple-quantile regression method to improve the statistical efficiency for extreme quantile regression. Simulation studies show that this method generally outperforms the Bayesian quantile regression model at a single quantile point, especially for the extreme quantile points. We also prove an Oracle property for this Bayesian weighted multiple-quantile regression model.

When the number of parameters is greater than the number of observations ($p \geq n$), the EL (4) is generally not well-defined (or there is only one set of $\{p_i \mid i = 1, \ldots, n\}$ that satisfies the constraints). Hence, the current method would not work for the case $p \geq n$. Recently, Lahiri et al. (2012) proposed a new penalized EL (PEL) for testing population means. Their PEL allows p greater than n and has good asymptotic properties. This PEL can be extended in regression contexts. We may develop a corresponding Bayesian quantile regression model based on the PEL, but its asymptotic property is still not clear. Another way of circumventing the problem is to first screen for the important factors using correlation measures such as the distance correlation (Székely et al., 2007;

Li et al., 2012) and then use the method developed in this paper for more precise inference.

We have developed an accurate method for statistical inference for extreme quantile regression for homogeneous error models and further proved its asymptotic property. The loss function (11) takes similar form as the one proposed in Zou and Yuan (2008). However, we restrict the non-intercept coefficients to be the same at multiple quantiles and thus significantly improve the inference accuracy for extreme quantiles for homogeneous error models. In addition, Zou and Yuan (2008) did not provide any asymptotic theory about their method. The homogenous error assumption is less restrictive than it seems, because under a monotonically increasing transformation, the quantile of the transformed data is the same as the transformed quantile of the data. The method developed here can be easily extended to the model $Y = g(X^T\boldsymbol{\beta}) + u$, where u is the error with its τth quantile being zero and g is a known function. If the data does not satisfy the homogeneous error assumption, we may first apply a monotonically increasing transformation to the response to make the homogeneous error assumption largely hold. After performing the extreme quantile regression on the transformed data, we can make the inverse transformation to get back to the original data. Note that this type of transformation is often performed in ordinary linear regression, but if such transformation is hard to find, one can always use the method developed in Section 3 to give estimate of extreme quantiles.

Appendix

Proof of the asymptotic results

The proof of Theorem 1 is based on the following lemma.

Lemma 1. *Assume that $\boldsymbol{\beta}_0$ and $\boldsymbol{\kappa}_1$ are as in Theorem 1, and $\bar{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}_0$ with $\sqrt{n}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d \mathcal{N}(\mathbf{0}, \Sigma_1)$. Consider the hierarchical model (6) with the likelihood $L(\boldsymbol{\beta}|\boldsymbol{X}, \boldsymbol{Y})$ replaced by $\tilde{L}(\boldsymbol{\beta}) = \exp\{-\frac{n}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\}$, where $\Sigma = \Omega^{-1}$ is a positive definite matrix. Suppose that $\boldsymbol{\beta}$ follows the posterior distribution of this model. Then, we have $P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} = 0) \to_p 1$ and $P(\sqrt{n}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t}) = F_{\Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1}}(\boldsymbol{t}) + o_p(1)$.*

Proof. With the likelihood $\tilde{L}$, we have the conditional distribution $\tilde{f}(\boldsymbol{\beta}|\boldsymbol{\theta}, \eta)$ is

$$\begin{aligned}
\tilde{f}(\boldsymbol{\beta}|\boldsymbol{\theta}, \eta) &\propto \tilde{L}(\boldsymbol{\beta}) \prod_{j=1}^{p} \pi(\beta_j|\theta_j, \eta) \\
&= \exp\{-\frac{n}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\} \\
&\quad \prod_{j=1}^{p} \left[\theta_j I_{\{\beta_j = 0\}} + (1 - \theta_j) I_{\{\beta_j \neq 0\}} \frac{1}{\sqrt{2\pi\eta^{-1}}} \exp\{-\frac{\beta_j^2}{2}\eta\}\right] \\
&\propto \frac{1}{(2n\pi)^{p/2}\det(\Sigma)^{1/2}} \exp\{-\frac{n}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\}
\end{aligned}$$

$$\prod_{j=1}^{p}\big[\theta_j I_{\{\beta_j=0\}}+(1-\theta_j)I_{\{\beta_j\neq 0\}}\frac{1}{\sqrt{2\pi\eta^{-1}}}\exp\{-\frac{\beta_j^2}{2}\eta\}\big]$$
$$\propto\quad \phi(\boldsymbol{\beta}|\ \bar{\boldsymbol{\beta}},n^{-1}\Sigma)\prod_{j=1}^{p}\big[\theta_j I_{\{\beta_j=0\}}+(1-\theta_j)I_{\{\beta_j\neq 0\}}\frac{1}{\sqrt{2\pi\eta^{-1}}}\exp\{-\frac{\beta_j^2}{2}\eta\}\big],$$

where $\phi(\boldsymbol{\beta}|\ \bar{\boldsymbol{\beta}},n^{-1}\Sigma)$ is the density of the normal distribution $\mathcal{N}(\bar{\boldsymbol{\beta}},n^{-1}\Sigma)$. Denote $\kappa_j=I(\beta_j\neq 0)$ $(j=1,\ldots,p)$, $\boldsymbol{\kappa}=(\kappa_1,\ldots,\kappa_p)^T$ and $\Lambda_{\boldsymbol{\kappa}}=\mathrm{diag}\{\kappa_1,\ldots,\kappa_p\}$. Using the convention $0^0=1$, the prior $\prod_{i=1}^{p}\pi(\beta_i|\ \theta_i,\eta)$ may be written as

$$\begin{aligned}
&\prod_{j=1}^{p}\big[\theta_j I_{\{\beta_j=0\}}+(1-\theta_j)I_{\{\beta_j\neq 0\}}\frac{1}{\sqrt{2\pi\eta^{-1}}}\exp\{-\frac{\beta_j^2}{2}\eta\}\big]\\
=\ &\sum_{\kappa_1,\ldots,\kappa_p\in\{0,1\}}\prod_{j=1}^{p}\big((1-\theta_j)I_{\{\beta_j\neq 0\}}\frac{1}{\sqrt{2\pi\eta^{-1}}}\exp\{-\frac{\beta_j^2}{2}\eta\}\big)^{\kappa_j}\big(\theta_j I_{\{\beta_j=0\}}\big)^{1-\kappa_j}\\
=\ &\sum_{\boldsymbol{\kappa}\in\{0,1\}^p}\Big[\prod_{j=1}^{p}\big(\theta_j I_{\{\beta_j=0\}}\big)^{1-\kappa_j}\big((1-\theta_j)I_{\{\beta_j\neq 0\}}\big)^{\kappa_j}\Big]\\
&\times(2\pi\eta^{-1})^{-\sum_{j=1}^{p}\kappa_j/2}\exp\{-\frac{1}{2}\eta\sum_{j=1}^{p}\kappa_j\beta_j^2\}\\
=\ &\sum_{\boldsymbol{\kappa}\in\{0,1\}^p}\Big[\prod_{j=1}^{p}\big(\theta_j I_{\{\beta_j=0\}}\big)^{1-\kappa_j}\big((1-\theta_j)I_{\{\beta_j\neq 0\}}\big)^{\kappa_j}\Big]\\
&\times(2\pi\eta^{-1})^{-\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})/2}\exp\{-\frac{1}{2}\eta\boldsymbol{\beta}^T\Lambda_{\boldsymbol{\kappa}}\boldsymbol{\beta}^T\}\\
=\ &\sum_{\boldsymbol{\kappa}\in\{0,1\}^p}\omega(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\kappa})g(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\kappa},\eta),
\end{aligned}$$

where

$$w(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\kappa})=\prod_{j=1}^{p}\big(\theta_j I_{\{\beta_j=0\}}\big)^{1-\kappa_j}\big((1-\theta_j)I_{\{\beta_j\neq 0\}}\big)^{\kappa_j}$$

and

$$g(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\kappa},\eta)=(2\pi\eta^{-1})^{-\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})/2}\exp\{-\frac{1}{2}\eta\boldsymbol{\beta}^T\Lambda_{\boldsymbol{\kappa}}\boldsymbol{\beta}^T\}.$$

Given $\boldsymbol{\kappa}$, after a proper permutation, we have $\boldsymbol{\beta}^T=(\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T,\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T)$ and the matrix $\Omega=\Sigma^{-1}$ is

$$\Omega=\begin{pmatrix}\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\end{pmatrix}.$$

We have $\boldsymbol{\beta}^T\Omega\boldsymbol{\beta}=\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{\beta}_{\boldsymbol{\kappa}}+2\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}+\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}$ and $\boldsymbol{\beta}^T\Omega\bar{\boldsymbol{\beta}}=\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}}+\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}}+\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}\bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}}+\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}}$. Thus,

$$
\begin{aligned}
& \phi(\boldsymbol{\beta}|\ \bar{\boldsymbol{\beta}}, n^{-1}\Sigma) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\kappa}, \eta) \\
= \ & C_{n,\boldsymbol{\kappa}} \exp\{-\frac{n}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\} \exp\{-\frac{1}{2}\eta\boldsymbol{\beta}^T\Lambda_{\boldsymbol{\kappa}}\boldsymbol{\beta}\} \\
= \ & C_{n,\boldsymbol{\kappa}} \exp\Big\{ -\frac{1}{2}\Big[n\boldsymbol{\beta}^T\Sigma^{-1}\boldsymbol{\beta} + \eta\boldsymbol{\beta}^T\Lambda_{\boldsymbol{\kappa}}\boldsymbol{\beta}^T - 2n\boldsymbol{\beta}^T\Sigma^{-1}\bar{\boldsymbol{\beta}} + n\bar{\boldsymbol{\beta}}^T\Sigma^{-1}\bar{\boldsymbol{\beta}}\Big]\Big\} \\
= \ & C_{n,\boldsymbol{\kappa}} \exp\Big\{ -\frac{1}{2}\Big[n\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{\beta}_{\boldsymbol{\kappa}} + 2\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} + \boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}\big) + \eta\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\boldsymbol{\beta}_{\boldsymbol{\kappa}} \\
& -2n\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}} + \boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}} + \boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}\bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}} + \boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}}^T\Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}}\big) + n\bar{\boldsymbol{\beta}}^T\Omega\bar{\boldsymbol{\beta}}\Big]\Big\}
\end{aligned}
$$

where $C_{n,\boldsymbol{\kappa}} = \frac{1}{\sqrt{(2\pi)^{p-1}\det(n^{-1}\Sigma)}}(2\pi\eta^{-1})^{-\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})/2}$. Define $C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}) = C_{n,\boldsymbol{\kappa}}\prod_{\kappa_i=0}\theta_i\prod_{\kappa_i\neq 0}(1-\theta_i)$, $\tilde{\Omega}_{n,\boldsymbol{\kappa}}(\eta) = \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} + n^{-1}\eta I_{\boldsymbol{\kappa}}$, $\boldsymbol{b} = \Omega\bar{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta) = \tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)\boldsymbol{b}_{\boldsymbol{\kappa}}$. Then, almost surely we have

$$
\begin{aligned}
& \phi(\boldsymbol{\beta}|\ \bar{\boldsymbol{\beta}}, n^{-1}\Sigma) w(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\kappa}) g(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\kappa}, \eta) \\
= \ & C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}) \exp\Big\{ -\frac{1}{2}\Big[n\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{\beta}_{\boldsymbol{\kappa}}\big) + \eta\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\boldsymbol{\beta}_{\boldsymbol{\kappa}} \\
& -2n\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}} + \boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}}\big) + n\bar{\boldsymbol{\beta}}^T\Omega\bar{\boldsymbol{\beta}}\Big]\Big\} I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0) \\
= \ & C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}) \exp\Big\{ -\frac{1}{2}\Big[\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\big(n\Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} + \eta I_{\boldsymbol{\kappa}}\big)\boldsymbol{\beta}_{\boldsymbol{\kappa}} - 2n\boldsymbol{\beta}_{\boldsymbol{\kappa}}^T\boldsymbol{b}_{\boldsymbol{\kappa}} + n\bar{\boldsymbol{\beta}}^T\Omega\bar{\boldsymbol{\beta}}\Big]\Big\} I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0) \\
= \ & C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}) \exp\Big\{ -\frac{1}{2}n\Big[\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}} - \tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta)\big)^T\tilde{\Omega}_{n,\boldsymbol{\kappa}}(\eta)\big(\boldsymbol{\beta}_{\boldsymbol{\kappa}} - \tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta)\big) \\
& -\tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta)^T\tilde{\Omega}_{n,\boldsymbol{\kappa}}(\eta)\tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta) + \bar{\boldsymbol{\beta}}^T\Omega\bar{\boldsymbol{\beta}}\Big]\Big\} I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0) \\
= \ & C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta) \exp\Big\{ -\frac{1}{2}n\big(\boldsymbol{b}^T\Omega^{-1}\boldsymbol{b} - \boldsymbol{b}_{\boldsymbol{\kappa}}^T\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)\boldsymbol{b}_{\boldsymbol{\kappa}}\big)\Big\} \\
& \times\phi(\boldsymbol{\beta}_{\boldsymbol{\kappa}}|\ \tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta), n^{-1}\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)) I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0) \\
= \ & \tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta)\phi(\boldsymbol{\beta}_{\boldsymbol{\kappa}}|\ \tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta), n^{-1}\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)) I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0)
\end{aligned}
$$

where $\tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta) = C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta})\sqrt{(2\pi)^{\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})}\det(n^{-1}\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta))}\exp\{-\frac{1}{2}n(\boldsymbol{b}^T\Omega^{-1}\boldsymbol{b} - \boldsymbol{b}_{\boldsymbol{\kappa}}^T\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)\boldsymbol{b}_{\boldsymbol{\kappa}})\}$ and $C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta) = C_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta})\sqrt{(2\pi)^{\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})}\det(n^{-1}\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta))}$. From this, we get that the posterior likelihood $\tilde{f}(\boldsymbol{\beta}|\boldsymbol{\theta}, \eta)$ is a mixture of normal distributions on the planes $\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0$, i.e.,

$$
\tilde{f}(\boldsymbol{\beta}|\boldsymbol{\theta}, \eta) = \sum_{\boldsymbol{\kappa}\in\{0,1\}^p} \pi_{n\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta)\phi(\boldsymbol{\beta}_{\boldsymbol{\kappa}}|\ \tilde{\boldsymbol{\beta}}_{n,\boldsymbol{\kappa}}(\eta), n^{-1}\tilde{\Omega}_{n,\boldsymbol{\kappa}}^{-1}(\eta)) I(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}} = 0),
$$

where $\pi_{n\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta) = \tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta)/\sum_{\boldsymbol{\kappa}\in\{0,1\}^p}\tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta}, \eta)$. Denote $\Omega_{\boldsymbol{\kappa},L}$ be a $p\times p$ matrix, with $\Omega_{\boldsymbol{\kappa},L}(i,j) = 0$ if either $\kappa_i = 0$ or $\kappa_j = 0$ and $\Omega_{\boldsymbol{\kappa},L}(i,j) = \Omega(i,j)$ for $\kappa_i = 1$ and

$\kappa_j = 1$. Let $\Omega^-_{\boldsymbol{\kappa},L}$ be the Moore–Penrose pseudoinverse matrix of $\Omega_{\boldsymbol{\kappa},L}$. Then, up to a permutation, we have

$$\Omega_{\boldsymbol{\kappa},L} = \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \Omega^-_{\boldsymbol{\kappa},L} = \begin{pmatrix} \Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & 0 \\ 0 & 0 \end{pmatrix}.$$

Denote $\Sigma_{\boldsymbol{\kappa},D} = \Omega^{-1} - \Omega^-_{\boldsymbol{\kappa},L}$ and $K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} = \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} - \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}\Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}}$. We have

$$\begin{aligned}
\Omega\Sigma_{\boldsymbol{\kappa},D}\Omega &= \Omega - \Omega\Omega^-_{\boldsymbol{\kappa},L}\Omega \\
&= \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} - \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} \begin{pmatrix} \Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} \\
&= \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} - \begin{pmatrix} I_{\boldsymbol{\kappa}} & 0 \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}\Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & 0 \end{pmatrix} \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} \\
&= \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix} - \begin{pmatrix} \Omega_{\boldsymbol{\kappa}\boldsymbol{\kappa}} & \Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \\ \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}} & \Omega_{\bar{\boldsymbol{\kappa}}\boldsymbol{\kappa}}\Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\Omega_{\boldsymbol{\kappa}\bar{\boldsymbol{\kappa}}} \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 \\ 0 & K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}} \end{pmatrix}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
S_{\boldsymbol{\kappa}}(\eta) &= \exp\left\{-\frac{1}{2}n\big(\boldsymbol{b}^T\Omega^{-1}\boldsymbol{b} - \boldsymbol{b}^T_{\boldsymbol{\kappa}}\tilde{\Omega}^{-1}_{n,\boldsymbol{\kappa}}(\eta)\boldsymbol{b}_{\boldsymbol{\kappa}}\big)\right\} \\
&= \exp\left\{-\frac{1}{2}n\big(\boldsymbol{b}^T\Omega^{-1}\boldsymbol{b} - \boldsymbol{b}^T\Omega^-_{\boldsymbol{\kappa},L}\boldsymbol{b} + \boldsymbol{b}^T_{\boldsymbol{\kappa}}\Omega^{-1}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{b}_{\boldsymbol{\kappa}} - \boldsymbol{b}^T_{\boldsymbol{\kappa}}\tilde{\Omega}^{-1}_{n,\boldsymbol{\kappa}}(\eta)\boldsymbol{b}_{\boldsymbol{\kappa}}\big)\right\} \\
&= \exp\left\{-\frac{1}{2}n\big(\boldsymbol{b}^T\Sigma_{\boldsymbol{\kappa},D}\boldsymbol{b} + n^{-1}\eta\boldsymbol{b}^T_{\boldsymbol{\kappa}}\Omega^{-2}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{b}_{\boldsymbol{\kappa}} + O(n^{-2})\big)\right\} \\
&= \exp\left\{-\frac{1}{2}n\big(\bar{\boldsymbol{\beta}}^T\Omega\Sigma_{\boldsymbol{\kappa},D}\Omega\bar{\boldsymbol{\beta}} + n^{-1}\eta\boldsymbol{b}^T_{\boldsymbol{\kappa}}\Omega^{-2}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{b}_{\boldsymbol{\kappa}} + O(n^{-2})\big)\right\} \\
&= \exp\left\{-\frac{1}{2}n\big(\bar{\boldsymbol{\beta}}^T_{\bar{\boldsymbol{\kappa}}}K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}} + n^{-1}\eta\boldsymbol{b}^T_{\boldsymbol{\kappa}}\Omega^{-2}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{b}_{\boldsymbol{\kappa}} + O(n^{-2})\big)\right\}.
\end{aligned}$$

Since $K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}$ is positive definite, if $\boldsymbol{\beta}_{0\bar{\boldsymbol{\kappa}}} = \mathbf{0}$, we have $\bar{\boldsymbol{\beta}}^T_{\bar{\boldsymbol{\kappa}}}K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}} = O_p(n^{-1})$ and $S_{\boldsymbol{\kappa}}(\eta) = \exp\{-\frac{1}{2}\eta\boldsymbol{b}^T_{\boldsymbol{\kappa}}\Omega^{-2}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}\boldsymbol{b}_{\boldsymbol{\kappa}} + O_p(1)\}$; otherwise, $S_{\boldsymbol{\kappa}}(\eta) = \exp\{-\frac{1}{2}n\bar{\boldsymbol{\beta}}^T_{\bar{\boldsymbol{\kappa}}}K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}} + O(1)\}$.

Take any $\boldsymbol{\kappa} \in \{0,1\}^p$ and $\boldsymbol{\kappa} \neq \boldsymbol{\kappa}_1$. If $\boldsymbol{\beta}_{0\bar{\boldsymbol{\kappa}}} = \mathbf{0}$, we must have $\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}_1}) < \mathrm{tr}(\Lambda_{\boldsymbol{\kappa}})$ and hence $\frac{\tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta},\eta)}{\tilde{C}_{n,\boldsymbol{\kappa}_1}(\boldsymbol{\theta},\eta)} = O_p(n^{(\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}_1})-\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}}))/2}) = o_p(1)$. On the other hand, if $\boldsymbol{\beta}_{0\bar{\boldsymbol{\kappa}}} \neq \mathbf{0}$, we have $\frac{\tilde{C}_{n,\boldsymbol{\kappa}}(\boldsymbol{\theta},\eta)}{\tilde{C}_{n,\boldsymbol{\kappa}_1}(\boldsymbol{\theta},\eta)} = O_p(n^{(\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}_1})-\mathrm{tr}(\Lambda_{\boldsymbol{\kappa}}))/2}\exp\{-\frac{1}{2}n\bar{\boldsymbol{\beta}}^T_{\bar{\boldsymbol{\kappa}}}K_{\bar{\boldsymbol{\kappa}}\bar{\boldsymbol{\kappa}}}\bar{\boldsymbol{\beta}}_{\bar{\boldsymbol{\kappa}}}\}) = o_p(1)$. Thus, we have $\pi_{n\boldsymbol{\kappa}_1}(\boldsymbol{\theta},\eta) \to_p 1$ as $n \to \infty$ and $\pi_{n\boldsymbol{\kappa}}(\boldsymbol{\theta},\eta) \to_p 0$ for $\boldsymbol{\kappa} \neq \boldsymbol{\kappa}_1$. This implies that $P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} = 0|\boldsymbol{\theta},\eta) \to_p 1$ and hence $P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} = 0) \to_p 1$. The other conclusion can be easily proved by using the fact that $\tilde{\Omega}^{-1}_{n,\boldsymbol{\kappa}_1}(\eta) = \Omega^{-1}_{\boldsymbol{\kappa}_1\boldsymbol{\kappa}_1} + o(1)$. □

We can now give the proof of Theorem 1.

Proof of Theorem 1. According to Lemma A.4 in Yang and He (2012), under assumptions of Theorem 1, we have

$$L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) = \exp\{-\frac{n}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T V_{12}^T V_{11}^{-1} V_{12}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}) + O_p(n^{-1/2})\}.$$

Let $\Sigma^{-1} = \Omega = V_{12}^T V_{11}^{-1} V_{12}$ and $\tilde{L}(\boldsymbol{\beta}) = \exp\{-\frac{n}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\}$. Suppose that $f(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$ and $\tilde{f}(\boldsymbol{\beta})$ are the posterior distributions corresponding to the likelihood $L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$ and $\tilde{L}(\boldsymbol{\beta})$, respectively. Thus, we have

$$f(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y}) = C_n \int L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta,$$

where C_n is such that $C_n \int L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta d\boldsymbol{\beta} = 1$. Similar equation holds for $\tilde{f}(\boldsymbol{\beta})$ and $\tilde{L}(\boldsymbol{\beta})$ with constant $\tilde{C}_n$. Suppose that $\boldsymbol{\beta}$ follows the distribution $f(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$, we have

$$\begin{aligned}
P(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} \neq \mathbf{0}|\ \boldsymbol{X},\boldsymbol{Y}) &= \int_{\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\neq\mathbf{0}} f(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})d\boldsymbol{\beta} \\
&= C_n \int_{\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\neq\mathbf{0}} \left[\int L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta\right] d\boldsymbol{\beta} \\
&\leq_p C_n \int_{\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\neq\mathbf{0}} \left[\int 2\tilde{L}(\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta\right] d\boldsymbol{\beta} \\
&= \frac{2C_n}{\tilde{C}_n}\tilde{P}(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} \neq \mathbf{0}),
\end{aligned}$$

where $\tilde{P}(\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1} \neq \mathbf{0})$ represents the probability of $\boldsymbol{\beta}$ not on the plane $\boldsymbol{\beta}_{\boldsymbol{\kappa}_1}$ with $\boldsymbol{\beta}$ following the distribution $\tilde{f}(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})$ and the inequality is in probability. Since $C_n/\tilde{C}_n \to 1$ in probability and $\tilde{P}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \neq \mathbf{0}) \to_p 0$ by Lemma 1, we have $P(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \neq \mathbf{0}|\ \boldsymbol{X},\boldsymbol{Y}) = o_p(1)$ and hence $P(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} = \mathbf{0}|\ \boldsymbol{X},\boldsymbol{Y}) \to_p 1$. For any measurable set $A \subset \mathbb{R}^q$, we have

$$\begin{aligned}
& P(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \in A|\boldsymbol{X},\boldsymbol{Y}) \\
= & \int_A \left[\int f(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})d\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\right] d\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \\
= & C_n \int_A \left[\int L(\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{Y})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta d\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\right] d\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \\
= & C_n \int_A \left[\int \tilde{L}(\boldsymbol{\beta})\exp(O_p(n^{-1/2}))\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta d\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\right] d\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \\
= & C_n \int_A \left[\int \tilde{L}(\boldsymbol{\beta})(1+O_p(n^{-1/2}))\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta d\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\right] d\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \\
= & \frac{C_n}{\tilde{C}_n}\int_A \left[\int \tilde{C}_n\tilde{L}(\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\ \boldsymbol{\theta},\eta)\pi(\boldsymbol{\theta})\pi(\eta)d\boldsymbol{\theta}d\eta d\boldsymbol{\beta}_{\bar{\boldsymbol{\kappa}}_1}\right] d\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} + \frac{C_n}{\tilde{C}_n}O_p(n^{-1/2}) \\
= & \frac{C_n}{\tilde{C}_n}\tilde{P}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \in A) + o_p(1) \\
= & \tilde{P}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} \in A) + o_p(1).
\end{aligned}$$

Thus, from Lemma A1 we have

$$\begin{aligned} P(\sqrt{n}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t} | \boldsymbol{X}, \boldsymbol{Y}) &= \tilde{P}(\sqrt{n}(\boldsymbol{\beta}_{\boldsymbol{\kappa}_1} - \bar{\boldsymbol{\beta}}_{\boldsymbol{\kappa}_1}) \le \boldsymbol{t}) + o_p(1) \\ &= F_{\Omega^{-1}_{\boldsymbol{\kappa}_1 \boldsymbol{\kappa}_1}}(\boldsymbol{t}) + o_p(1). \end{aligned}$$ □

Proof of Theorem 2. By using Lemma 1 and Theorem 4 in Qin and Lawless (1994), Theorem 2 can be proved similarly as Theorem 1. □

In order to prove Theorem 3, we first need to prove the following lemma.

Lemma 2. *Under Assumptions 2 and 3, we have*

(S1) $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})\}$ *and* $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q^T(\boldsymbol{X}, Y, \boldsymbol{\zeta})\}$ *are twice continuously differentiable with respect to* $\boldsymbol{\beta}$ *and* $\boldsymbol{\mu}$*, where* $\boldsymbol{\zeta} = c(\boldsymbol{\beta}^T, \boldsymbol{\mu}^T)^T$*.*

(S2) There exist compact neighborhoods $\mathcal{C}_{\boldsymbol{\xi}}$ *around 0 and* $\mathcal{C}_{\boldsymbol{\beta}}$*,* $\mathcal{C}_{\boldsymbol{\mu}}$ *such that* $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})/(1 + \boldsymbol{\xi}^T Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}))\}$ *has continuous partial derivatives with respect to* $\boldsymbol{\xi}$*,* $\boldsymbol{\beta}$ *and* $\boldsymbol{\mu}$*, and* $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})^T / (1 + \boldsymbol{\xi}^T Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}))\}$ *is uniformly continuous with respect to* ξ*,* β *and* μ*.*

Proof. We first prove (S1). Let $Q_j(\boldsymbol{X}, Y, \boldsymbol{\zeta})$ be the jth element of $Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})$. Then, for $j = 1, 2, \ldots, p$,

$$\begin{aligned} E\{Q_j(\boldsymbol{X}, Y, \boldsymbol{\zeta})\} &= E\{\sum_{k=1}^{m} a_k (\tau_k - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_k\}}) x_j\} \\ &= E_X\{\sum_{k=1}^{m} a_k E_{Y|X} (\tau_k - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_k\}}) x_j\} \\ &= E_X\{\sum_{k=1}^{m} a_k (\tau_k - F_X(\boldsymbol{X}^T \boldsymbol{\beta} + \mu_k)) x_j\}. \end{aligned}$$

For $j = p+1, p+2, \ldots, p+m$, we have

$$\begin{aligned} E\{Q_j(\boldsymbol{X}, Y, \boldsymbol{\zeta})\} &= E\{\tau_{j-p} - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{j-p}\}}\} \\ &= E_X\{E_{Y|X}(\tau_{j-p} - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{j-p}\}})\} \\ &= E_X\{\tau_{j-p} - F_X(\boldsymbol{X}^T \boldsymbol{\beta} + \mu_{j-p})\}. \end{aligned}$$

Under Assumptions 2 and 3, we have that $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})\}$ is twice continuously differentiable with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. Similarly, for $1 \le j_1 \le j_2 \le p$,

$$\begin{aligned} & E\{Q_{j_1}(\boldsymbol{X}, Y\boldsymbol{\zeta}) Q_{j_2}(\boldsymbol{X}, Y, \boldsymbol{\zeta})\} \\ = \; & E_X\{E_{Y|X}(\sum_k a_k (\tau_k - I_{\{Y \le \boldsymbol{X}^T + \boldsymbol{\beta} + \mu_k\}}))^2 x_{j_1} x_{j_2}\} \\ = \; & E_X\{E_{Y|X}(\sum_{k_1} \sum_{k_2} a_{k_1} a_{k_2} (\tau_{k_1} - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{k_1}\}})(\tau_{k_2} - I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{k_2}\}})) x_{j_1} x_{j_2}\} \\ = \; & E_X\{E_{Y|X}(\sum_{k_1} \sum_{k_2} a_{k_1} a_{k_2} (\tau_{k_1} \tau_{k_2} - \tau_{k_1} I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{k_2}\}} \end{aligned}$$

$$
\begin{aligned}
&- \tau_{k_2} I_{\{Y \le \boldsymbol{X}^T + \boldsymbol{\beta} + \mu_{k_1}\}} + I_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_{k_1 \wedge k_2}\}}) x_{j_1} x_{j_2})\} \\
= \;& E_X\{\sum_{k_1}\sum_{k_2} a_{k_1} a_{k_2} (\tau_{k_1}\tau_{k_2} - \tau_{k_1} F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k_2}) \\
&- \tau_{k_2} F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k_1}) + F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k_1 \wedge k_2})) x_{j_1} x_{j_2}\}.
\end{aligned}
$$

For $1 \le j_1 \le p$ and $p+1 \le j_2 \le p+m$,

$$
\begin{aligned}
& E\{Q_{j_1}(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q_{j_2}(\boldsymbol{X}, Y, \boldsymbol{\zeta})\} \\
= \;& E_X\{E_{Y|X}(\sum_k a_k(\tau_k - I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_k\}}))(\tau_{j_2 - p} - I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_2 - p}\}}) x_{j_1}\} \\
= \;& E_X\{E_{Y|X}(\sum_k a_k(\tau_k \tau_{j_2-p} - \tau_k I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_2-p}\}} - \tau_{j_2-p} I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_k\}} \\
& + I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k \wedge (j_2 - p)}\}})) x_{j_1}\} \\
= \;& E_X\{\sum_k a_k(\tau_k\tau_{j_2-p} - \tau_k F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_2-p}) - \tau_{j_2-p} F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_k) \\
& + F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k \wedge (j_2-p)})) x_{j_1}\}.
\end{aligned}
$$

For $p+1 \le j_1 \le j_2 \le p+m$,

$$
\begin{aligned}
& E\{Q_{j_1}(\boldsymbol{X}, Y\boldsymbol{\zeta}) Q_{j_2}(\boldsymbol{X}, Y, \boldsymbol{\zeta})\} \\
= \;& E_X\{E_{Y|X}(\tau_{j_1-p} - I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_1-p}\}})(\tau_{j_2-p} - I_{\{Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_2-p}\}})\} \\
= \;& E_X\{\tau_{j_1-p}\tau_{j_2-p} - \tau_{j_1-p} F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_2-p}) - \tau_{j_2-p} F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{j_1-p}) \\
& + F_X(\boldsymbol{X}^T\boldsymbol{\beta} + \mu_{(j_1 \wedge j_2)-p})\}.
\end{aligned}
$$

To prove (S2), we first choose a neighborhood $\mathcal{C}_\mu$ of $\boldsymbol{\mu}_0$ sufficiently small such that $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m) \in \mathcal{C}_\mu$ implies $\mu_1 < \mu_2 < \cdots < \mu_m$ (since $\boldsymbol{\mu}_0$ has a similar property). Let $\mu_0 = -\infty$ and $\mu_{m+1} = \infty$. Then, $F_X(X^T\beta + \mu_0) = 0$, $F_X(X^T\beta + \mu_{m+1}) = 1$. Define $Q_j^{(k)} = Q_j(\boldsymbol{X}, Y, \boldsymbol{\beta}, \boldsymbol{\mu}) I_{\{\boldsymbol{X}^T\boldsymbol{\beta} + \mu_k < Y \le \boldsymbol{X}^T\boldsymbol{\beta} + \mu_{k+1}\}}$. For $j \in \{1, 2, \ldots, p\}$, let $q_j^{(k)} = \sum_{s=1}^k a_s \tau_s x_j + \sum_{s=k+1}^m a_s(\tau_s - 1) x_j$, and for $j \in \{p+1, p+2, \ldots, p+m\}$, let $q_j^{(k)} = \tau_{j-p} - I_{\{j-p>k\}}$. Then, $E_{Y|X}(Q_j^{(k)}) = F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_{k+1}) - F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_k))$. Define $q^{(k)} = (q_1^{(k)}, \ldots, q_{p+m}^{(k)})$. By the conditional expectation, for each $j \in \{1, 2, \ldots, p+m\}$, we have

$$
E \frac{Q_j(\boldsymbol{X}, Y, \boldsymbol{\zeta})}{1 + \xi^T Q(\boldsymbol{X}, Y, \boldsymbol{\beta}, \boldsymbol{\mu})} = E_X\big[\sum_{0 \le k \le m} \frac{q_j^{(k)}}{1 + \xi^T q^{(k)}} (F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_{k+1}) - F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_k))\big],
$$

and for $j_1, j_2 \in \{1, 2, \ldots, p+m\}$

$$
E \frac{Q_{j_1}(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q_{j_2}(\boldsymbol{X}, Y, \boldsymbol{\zeta})}{1 + \xi^T Q(\boldsymbol{X}, Y, \boldsymbol{\beta}, \boldsymbol{\mu})} = E_X[\sum_{0 \le k \le m} \frac{q_{j_1}^{(k)} q_{j_2}^{(k)}}{1 + \xi^T q^{(k)}} (F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_{k+1}) - F_X(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta} + \mu_k))].
$$

Hence, when ξ is sufficiently close to 0, $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})/(1 + \boldsymbol{\xi}^T Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}))\}$ has continuous partial derivative and $E\{Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q(\boldsymbol{X}, Y, \boldsymbol{\zeta})^T / (1 + \boldsymbol{\xi}^T Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}))\}$ is uniformly continuous with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\mu}$. □

Define

$$\Gamma_n(\boldsymbol{\zeta}) = -n^{-1}\sum_{i=1}^{n}\log\{1+\lambda_n(\boldsymbol{\zeta})^T Q(\boldsymbol{x}_i, y_i, \boldsymbol{\zeta})\} \tag{14}$$

where $\lambda_n(\boldsymbol{\zeta})$ satisfies

$$\sum_{i=1}^{n}\frac{Q(\boldsymbol{x}_i, y_i, \boldsymbol{\zeta})}{1+\lambda_n(\boldsymbol{\zeta})^T Q(\boldsymbol{x}_i, y_i, \boldsymbol{\zeta})} = 0. \tag{15}$$

Recall that the MELE $\hat{\boldsymbol{\zeta}}$ that maximizes the EL (12) satisfies $\hat{\boldsymbol{\zeta}} = \text{argmax}\{\Gamma_n(\boldsymbol{\zeta})\}$. By Lemma A2 and Lemma 3 in Yang and He (2012), we can prove the following consistency result. Its proof is almost exactly the same as the proof of Theorem 3.1 in Yang and He (2012) and we omit it.

Lemma 3. *Let $\hat{\boldsymbol{\zeta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\mu}}^T)^T$ be the estimator that maximizes the EL (12). The maximum EL estimator (MELE) $\hat{\boldsymbol{\zeta}}$ is a consistent estimator of $\boldsymbol{\zeta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{\mu}_0^T)^T$ under Assumptions 1′, 2, 3, 4′ and 5′.*

We also need the following lemma for proving Theorem 2.

Lemma 4. *Under Assumptions 1′, 2, 3, 4′ and 5′, we have*

$$\begin{aligned}\Gamma_n(\boldsymbol{\zeta}) = &-\frac{1}{2}(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0)^t V_{12}^t V_{11} V_{12}(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0) + n^{-1/2}(\boldsymbol{\zeta}-\boldsymbol{\zeta}_0)^t V_{12}^t V_{11}^{-1} M_n \\ &- \frac{1}{2}n^{-1} M_n^t V_{11}^{-1} M_n + o_p(n^{-1})\end{aligned} \tag{16}$$

uniformly in $\boldsymbol{\zeta}$, for $\boldsymbol{\zeta}-\boldsymbol{\zeta}_0 = O(n^{-1/2})$ and

$$\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0 = n^{-1/2}(V_{12}^t V_{11} V_{12})^{-1} V_{12}^t V_{11}^{-1} M_n + o_p(n^{-1/2}), \tag{17}$$

where $M_n = n^{-1/2}\sum_{i=1}^{n} Q(\boldsymbol{x}_i, y_i, \boldsymbol{\zeta}_0)$, $V_{11} = E(Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0)Q^T(\boldsymbol{X}, Y, \boldsymbol{\zeta}_0))$ and $V_{12} = -\frac{\partial}{\partial\boldsymbol{\zeta}}E(Q(\boldsymbol{X}, Y, \boldsymbol{\zeta}))|_{\boldsymbol{\zeta}=\boldsymbol{\zeta}_0}$.

Proof. We will use Lemma 6 in Molanes Lopez et al. (2009) to prove Lemma 4. Clearly, Conditions (C0), (C1), (C2), and (C3) are satisfied from the assumptions of Lemma 4 and the previous lemmas. We only need to check the following conditions as in Molanes Lopez et al. (2009):

(C4) $\|\sum_{i=1}^{n}[Q(X_i, Y_i, \boldsymbol{\zeta}) - E\{Q(X_i, Y_i, \boldsymbol{\zeta})\}]\| = O_p(n^{1/2})$, uniformly in $(\boldsymbol{\beta}, \boldsymbol{\mu})$ in a $o(1)$-neighborhood of $\boldsymbol{\zeta}_0$.

(C5) $\|\sum_{i=1}^{n}[Q(X_i, Y_i, \boldsymbol{\zeta})Q(X_i, Y_i, \boldsymbol{\zeta})^T - E\{Q(X_i, Y_i, \boldsymbol{\zeta})Q(X_i, Y_i, \boldsymbol{\zeta})^T\}]\| = o_p(n)$, uniformly in $\boldsymbol{\zeta}$ in a $o(1)$-neighborhood of $\boldsymbol{\zeta}_0$.

(C6) $\|\sum_{i=1}^{n}[Q(X_i, Y_i, \boldsymbol{\zeta}) - E\{Q(X, Y, \boldsymbol{\zeta})\} - Q(X_i, Y_i, \boldsymbol{\zeta}_0) + E\{Q(X, Y, \boldsymbol{\zeta}_0)\}]\| = o_p(n^{1/2})$, uniformly in $\boldsymbol{\zeta}$ in a $O_p(n^{-1/2})$ neighborhood of $\boldsymbol{\zeta}_0$.

By the definition of P-Glivenko–Cantelli (P-GC) and P-Donsker class, a sufficient condition for (C4) is that $\mathcal{F}_0 = \{Q_j(\boldsymbol{X}, Y, \boldsymbol{\zeta}) : \boldsymbol{\zeta} \in \mathbb{R}^{p+m},\ j = 1, 2, \ldots, p+m,\}$ is P-Donsker, and for (C5) is that $\mathcal{F}_1 = \{Q_{j_1}(\boldsymbol{X}, Y, \boldsymbol{\zeta}) Q_{j_2}(\boldsymbol{X}, Y, \boldsymbol{\zeta}) : \boldsymbol{\zeta} \in \mathbb{R}^{p+m},\ j_1, j_2 = 1, 2, \ldots, p+m, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\mu} \in \mathbb{R}^m\}$ is a P-GC class.

From Lemma 9.12 and Lemma 9.8 of Korosok (2008), we know that the class of indicator functions $\mathcal{G}_{1k} = \{1_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_k\}} - \tau_k : \boldsymbol{\beta} \in \mathbb{R}^p, \mu_k \in \mathbb{R}\}$ is a VC-class, and therefore it is a P-Donsker class (Theorem 9.3 and Theorem 8.19 of Korosok 2008). Similarly, we can show that $\mathcal{G}_{2kj} = \{(1_{\{Y \le \boldsymbol{X}^T \boldsymbol{\beta} + \mu_k\}} - \tau_k) X_j : \boldsymbol{\beta} \in \mathbb{R}^p, \mu_k \in \mathbb{R}\}$ is a VC-class (Lemma 9.9 of Korosok 2008) and thus is P-Donsker. By Corollary 9.31 and Theorem 9.29 of Korosok (2008), we get that $\mathcal{F}_0$ is P-Donsker. By Theorem 8.14 of Korosok (2008), $\mathcal{G}_{1k}$ and $\mathcal{G}_{2kj}$ are also P-GC, and therefore, by Corollary 9.26 of Korosok (2008), $\mathcal{F}_1$ is P-GC. Finally, (C6) can be obtained by applying lemma 4.1 of He and Shao (1996) to $Q(X, Y, \boldsymbol{\zeta})$. □

Based on these results, Theorem 3 can be proved using the same technique as in the proof of Theorem 1.

On the hyper-prior $\Gamma(a, b)$

Berger (2006) argued that it is problematic to use the vague prior $\pi(\tau^2) \propto \tau^{-2(\epsilon+1)} \times \exp(-\epsilon'/\tau^2)$ for the "higher level" variance τ^2 in normal hierarchical models, where ϵ and ϵ' are small. Here, the parameter τ^2 corresponds to the parameter σ^2 in our paper since both of them are higher level variances, i.e., the variances of the prior for the regression coefficients β_i. Simple calculation shows that the prior $\pi(\tau^2) \propto \tau^{-2(\epsilon+1)} \exp(-\epsilon'/\tau^2)$ on τ^2 is just the inverse Gamma distribution Inv-$\Gamma(\epsilon, \epsilon')$ distribution, which is equivalent to putting a Gamma prior $\Gamma(\epsilon, \epsilon')$ on τ^{-2}. The prior we used for σ^{-2} is also a Gamma distribution $\Gamma(a, b)$. Berger (2006) states that as $\epsilon \to 0$ "the posterior for τ^2 will pile up its mass near 0, so that the answer can be ridiculous if ϵ is too small". If this is the case, we should be able to see that the τ^2 parameter would have a lot of points near 0 in the MCMC chain. Therefore, we performed a simulation study with the following normal hierarchical model:

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \quad (i = 1, \ldots, n), \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2), \\ (\beta_1, \beta_2)^T &\sim \mathcal{N}(0, \tau^2 I_2), \\ \sigma^2 &\sim \text{Inv-}\Gamma(a_1, b_1), \\ \tau^2 &\sim \text{Inv-}\Gamma(\epsilon, \epsilon'), \end{aligned}$$

where I_2 is the 2×2 identity matrix (note that the priors are on σ^2 and τ^2). In this simulation, we set $(\beta_1, \beta_2) = (1, 2)$, $\sigma^2 = 1$ and generated 100 ($n = 100$) data points from the linear model. In the Gibbs sampler, we always used $a_1 = 10^{-5}, b_1 = 10^{-5}$ and $\epsilon' = 10^{-5}$. Figure 3(a,b) shows the sampling chains of τ^2 with $\epsilon = 10^{-5}$ and $\epsilon = 10^5$, respectively. On the contrary to Berger (2006), τ^2 does not pile up its mass near 0 for $\epsilon = 10^{-5}$ (ϵ small) but does so for $\epsilon = 10^5$ (ϵ large). This is, in fact, expected because

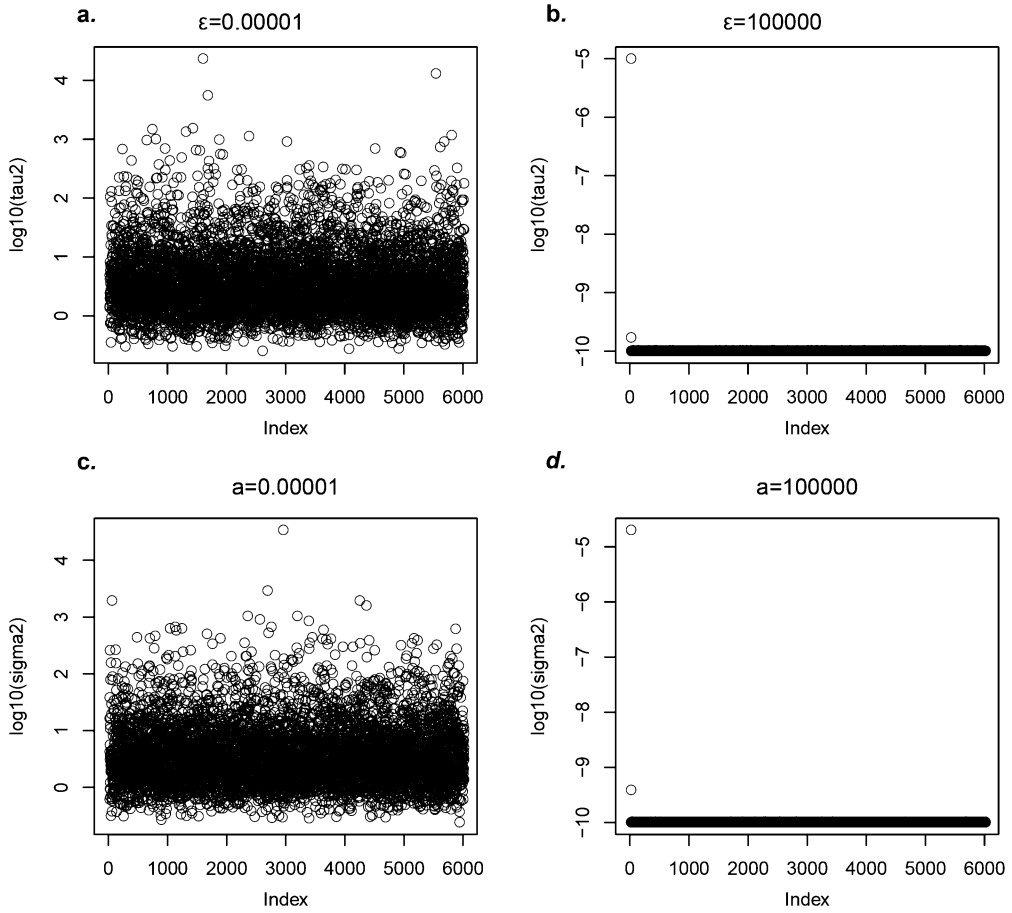

Figure 3: (a,b) The sampling chain of τ^2 for the Bayesian linear hierarchical model with $\epsilon = 10^{-5}$ and $\epsilon = 10^5$; (c,d) The sampling chain of σ^2 for the Bayesian empirical likelihood quantile regression. The y-axis is in log 10 scale.

when ϵ is small, the prior Inv-$\Gamma(\epsilon, \epsilon')$ puts most of its mass on relatively large values, but when ϵ is large it puts most of its mass near 0. For comparison, we also applied our Bayesian empirical quantile regression method to this simulation data (b is always 10^{-5}), we also got similar results (Figure 3(c,d)). The posterior of σ^2 piles up its mass near zero for $a = 10^5$ (corresponding to ϵ in the above linear normal model) but not for $a = 10^{-5}$.

In addition, the inverse gamma distribution Inv-$\Gamma(\epsilon, \epsilon')$ satisfies Condition 2 in Berger and Strawderman (1996). According to Theorem 3 in Berger and Strawderman (1996), the inverse Gamma prior gives admissible Bayesian estimators for all $\epsilon, \epsilon' > 0$ for the above normal hierarchical model. Although Theorem 3 in Berger and Strawderman (1996) cannot be applied to the linear quantile regression models considered

in this paper, it can shed light on the validity of the gamma prior on σ^{-2} (or equivalently, the inverse gamma prior on σ^2). Finally, Berger (2006) suggested that the prior $\pi(\tau^2) \propto \tau^{-(\epsilon+1)} \exp(-\epsilon'/\tau^2)$ is a good proper prior for τ^2. However, this prior is not integrable on $(0, \infty)$ when $\epsilon > 0$ is less than 1, and thus is not a proper prior. Based on the above observation, we conclude that the prior $\Gamma(a, b)$ is a reasonable prior.

Supplementary Material

Supplement to: Bayesian quantile regression based on the empirical likelihood with spike and slab priors (DOI: 10.1214/15-BA975SUPP; .zip).

References

Berger, J. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1(3): 385–402. MR2221271. 826, 850, 852

Berger, J. O. and Strawderman, W. E. (1996). "Choice of hierarchical priors: admissibility in estimation of normal means." *The Annals of Statistics*, 931–951. MR1401831. doi: http://dx.doi.org/10.1214/aos/1032526950. 851

Buchinsky, M. (1994). "Changes in the US wage structure 1963–1987: Application of quantile regression." *Econometrica: Journal of the Econometric Society*, 405–458. 821

Chen, J., Variyath, A. M., and Abraham, B. (2008). "Adjusted empirical likelihood and its properties." *Journal of Computational and Graphical Statistics*, 17(2): 426–443. MR2439967. doi: http://dx.doi.org/10.1198/106186008X321068. 829

Chen, S. X. and Hall, P. (1993). "Smoothed empirical likelihood confidence intervals for quantiles." *The Annals of Statistics*, 1166–1181. MR1241263. doi: http://dx.doi.org/10.1214/aos/1176349256. 822

Chen, S. X. and Van Keilegom, I. (2009). "A review on empirical likelihood methods for regression." *Test*, 18(3): 415–447. MR2566404. doi: http://dx.doi.org/10.1007/s11749-009-0159-5. 825

Chen, S. X. and Wong, C. (2009). "Smoothed block empirical likelihood for quantiles of weakly dependent processes." *Statistica Sinica*, 19(1): 71. MR2487878. 822

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. MR1946581. doi: http://dx.doi.org/10.1198/016214501753382273. 823

Fenske, N., Kneib, T., and Hothorn, T. (2011). "Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression." *Journal of the American Statistical Association*, 106(494). MR2847965. doi: http://dx.doi.org/10.1198/jasa.2011.ap09272. 821

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 823

Hornstein, E. and Shomron, N. (2006). "Canalization of development by microRNAs." *Nature genetics*, 38: S20–S24. 837

Hulmán, A., Witte, D. R., Kerényi, Z., Madarász, E., Tänczer, T., Bosnyák, Z., Szabó, E., Ferencz, V., Péterfalvi, A., and Tabák, A. G., et al. (2015). "Heterogeneous effect of gestational weight gain on birth weight: quantile regression analysis from a population-based screening." *Annals of Epidemiology* 25(2):133–137. 821

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *Annals of Statistics*, 730–773. MR2163158. doi: http://dx.doi.org/10.1214/009053604000001147. 823

Ishwaran, H. and Rao, J. S. (2011). "Consistency of spike and slab regression." *Statistics & Probability Letters*, 81(12): 1920–1928. MR2845909. doi: http://dx.doi.org/10.1016/j.spl.2011.08.005. 823

Kim, M.-O. and Yang, Y. (2011). "Semiparametric approach to a random effects quantile regression model." *Journal of the American Statistical Association*, 106(496). MR2896845. doi: http://dx.doi.org/10.1198/jasa.2011.tm10470. 823, 828

Koenker, R. (2005). *Quantile Regression*. New York: Cambridge University Press. MR2268657. doi: http://dx.doi.org/10.1017/CBO9780511754098. 821

Kottas, A. and Gelfand, A. E. (2001). "Bayesian semiparametric median regression modeling." *Journal of the American Statistical Association*, 96(456): 1458–1468. MR1946590. doi: http://dx.doi.org/10.1198/016214501753382363. 822

Lahiri, S. N., Mukhopadhyay, S., et al. (2012). "A penalized empirical likelihood method in high dimensions." *The Annals of Statistics*, 40(5): 2511–2540. MR3097611. doi: http://dx.doi.org/10.1214/12-AOS1040. 841

Lazar, N. A. (2003). "Bayesian empirical likelihood." *Biometrika*, 90(2): 319–326. MR1986649. doi: http://dx.doi.org/10.1093/biomet/90.2.319. 822

Lewis, B., Burge, C., and Bartel, D. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." *Cell*, 120: 15–20. 837

Li, Q., Xi, R., and Lin, N. (2010). "Bayesian regularized quantile regression." *Bayesian Analysis*, 5(3): 533–556. MR2719666. doi: http://dx.doi.org/10.1214/10-BA521. 832

Li, R., Zhong, W., and Zhu, L. (2012). "Feature screening via distance correlation learning." *Journal of the American Statistical Association*, 107(499): 1129–1139. MR3010900. doi: http://dx.doi.org/10.1080/01621459.2012.695654. 841

Li, Y. and Zhu, J. (2008). "L_1-norm quantile regression." *Journal of Computational and Graphical Statistics*, 17: 163–185. MR2424800. doi: http://dx.doi.org/10.1198/106186008X289155. 832

Lin, N. and Chang, C. (2012). "Comment on article by Lum and Gelfand." *Bayesian Analysis*, 7(2): 263–270. MR2934949. doi: http://dx.doi.org/10.1214/12-BA708B. 822

Lu, J. and Clark, A. (2012). "Impact of microRNA regulation on variation in human gene expression." *Genome Research*, 22: 1243–1254. 837, 839, 840

Machado, J. A. and Mata, J. (2005). "Counterfactual decomposition of changes in wage distributions using quantile regression." *Journal of applied Econometrics*, 20(4): 445–465. MR2143445. doi: http://dx.doi.org/10.1002/jae.788. 821

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 823

Molanes Lopez, E., Keilegom, I., and Veraverbeke, N. (2009). "Empirical likelihood for non-smooth criterion functions." *Scandinavian Journal of Statistics*, 36(3): 413–432. MR2549702. doi: http://dx.doi.org/10.1111/j.1467-9469.2009.00640.x. 829, 849

Narisetty, N. N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: http://dx.doi.org/10.1214/14-AOS1207. 823

Okada, K. and Samreth, S. (2012). "The effect of foreign aid on corruption: a quantile regression approach." *Economics Letters*, 115(2): 240–243. 821

Owen, A. (1991). "Empirical likelihood for linear models." *The Annals of Statistics*, 19(4): 1725–1747. MR1135146. doi: http://dx.doi.org/10.1214/aos/1176348368. 822, 824

Owen, A. B. (1988). "Empirical likelihood ratio confidence intervals for a single functional." *Biometrika*, 75(2): 237–249. MR0946049. doi: http://dx.doi.org/10.1093/biomet/75.2.237. 822, 824

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC. 825

Qin, J. and Lawless, J. (1994). "Empirical likelihood and general estimating equations." *The Annals of Statistics*, 300–325. MR1272085. doi: http://dx.doi.org/10.1214/aos/1176325370. 825, 847

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms." *The Annals of Applied Probability*, 7(1): 110–120. MR1428751. doi: http://dx.doi.org/10.1214/aoap/1034625254. 828

Roberts, G. O. and Rosenthal, J. S. (2001). "Optimal scaling for various Metropolis–Hastings algorithms." *Statistical Science*, 16(4): 351–367. MR1888450. doi: http://dx.doi.org/10.1214/ss/1015346320. 828

Schennach, S. M. (2005). "Bayesian exponentially tilted empirical likelihood." *Biometrika*, 92(1): 31–46. MR2158608. doi: http://dx.doi.org/10.1093/biomet/92.1.31. 823

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). "Measuring and testing dependence by correlation of distances." *The Annals of Statistics*, 35(6): 2769–2794. MR2382665. doi: http://dx.doi.org/10.1214/009053607000000505. 841

Taddy, M. A. and Kottas, A. (2010). "A Bayesian nonparametric approach to inference for quantile regression." *Journal of Business & Economic Statistics*, 28(3). 822

Tang, C. Y. and Leng, C. (2010). "Penalized high-dimensional empirical likelihood." *Biometrika*, 97(4): 905–920. MR2746160. doi: http://dx.doi.org/10.1093/biomet/asq057. 823

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58: 267–288. MR1379242. 823, 832

Tierney, L. (1994). "Markov chains for exploring posterior distributions." *the Annals of Statistics*, 1701–1728. MR1329166. doi: http://dx.doi.org/10.1214/aos/1176325750. 828

Wu, C., Shen, Y., and Tang, T. (2009). "Evolution under canalization and the dual roles of microRNAs: a hypothesis." *Genome Research*, 19: 734–743. 837

Xi, R., Li, Y., and Hu, Y. (2015). "Supplement to: Bayesian Quantile Regression Based on the Empirical Likelihood with spike and slab priors." Bayesian Analysis. doi: http://dx.doi.org/10.1214/15-BA975SUPP. 828

Yang, Y. and He, X. (2012). "Bayesian empirical likelihood for quantile regression." *The Annals of Statistics*, 40(2): 1102–1131. MR2985945. doi: http://dx.doi.org/10.1214/12-AOS1005. 823, 827, 829, 846, 849

Yu, K. and Moyeed, R. A. (2001). "Bayesian quantile regression." *Statistics & Probability Letters*, 54(4): 437–447. MR1861390. doi: http://dx.doi.org/10.1016/S0167-7152(01)00124-9. 822

Zou, H. and Hastie, T. (2005). "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 67: 301–320. MR2137327. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x. 823

Zou, H. and Yuan, M. (2008). "Regularized simultaneous model selection in multiple quantiles regression." *Computational Statistics & Data Analysis*, 52(12): 5296–5304. MR2526595. doi: http://dx.doi.org/10.1016/j.csda.2008.05.013. 842

Acknowledgments

The authors would like to thank Prof. Songxi Chen for his insightful suggestions. This work is partially supported by the National Natural Science Foundation of China (11471022) and National Key Basic Research Program of China (2015CB856000).