12-2015

Objective Bayesian analysis on the quantile regression

Shiyi Tu
Clemson University, stu@clemson.edu

OBJECTIVE BAYESIAN ANALYSIS ON THE QUANTILE REGRESSION

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Shiyi Tu
December 2015

Accepted by:
Dr. Xiaoqian Sun, Committee Chair
Dr. Derek Brown
Dr. Colin Gallagher
Dr. Yingbo Li

Abstract

The dissertation consists of two distinct but related research projects. First of all, we study the Bayesian analysis on the two-piece location-scale models, which contain several well-known sub-distributions, such as the asymmetric Laplace distribution, the ϵ-skew normal distribution, and the skewed Student-t distribution. The use of two-piece location-scale models is an attractive method to model non-symmetric data. From a practical point of view, a prior with some objective information may be more reasonable due to the lack of prior information in many applied situations. It has been shown that several common used objective priors, such as the Jeffreys prior, result in improper posterior distributions for the case of two-piece location-scale models. This motivates us to consider alternative priors. Specifically, we develop reference priors with partial information which lead to proper posterior distributions. Based on those priors, we extend our prior to a general class of priors. A sufficient and necessary condition is provided to ensure the propriety of the posterior distribution under such general priors. Our results show that the proposed Bayesian approach outperforms the frequentist method in terms of mean squared error. It is noteworthy that the proposed Bayesian method can be applied to the quantile regression due to the close relationship between the asymmetric Laplace distribution and the quantile regression.

The second project deals with the Bayesian variable selection for the maximum entropy quantile regression. Quantile regression has gained increasing popularity in many areas as it provides richer information than the regular mean regression, and variable selection plays an important role in quantile regression model building process, as it can improve the prediction accuracy by choosing an appropriate subset of regression predictors. Most existing methods in quantile regression consider quantile at some fixed value. However, if our purpose is, among all the fitted quantile regression models, to identify which one fits the data best, then the traditional quantile regression may not be appropriate. Therefore, we consider the quantile as an unknown parameter and estimate it

jointly with other regression coefficients. In particular, we consider the maximum entropy quantile regression whose error distribution is obtained by maximizing Shannon's entropy measure subject to two moment constraints. We apply the Bayesian adaptive Lasso to the model and put a flat prior on the quantile parameter due to the lack of information on it. Our proposed method not only addresses the problem about which quantile would be the most probable one among all the candidates, but also reflects the inner relationship of the data through the estimated quantile. We develop an efficient Gibbs sampler algorithm and show that the results of our proposed method are better than the ones under the Bayesian Lasso and Bayesian adaptive Lasso with fixed quantile values through both simulation studies and real data analysis.

Dedication

This dissertation is dedicated to my parents, Liping Peng and Jiangang Tu, who love me, believe in me, inspire me and have supported me every step of the way.

Acknowledgments

First and foremost, I would like to show my deepest gratitude to my advisor, Dr. Xiaoqian Sun, a respectable, responsible and resourceful scholar, who has provided me with valuable guidance on my graduate studies and research projects. I could not have completed my work without his inspiration and support.

I would specially like to thank Dr. Derek Brown, Dr. Colin Gallagher and Dr. Yingbo Li for their advice, support and serving on my advisory committee. They provided many valuable suggestions in my proposal defense for me to complete this work. In addition, I would thank to Dr. Cox and Kris for their help during my stay at the Department of Mathematical Sciences, Clemson University.

I have recently submitted one dissertation-derived paper for peek review. In addition, one dissertation-derived paper has been accepted for publication by *Computational Statistics and Data Analysis*. I am very grateful to the editor, the associate editor, and the anonymous referees of this journal for their constructive suggestions that led to the improvements of this paper.

I would also like to thank Dewei Wang and his wife Chendi Jiang, Dongmei Wang, Min Wang, Qi Zheng, and Haiming Zhou for continuously offering me help since I came to U.S.A. Especially, I am thankful to Xiaohua Bai, Yan Liu, Yanbo Xia, Honghai Xu, Shuhan Xu, Yibo Xu and his wife Xun Dong, Jing Zhao and his wife Jinghua Zhao for their support and encouragement.

Last but not least, I would like to thank my parents, my wife Jingshu for their love, understanding, support and assistance through my study.

Table of Contents

Title Page . . . . . **i**

Abstract . . . . . **ii**

Dedication . . . . . **iv**

Acknowledgments . . . . . **v**

List of Tables . . . . . **vii**

List of Figures . . . . . **viii**

1 Introduction **1**

2 Bayesian Analysis of Two-piece Location-scale Models under Reference Priors with Partial Information **4**
2.1 Introduction 4
2.2 Two-piece location-scale models and reparametrization 6
2.3 Reference priors with partial information 8
2.4 Simulation study and real data analysis 21
2.5 Concluding remarks 25
2.6 Proofs 26

3 Variable Selection in Quantile Regression **29**
3.1 Quantile regression 29
3.2 Frequentist variable selection in the quantile regression 31
3.3 Bayesian variable selection in the quantile regression 38
3.4 Discussion 57

4 Variable Selection in Bayesian Maximum Entropy Quantile Regression **59**
4.1 Introduction 59
4.2 Maximum entropy quantile regression 62
4.3 Bayesian adaptive Lasso on the maximum entropy quantile regression 64
4.4 Simulation study 69
4.5 Real data analysis 74
4.6 Concluding Remarks 81

5 Future Work **83**
5.1 Bayesian binary quantile regression 83
5.2 Bayesian logistic quantile regression 85

Bibliography **87**

List of Tables

2.1 Scenarios of the RPPI in two-piece location-scale models. 10
2.2 MSEs of each parameter when $n = 100$. 22
2.3 Bayesian estimates and MLEs with corresponding standard deviations (sd) in parenthesis when sample size is 100. 22
2.4 Acceptance rates for the Metropolis step when n=100. 24
2.5 Frequentist converge probability of 95% CI. 25

4.1 The parameter estimates for the simulated data with normally distributed errors and the corresponding standard deviations in the parenthesis. The true value of γ is 0.95, whereas we set it to be 0.5 for BALQR and BLQR. 73

List of Figures

2.1 Examples of the ALD 14
2.2 The marginal priors for γ 15
2.3 Boxplots summarizing the Bayesian estimates and MLE of μ, and the corresponding standard deviations when $\gamma = 0.2$ and sample size is 100. 23
2.4 Boxplots summarizing the Bayesian estimates and MLE of σ, and the corresponding standard deviations when $\gamma = 0.2$ and sample size is 100. 23
2.5 Boxplots summarizing the Bayesian estimates and MLE of γ, and the corresponding standard deviations of them when $\gamma = 0.2$ and sample size is 100. 24
2.6 Fitted curves and data histogram: Bayesian predictive density curve (Bayesian), and MLEs based curve (MLE). 25

4.1 Graph of $\gamma(r_2)$ when $r_1 = 10$. 64
4.2 Boxplots summarizing the MMADs and the corresponding standard deviations under the three methods for the six error distributions in Simulation 1 when γ is 0.5. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$). 71
4.3 Boxplots summarizing the MMADs and the corresponding standard deviations under the three methods for the six error distributions in Simulation 3 when γ is 0.95. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$). 72
4.4 Posterior distributions of $(\beta_1, \beta_2, \beta_3, \beta_4)$ overplotted with the histogram of simulated values in BMEQR. 74
4.5 Posterior distributions of $(\beta_5, \beta_6, \beta_7, \beta_8)$ overplotted with the histogram of simulated values in BMEQR. 75
4.6 Boxplots summarizing simulated γ and the corresponding standard deviation for BMEQR using the six error distributions in Simulation 1. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$). 76
4.7 Boxplots summarizing simulated γ and the corresponding standard deviation for BMEQR using the six error distributions in Simulation 2. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$). 77
4.8 Boxplots summarizing simulated γ and corresponding standard deviation for BMEQR using the six error distributions in Simulation 3. Overlaid are the AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$). 78
4.9 Boxplots summarizing the MSE of the three methods when $\gamma = 0.1$. 79
4.10 Boxplots summarizing the MSE of the three methods when $\gamma = 0.3$. 80
4.11 Boxplots summarizing the MSE of the three methods when $\gamma = 0.5$. 80
4.12 Boxplot summarizing the estimated γ in BMEQR. 81

Chapter 1

Introduction

The dissertation consists of two distinct but related research projects. The first project deals with the problems of point estimation and frequentist coverage probability for the parameters in the two-piece location-scale models. The second project discusses the variable selection for the maximum entropy quantile regression through Bayesian adaptive Lasso.

In Chapter 2, we consider the Bayesian approach to the point estimation for the parameters in the two-piece location-scale models. Numerous researchers have discussed those models by adopting subjective priors for the parameters. However, from a practical point of view, a prior with some objective information may be more reasonable due to the lack of prior information in many situations. The most common used objective prior is the Jeffreys prior. Recently, [47] derived the Jeffreys and independence Jeffreys priors for the two-piece location-scale models. Unfortunately, it has been shown that these Jeffreys priors result in improper posterior distributions for some sub-models, such as the inverse scale factors model. It is well known that the Bayesian inference on an improper posterior distribution is invalid. Therefore, the Jeffrey priors can not be used for the two-piece location-scale models. Another common used objective prior is the reference priors, which are very difficult to calculate for the two-piece location-scale models. As an alternative way, we consider using reference priors with partial information, which share the same idea as the reference priors, and are easier to derive. We derive several references priors with partial information for the two-piece location-scale models, and discuss the propriety of the posterior distributions for one special case of the models, the asymmetric Laplace distribution (ALD). We have also proposed similar results for the other two sub-models, the inverse scale factors model, and the ϵ-skew model. A sufficient and

necessary condition has been established to ensure the propriety of the posterior distribution under a general class of priors for the ALD. The results of the Bayesian approach are compared with the maximum likelihood estimators.

The close relationship between the ALD and the quantile regression is studied by [61]. Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the γth quantile regression is defined as any solution $\boldsymbol{\beta}(\gamma)$ that minimizing $\sum_i \rho_\gamma(y_i - \mathbf{x}_i\boldsymbol{\beta})$, where $\rho_\gamma(\cdot)$ is the check function. If we consider the ALD as the error distribution for the above linear model, then the maximum likelihood estimators (MLE) of regression coefficients $\boldsymbol{\beta}$ is the solution to the above minimization problem. Thus we can use Bayesian approach to estimate $\boldsymbol{\beta}$ and the method we discuss in Chapter 2 can be naturally extended to the quantile regression.

When the model contains many predictors, variable selection plays an important role in the model building process to obtain a better interpretation and to improve the precision of model fit. In Chapter 3, we review several variable selection methods in the quantile regression, ranging from the frequentist approaches to the Bayesian procedures. All of those methods estimate the regression coefficients at some fixed quantile value. However, if our purpose is, among all the quantile regression models, to identify which one fits the data best, then the traditional quantile regression may not be appropriate. For example, given a range of quantile, $(0.1, 0.2, \cdots, 0.9)$, we could fit 9 different regression models according to each quantile value, we are interested in which one is the most probable one to exact the most information from the data. That is, which model could reflect the inner relationship of the data and which quantile would be the most likely one. In such cases, those questions can be easily answered if we consider the quantile as an unknown parameter and estimate it from the data. Therefore, in order to extract important information from the data itself, we consider the quantile as an unknown parameter and estimate it jointly with other regression coefficients. The detail algorithm is discussed in Chapter 4.

In Chapter 4, we consider the Bayesian approach to the problem of variable selection on the maximum entropy quantile regression. Although the error distribution of the quantile regression is usually unknown, it is restricted to be 0 at a specified quantile level. We consider a special error distribution by maximizing Shannon's entropy measure subject to two moment constraints, and refer the resulting model to the maximum entropy quantile regression. The Bayesian adaptive Lasso has been employed to the variable selection on the maximum entropy quantile regression. We consider the quantile as an unknown parameter and put a uniform prior on it. Our proposed method not

only addresses the problem about which quantile would be the most probable one among all the candidates, but also reflects the inner relationship of the data through the estimated quantile. The results presented here are compared with the ones through Bayesian Lasso and Bayesian adaptive Lasso with fixed quantile value.

Some future work are discussed in Chapter 5. We consider extending the method in Chapter 4 to other types of quantile regression, such as the binary quantile regression, and the logistic quantile regression.

Chapter 2

Bayesian Analysis of Two-piece Location-scale Models under Reference Priors with Partial Information

2.1 Introduction

The use of skewed distributions is an attractive option for modeling data when symmetry is not appropriate; see, for example, [6], [49], [26], among others. As an illustration, it is widely known that the asymmetric Laplace distribution (ALD), a special case of this family, has received much attention in a wide range of disciplines, such as economics ([65]), engineering ([32]), financial analysis ([33]), medical study ([42]), and microbiology ([46]). In recent years, numerous techniques have been developed to derive new skewed distributions, mainly based on a modification of various symmetric distributions, such as adding a scale parameter to the symmetric density ([18]), multiplying the original density by a cumulative density function of a symmetric random variable ([40]).

Due to their simplicity and fitting real data quite well in practice, the two-piece location-scale models have been paid considerable attention in the literature. Besides the ALD, other two of

their special sub-models, the inverse scale factors model and the ϵ-skew model, have been discussed extensively by [18] and then by [39]. In the absence of prior knowledge, an objective prior such as the Jeffreys prior is often preferred to conduct Bayesian inference. Recently, [47] derived the Jeffreys and independence Jeffreys priors for several families of the skewed distributions. Unfortunately, it has been shown that these Jeffreys priors result in improper posterior distributions for some sub-models, such as the inverse scale factors model. Of particular note is that from [47], several discussants advocated the use of reference priors proposed by [10], which are very difficult to calculate for the two-piece location-scale models. However, reference priors with partial information (for short, RPPI) firstly proposed by [52] share the same idea as reference priors, and are easier to derive. Therefore, we are interested in deriving RPPI to see if they result in proper posterior distributions.

The use of RPPI is also quite attractive in applied situations because we usually have some partial prior information for several parameters. Thus, we just need to find a conditional prior for the remaining unknown parameters based on available information. For instance, [11] showed that for the range parameter in the spatial model, the frequentist coverage probability of the credible intervals based on the RPPI is better than the one in terms of the Jeffreys prior. [19] illustrated that for elapsed times in continuous-time Markov chains, the frequentist coverage of the credible intervals of the parameters based on the RPPI are better than the ones from other priors. [14] discussed Bayesian inference for the high energy physics problems by applying the RPPI for both single-count and multiple-count models and obtained a nice frequentist coverage probability. In this chapter, we derive RPPI for the two-piece location-scale models and show that some of them lead to proper posterior distributions. In particular, a sufficient and necessary condition for the propriety of the posterior distribution is provided under a general class of priors.

The reminder of this chapter is organized as follows. In Section 2.2, we describe the two-piece location-scale models and present several skewed distributions from different reparametrizations. In Section 2.3, we derive several RPPI for these distributions and study the propriety of the posterior distributions for the ALD in detail. In Section 2.4, the performance of our approach is illustrated through extensive simulation studies and one real data application. Finally, some concluding remarks are provided in Section 2.5, with proofs given in Section 2.6.

2.2 Two-piece location-scale models and reparametrization

The framework of the two-piece location-scale models was established by [47]. For completeness, we firstly overview such models as follows. Let $f(y \mid \mu, \sigma)$ be a symmetric and absolutely continuous density with support on $\mathbb{R}$, location parameter $\mu \in \mathbb{R}$, and scale parameter $\sigma \in \mathbb{R}^+$. The probability density function (pdf) of these models has the form:

$$h(y \mid \mu, \sigma_1, \sigma_2, \epsilon) = \frac{2\epsilon}{\sigma_1} f(y \mid \mu, \sigma_1) I_{(-\infty,\mu)}(y) + \frac{2(1-\epsilon)}{\sigma_2} f(y \mid \mu, \sigma_2) I_{[\mu,\infty)}(y),$$

where $\sigma_1 \in \mathbb{R}^+$ and $\sigma_2 \in \mathbb{R}^+$ are two separate scale parameters, and $0 < \epsilon < 1$. Note that the density $h(\cdot)$ is the finite mixtures of the densities obtained by truncating the location-scale densities $f(y \mid \mu, \sigma_1)$ and $f(y \mid \mu, \sigma_2)$ at the intervals $(-\infty, \mu]$ and $[\mu, \infty)$, respectively. Therefore, the density $h(\cdot)$ may not be continuous at $y = \mu$. In order to ensure the continuity of the density, we set ϵ to be $\sigma_1/(\sigma_1 + \sigma_2)$. Consequently, the above density can be rewritten as

$$g(y \mid \mu, \sigma_1, \sigma_2) = \frac{2}{\sigma_1 + \sigma_2} \left\{ f(y \mid \mu, \sigma_1) I_{(-\infty,\mu)}(y) + f(y \mid \mu, \sigma_2) I_{[\mu,\infty)}(y) \right\}. \tag{2.1}$$

Note that

$$\int_{-\infty}^{\mu} g(y \mid \mu, \sigma_1, \sigma_2) dy = \frac{\sigma_1}{\sigma_1 + \sigma_2},$$

which indicates $g(\cdot)$ is skewed about μ if $\sigma_1 \neq \sigma_2$ and the ratio σ_1/σ_2 controls the allocation of mass to each side of μ.

In a similar way as done by [47], we consider a one-to-one transformation between $(\mu, \sigma_1, \sigma_2)$ and (μ, σ, γ):

$$\mu \;=\; \mu, \quad \sigma_1 \;=\; \sigma b(\gamma), \quad \text{and} \quad \sigma_2 \;=\; \sigma a(\gamma), \tag{2.2}$$

where $\sigma > 0$, $\gamma \in \Gamma$ is an asymmetry parameter with the set Γ depending on the choice of $\{a(\cdot), b(\cdot)\}$, $a(\cdot)$ and $b(\cdot)$ are known and positive functions, and both are differentiable such that

$$0 < |\lambda(\gamma)| < \infty, \quad \text{with} \quad \lambda(\gamma) = \frac{d}{d\gamma} \log\left[\frac{a(\gamma)}{b(\gamma)}\right].$$

The density function in (2.1) can thus be written as

$$g(y \mid \mu, \sigma, \gamma) = \frac{2}{\sigma\big[a(\gamma) + b(\gamma)\big]} \left\{ f(y \mid \mu, \sigma b(\gamma)) I_{(-\infty,\mu)}(y) + f(y \mid \mu, \sigma a(\gamma)) I_{[\mu,\infty)}(y) \right\}. \tag{2.3}$$

The density in (2.3) was also presented by [5] as a general class of asymmetric distributions, including the entire family of univariate symmetric unimodal distributions as a special case. Several interesting properties of this density have been discussed by [5], one is that any random variable X with density function given by (2.3) can be represented as the product of two independent random variables as described below

Remark 1 *(Proposition 2, [5]) Let $f(\cdot)$ be a symmetric density and consider known and positive asymmetry functions $a(\gamma)$ and $b(\gamma)$. Then a random variable X has density function (2.3) if and only if there are two independent random variables V and U_γ with $V \sim 2f(x \mid \mu, \sigma) I\{x \geq \mu\}$ and $P(U_\gamma = a(\gamma)) = a(\gamma)/(a(\gamma) + b(\gamma))$, $P(U_\gamma = -b(\gamma)) = b(\gamma)/(a(\gamma) + b(\gamma))$ such that $X = U_\gamma V$.*

Remark 1 provides an alternative way of constructing random variable with the density function (2.3). Let $F(\cdot)$ be the cumulative density function of $f(\cdot)$. The median of the two-piece location-scale models is given by

$$Q^{-1}\left(\frac{1}{2} \mid \gamma\right) = \begin{cases} \mu + \sigma b(\gamma) F^{-1}\left(\frac{a(\gamma)+b(\gamma)}{4b(\gamma)}\right), & \text{if } \ a(\gamma) < b(\gamma) \\ \mu + \sigma a(\gamma) F^{-1}\left(\frac{3a(\gamma)-b(\gamma)}{4a(\gamma)}\right), & \text{if } \ a(\gamma) \geq b(\gamma) \end{cases},$$

where $Q(\cdot)$ is the cumulative density function of the two-piece location-scale random variable. Note that the above median is always greater than μ.

In this chapter, we mainly focus on the case in which $f(\cdot)$ belongs to the class of scale mixture of normals, which includes three common models in terms of $\{a(\gamma),\ b(\gamma)\}$: the inverse scale factors model with $\{a(\gamma) = \gamma,\ b(\gamma) = 1/\gamma\}$ ([18]), the ϵ-skew model with $\{a(\gamma) = 1-\gamma,\ b(\gamma) = 1+\gamma\}$ ([39]), and the ALD with $f(\cdot)$ being the standard Laplace distribution, and $\{a(\gamma) = 1/\gamma,\ b(\gamma) = 1/(1-\gamma)\}$ ([62]).

[39] discussed a particular case of the ϵ-skew model, the so-called ϵ-skew-normal distribution, and they considered Bayesian analysis by adopting a subjective prior for μ, with fixed σ and γ. [61] considered Bayesian quantile regression by employing a likelihood function which is based on the ALD. From a practical point of view, a prior with some objective information is more reasonable due

to the lack of prior information in various applications. These observations motivate us to consider alternative priors with objective information for all model parameters.

2.3 Reference priors with partial information

Due to the lack of prior knowledge about the unknown parameters, we often have a preference for the use of objective priors. One of the most widely used noninformative priors is the Jeffreys prior, which is proportional to the square root of the determinant of the Fisher information matrix of the model. The Jeffreys prior enjoys the invariant property under any one-to-one reparameterization of the model. For notational simplicity, we use the same notations as in [47]. Define

$$\begin{aligned}
\alpha_1 &= \int_0^\infty \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt, \\
\alpha_2 &= 2\int_0^\infty \left[1 + t\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt, \\
\alpha_3 &= \int_0^\infty t\left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt.
\end{aligned}$$

Remark 2 *(Theorems 1 and 3, [47]) Let $g(y \mid \mu, \sigma, \gamma)$ be as in (2.3). Under the following three conditions*

(i) $\int_0^\infty \left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt < \infty$,

(ii) $\int_0^\infty t^2\left[\frac{f'(t)}{f(t)}\right]^2 f(t)\,dt < \infty$,

(iii) $\lim_{t\to\infty} tf(t) = 0$ *or* $\int_0^\infty tf'(t) = -\frac{1}{2}$,

the Fisher information matrix of the model (2.3) is given by

$$I(\mu,\sigma,\gamma) = \begin{pmatrix}
\frac{2\alpha_1}{a(\gamma)b(\gamma)\sigma^2} & 0 & \frac{2\alpha_3}{\sigma\left[a(\gamma)+b(\gamma)\right]}\left[\frac{a'(\gamma)}{a(\gamma)} - \frac{b'(\gamma)}{b(\gamma)}\right] \\
0 & \frac{\alpha_2}{\sigma^2} & \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] \\
\frac{2\alpha_3}{\sigma\left[a(\gamma)+b(\gamma)\right]}\left[\frac{a'(\gamma)}{a(\gamma)} - \frac{b'(\gamma)}{b(\gamma)}\right] & \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] & \frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)} + \frac{a'(\gamma)^2}{a(\gamma)}\right] - \left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2
\end{pmatrix}.$$

If the Fisher information matrix is non-singular, then the Jeffreys prior $\pi_J(\mu,\sigma,\gamma)$ and the independence Jeffreys prior $\pi_I(\mu,\sigma,\gamma)$ are respectively given by

$$\begin{aligned}\pi_J(\mu,\sigma,\gamma) &\propto \frac{|\lambda(\gamma)|}{\sigma^2\big[a(\gamma)+b(\gamma)\big]};\\ \pi_I(\mu,\sigma,\gamma) &\propto \frac{1}{\sigma}\sqrt{\frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2}.\end{aligned}$$

As shown in [47], the posterior distribution is improper under the Jeffreys prior for any choice of $\{a(\gamma), b(\gamma)\}$ if the mapping in (2.2) is one-to-one. Under the independence Jeffreys prior, the posterior distribution is proper for the ϵ-skew model, whereas it is improper for the inverse scale factors model. Also from [47], several discussants advocated the use of reference priors proposed by [10], which are very difficult to calculate for the two-piece location-scale models. However, RPPI firstly proposed by [52] share the same idea as reference priors, and are easier to derive. Therefore, we are interested in deriving RPPI to see if they result in proper posterior distributions. It deserves mentioning that the use of RPPI for the unknown parameters is not unreasonable, because in many practical problems, we may have partial prior information for some of the parameters. For example, in a skewed asymmetric population, we might possess sensible prior information about γ even though the prior information for other parameters is unknown.

Let $\mathbf{X}=(x_1,x_2,\cdots,x_n)$ be a random sample from the density $p(x \mid \boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$ and $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}_2$. Given that the subjective prior $\pi(\boldsymbol{\theta}_1)$ for $\boldsymbol{\theta}_1$ is known, consider the expected Kullback-Leibler divergence between the conditional posterior density $\boldsymbol{\theta}_2$, given $\boldsymbol{\theta}_1$ and $\mathbf{X}$, and the conditional prior of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$,

$$\Psi(\mathbf{X},\pi(\cdot \mid \boldsymbol{\theta}_1)) = E\left[\int_{\boldsymbol{\Theta}_1}\pi(\boldsymbol{\theta}_1 \mid \mathbf{X})\int_{\boldsymbol{\Theta}_2}\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1,\mathbf{X})\log\left\{\frac{\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1,\mathbf{X})}{\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)}\right\}d\boldsymbol{\theta}_2 d\boldsymbol{\theta}_1\right]. \tag{2.4}$$

The conditional prior for $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$, $\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)$, is derived through maximizing the asymptotic expansion of (2.4), which yields the following lemma that plays an important role in deriving the RPPI.

Lemma 1 *Given a density function $p(x \mid \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, let $\Sigma_{22}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ denote the Fisher information matrix of $\boldsymbol{\theta}_2$. Assume*

$$|\Sigma_{22}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)| = g_1(\boldsymbol{\theta}_1) g_2(\boldsymbol{\theta}_2),$$

for some functions $g_1(\cdot)$ and $g_2(\cdot)$. Then the conditional reference prior of $\boldsymbol{\theta}_2$ given $\boldsymbol{\theta}_1$ satisfies

$$\pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1) \propto |\Sigma_{22}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)|^{\frac{1}{2}} \propto \left\{ g_2(\boldsymbol{\theta}_2) \right\}^{\frac{1}{2}}.$$

Note that the subjective marginal prior $\pi(\boldsymbol{\theta}_1)$ could be improper, and thus the full prior for $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \pi(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_1)$, may also be improper. Therefore, it is essential to check the propriety of the posterior distribution under the full prior $\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

There are three unknown parameters μ, σ, and γ in the two-piece location-scale models. Since μ is the location parameter, $\pi(\mu) \propto 1$, the most commonly used objective (noninformative) prior for μ, could be viewed as a subjective marginal prior for μ. Of course, some other priors such as a normal prior with known mean and variance as a subjective marginal prior for μ can also be considered. Similarly, the most commonly used noninformative prior for the scale parameter σ is $\pi(\sigma) \propto 1/\sigma$, which is also a standard objective prior in many models and could be viewed as a subjective marginal prior for σ. Furthermore, a Gamma or inverse Gamma prior for σ can also be considered. Since the range of γ depends on parametrization, the subjective marginal prior for γ may be different for the different choices of $\{a(\gamma),\ b(\gamma)\}$. Nonetheless, a uniform prior of γ within an appropriate interval could be a good choice as a subjective marginal prior on γ. We summarize these scenarios in Table 2.1:

Scenario	Known subjective prior	Unknown conditional noninformative prior
1	$\pi(\mu)$	$\pi(\sigma, \gamma \mid \mu)$
2	$\pi(\sigma)$	$\pi(\mu, \gamma \mid \sigma)$
3	$\pi(\gamma)$	$\pi(\mu, \sigma \mid \gamma)$
4	$\pi(\mu, \sigma)$	$\pi(\gamma \mid \mu, \sigma)$

Table 2.1: Scenarios of the RPPI in two-piece location-scale models.

For the two-piece location-scale models, we have the following theorem (Proofs are provided in Section 2.6):

Theorem 1 *Consider the two-piece location-scale models in (2.3).*

(a) Assume that the subjective marginal prior $\pi(\mu)$ is available. The RPPI of (μ, σ, γ) is given by

$$\begin{aligned}\pi_1(\mu,\sigma,\gamma) &= \pi(\mu)\pi(\sigma,\gamma \mid \mu)\\ &\propto \frac{\pi(\mu)}{\sigma}\sqrt{\frac{1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2};\end{aligned}$$

(b) Assume that the subjective marginal prior $\pi(\sigma)$ is available. The RPPI of (μ, σ, γ) is given by

$$\begin{aligned}&\pi_2(\mu,\sigma,\gamma) = \pi(\sigma)\pi(\mu,\gamma \mid \sigma)\\ \propto\ &\pi(\sigma)\sqrt{\frac{2\alpha_1}{a(\gamma)b(\gamma)}\left\{\frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2\right\}-\frac{4\alpha_3^2\lambda(\gamma)^2}{\left[a(\gamma)+b(\gamma)\right]^2}};\end{aligned}$$

(c) Assume that the subjective marginal prior $\pi(\gamma)$ is available. The RPPI of (μ, σ, γ) is given by

$$\pi_3(\mu,\sigma,\gamma) = \pi(\gamma)\pi(\mu,\sigma \mid \gamma) \propto \frac{\pi(\gamma)}{\sigma^2};$$

(d) Assume that the subjective marginal prior $\pi(\mu, \sigma)$ is available. The RPPI of (μ, σ, γ) is given by

$$\begin{aligned}\pi_4(\mu,\sigma,\gamma) &= \pi(\mu,\sigma)\pi(\gamma \mid \mu,\sigma)\\ &\propto \pi(\mu,\sigma)\sqrt{\frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2}.\end{aligned}$$

The above four priors are closely related to several independence Jeffreys priors based on a particular group partition of parameters discussed in [48]. In fact, $\pi_1(\mu, \sigma, \gamma)$ with $\pi(\mu) \propto 1$ is the same as the independence Jeffreys prior when the parameters are grouped as $\{\mu, (\sigma, \gamma)\}$, and it is also the same as the modified Jeffreys prior in [47] if $a(\gamma) \cdot b(\gamma)$ is constant; $\pi_2(\mu, \sigma, \gamma)$ with $\pi(\sigma) \propto 1/\sigma$ is the same as the independence Jeffreys prior when the parameters are grouped as $\{\sigma, (\mu, \gamma)\}$; $\pi_3(\mu, \sigma, \gamma)$ with $\pi(\gamma) \propto \sqrt{\frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2}$ or $\pi_4(\mu, \sigma, \gamma)$ with $\pi(\mu, \sigma) \propto 1/\sigma^2$ is the same as the independence Jeffreys prior when the parameters are grouped as $\{(\mu, \sigma), \gamma\}$. In addition, the Jeffreys prior $\pi_J(\mu, \sigma, \gamma)$ is a special case of $\pi_3(\mu, \sigma, \gamma)$ with $\pi(\gamma) \propto |\lambda(\gamma)|/(a(\gamma) + b(\gamma))$, while the independence Jeffreys prior $\pi_I(\mu, \sigma, \gamma)$ is a special case of $\pi_4(\mu, \sigma, \gamma)$

with $\pi(\mu, \sigma) \propto 1/\sigma$.

To further study the propriety of the corresponding posterior distributions under the priors in Theorem 1, we firstly provide a useful lemma, whose proof is similar to that of Theorem 6 in [47] and is thus omitted for simplicity.

Lemma 2 *Let* $\mathbf{y} = (y_1, \cdots, y_n)$ *be a random sample from the population with the pdf in (2.3), where* $f(\cdot)$ *is a scale mixture of normals, that is,* $f(\cdot)$ *could be written as*

$$f(x) = \int_0^\infty \omega \cdot \phi(\omega x) dP(\omega), \tag{2.5}$$

where $\phi(\cdot)$ *is the standard normal pdf and* $P(\omega)$ *is the cumulative distribution function of any positive random variable. Consider a prior of* (μ, σ, γ)*,*

$$\pi(\mu, \sigma, \gamma) \propto \frac{1}{\sigma^d} \pi(\mu)\pi(\gamma), \tag{2.6}$$

where $d \geq 1$*,* $\pi(\mu)$ *is any bounded prior for* μ*, and* $\pi(\gamma) > 0$ *for* $\gamma \in \Gamma$*. A necessary condition for the propriety of the joint posterior distribution of* (μ, σ, γ) *under the prior in (2.6) for* $n \geq 2$ *is*

$$\int_\Gamma \frac{a(\gamma)^{n+d-1}}{\big[a(\gamma) + b(\gamma)\big]^n} \pi(\gamma)\, d\gamma < \infty. \tag{2.7}$$

In addition, the posterior distribution of (μ, σ, γ) *is proper provided that all the observations are different and* $\pi(\gamma)$ *is proper when* $d = 1$*.*

For the three models we considered: the ALD, the ϵ-skew model, and the inverse scale factors model, we choose the subjective marginal priors that specified in Theorem 1 as follows, $\pi(\sigma) \propto 1/\sigma^d$, $\pi(\gamma) \propto 1$, and $\pi(\mu, \sigma) \propto \pi(\mu)/\sigma^d$, where $d \geq 1$ and $\pi(\mu)$ being any bounded prior for μ. Since the location parameter μ does not affect the propriety of the posterior distribution, there is no need to specify the prior for μ. We discuss the propriety of the corresponding posterior distributions under the RPPI for those three special cases.

2.3.1 Asymmetric Laplace distribution

The pdf of the ALD can be obtained from (2.3) by specifying $\{a(\gamma) = 1/\gamma, b(\gamma) = 1/(1-\gamma)\}$ and $f(\cdot)$ being the standard Laplace distribution

$$q(y \mid \mu, \sigma, \gamma) = \begin{cases} \frac{\gamma(1-\gamma)}{\sigma} \exp\left\{\frac{1-\gamma}{\sigma}(y-\mu)\right\}, & y \leq \mu, \\ \\ \frac{\gamma(1-\gamma)}{\sigma} \exp\left\{-\frac{\gamma}{\sigma}(y-\mu)\right\}, & y > \mu, \end{cases} \tag{2.8}$$

where $-\infty < \mu < \infty$, $\sigma > 0$, and γ is the skewness parameter whose value is between 0 and 1. The ALD becomes the standard Laplace distribution when γ is equal to 0.5. It's obviously to see that the mode of the ALD is μ. The mean and variance of an asymmetric Laplace random variable are given by

$$\begin{aligned} E(y) &= \mu + \frac{1-2\gamma}{\gamma(1-\gamma)}\sigma, \\ V(y) &= \left(\frac{1}{\gamma^2} + \frac{1}{(1-\gamma)^2}\right)\sigma^2, \end{aligned}$$

respectively. Note that the mean $E(y)$ is greater than the mode μ when γ is less than 0.5, which indicates that the density curve is skewed to the right. Similarly, the mean $E(y)$ is less than the mode μ when γ is greater than 0.5, so the density curve is skewed to the left. Two examples of the ALD could be seen from Figure 2.1.

It is easy to calculate that $\alpha_1 = 1/2$, $\alpha_2 = 1$, and $\alpha_3 = 1/2$. Thus the Fisher information matrix of the ALD is given by

$$I^{AL}(\mu, \sigma, \gamma) = \begin{pmatrix} \frac{\gamma(1-\gamma)}{\sigma^2} & 0 & -\frac{1}{\sigma} \\ 0 & \frac{1}{\sigma^2} & \frac{1}{\sigma}\frac{2\gamma-1}{\gamma(1-\gamma)} \\ -\frac{1}{\sigma} & \frac{1}{\sigma}\frac{2\gamma-1}{\gamma(1-\gamma)} & \frac{1}{\gamma^2} + \frac{1}{(1-\gamma)^2} \end{pmatrix},$$

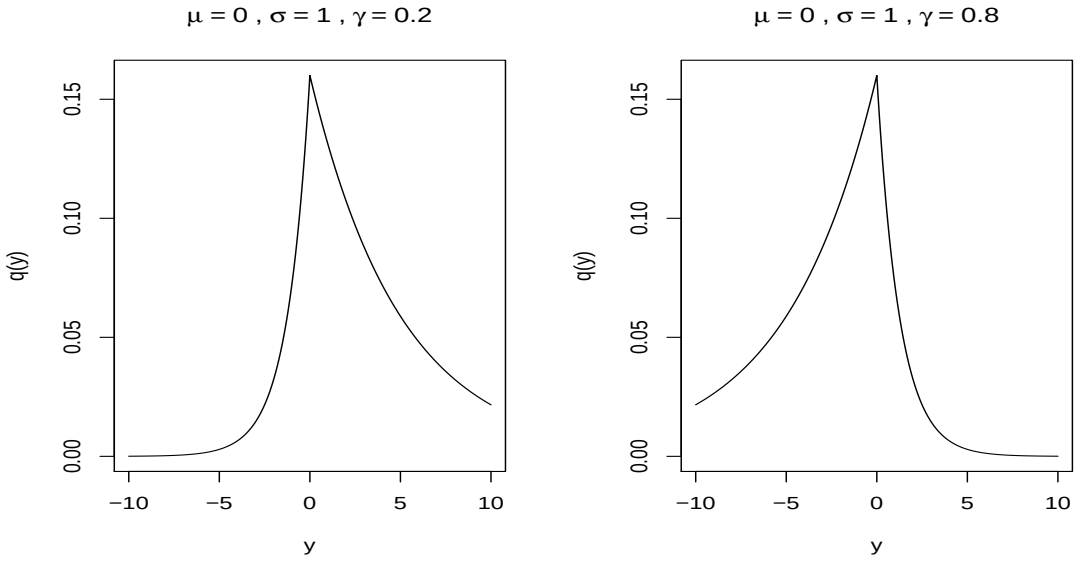

Figure 2.1: Examples of the ALD

which yields the following RPPI

$$\begin{aligned}
\pi_1^{AL}(\mu,\sigma,\gamma) &= \pi(\mu)\pi(\sigma,\gamma \mid \mu) \propto \frac{\pi(\mu)}{\sigma\sqrt{\gamma(1-\gamma)}};\\
\pi_2^{AL}(\mu,\sigma,\gamma) &= \pi(\sigma)\pi(\mu,\gamma \mid \sigma) \propto \frac{1}{\sigma^d}\sqrt{\frac{3\gamma^2-3\gamma+1}{\gamma(1-\gamma)}};\\
\pi_3^{AL}(\mu,\sigma,\gamma) &= \pi(\gamma)\pi(\mu,\sigma \mid \gamma) \propto \frac{1}{\sigma^2};\\
\pi_4^{AL}(\mu,\sigma,\gamma) &= \pi(\mu,\sigma)\pi(\gamma \mid \mu,\sigma) \propto \frac{\pi(\mu)}{\sigma^d}\sqrt{\frac{1}{\gamma^2}+\frac{1}{(1-\gamma)^2}}.
\end{aligned}$$

Corollary 1 *Consider sampling from (2.3) with $f(\cdot)$ being the standard Laplace distribution, $a(\gamma) = 1/\gamma$, and $b(\gamma) = 1/(1-\gamma)$ for $\gamma \in (0,1)$. Then the posterior distribution is proper under prior $\pi_1^{AL}(\mu,\sigma,\gamma)$ or $\pi_2^{AL}(\mu,\sigma,\gamma)$ if $d = 1$, and the posterior distribution is improper under prior $\pi_3^{AL}(\mu,\sigma,\gamma)$ or $\pi_4^{AL}(\mu,\sigma,\gamma)$.*

Consider the RPPI for the ALD, the three marginal priors for γ corresponding to $\pi_1^{AL}(\mu,\sigma,\gamma)$, $\pi_2^{AL}(\mu,\sigma,\gamma)$ and $\pi_4^{AL}(\mu,\sigma,\gamma)$ are

$$\pi_1^{AL}(\gamma) \propto \frac{1}{\sqrt{\gamma(1-\gamma)}}, \quad \pi_2^{AL}(\gamma) \propto \sqrt{\frac{3\gamma^2-3\gamma+1}{\gamma(1-\gamma)}}, \quad \text{and} \quad \pi_4^{AL}(\gamma) \propto \sqrt{\frac{1}{\gamma^2}+\frac{1}{(1-\gamma)^2}}.$$

It is not difficult to verify that both $\pi_1^{AL}(\gamma)$ and $\pi_2^{AL}(\gamma)$ are proper while $\pi_4^{AL}(\gamma)$ is improper. Figure 2.2 displays $\pi_1^{AL}(\gamma), \pi_2^{AL}(\gamma), \pi_4^{AL}(\gamma)$ along with the uniform prior on $(0,1)$ for γ, denoted by $\pi_3^{AL}(\gamma)$ for convenience.

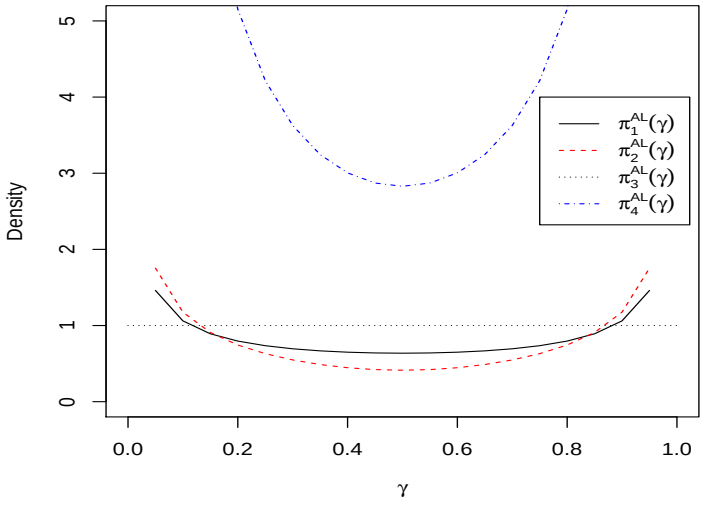

Figure 2.2: The marginal priors for γ

Corollary 1 states that among the four priors of (μ, σ, γ), both $\pi_1^{AL}(\mu, \sigma, \gamma)$ and $\pi_2^{AL}(\mu, \sigma, \gamma)$ result in a valid Bayesian inference. More generally, we consider a general class of priors given by

$$\pi(\mu, \sigma, \gamma) \propto \frac{\gamma^s (1-\gamma)^t}{\sigma^h} c(\mu, \sigma, \gamma), \tag{2.9}$$

where $h, s, t, \in \mathbb{R}$, and $c(\mu, \sigma, \gamma)$ is any positive and bounded function.

Theorem 2 *Let* $\mathbf{y} = (y_1, \cdots, y_n)$ *be a random sample from the ALD.*

(a) For the prior in (2.9), a sufficient condition for the posterior distribution to be proper is

$$h < \min\{s, t\} + 2 \quad \textit{and} \quad n > \max\{-h + 1, -\min\{s, t\} - 1\}. \tag{2.10}$$

(b) If $c(\mu, \sigma, \gamma) \propto c_1(\mu)$*, and* $c_1(\mu)$ *is any positive and bounded function. The condition (2.10) for the prior in (2.9) is also necessary.*

[47] proposed an AG-Beta prior, which has form

$$\pi_{AG}(\mu,\sigma,\gamma) \propto \frac{1}{\sigma}\frac{|a'(\gamma)b(\gamma)-a(\gamma)b'(\gamma)|}{[a(\gamma)+b(\gamma)]^{\alpha_0+\beta_0}}a^{\alpha_0-1}(\gamma)b^{\beta_0-1}(\gamma),$$

where α_0 and β_0 are positive numbers. This prior is the same as the one $\pi_1(\mu,\sigma,\gamma)$ if we choose $\alpha_0=\beta_0=1/2$ and $\pi(\mu)\propto 1$. Moreover, for the ALD, we could simplify the AG-Beta prior as

$$\pi_{AG}^{ALD}(\mu,\sigma,\gamma) \propto \frac{\gamma^{\beta_0-1}(1-\gamma)^{\alpha_0-1}}{\sigma},$$

which is a special case of our prior in (2.9). The AG-Beta prior has restrictions that $\alpha_0>0$ and $\beta_0>0$, however, the prior in (2.9) is applicable to more general cases, s, t, and h can be any real numbers as long as they satisfy the condition (2.10).

2.3.2 The ϵ-skew model

The ϵ-skew model corresponds to (2.3) by specifying $a(\gamma)=1-\gamma$ and $b(\gamma)=a+\gamma$ with $\gamma\in(-1,1)$. A special case of the ϵ-skew model, the ϵ-normal density has been extensively discussed by [39]. A crucial idea in [39] is to form a distribution by joining at $x=0$ two half normals with different scale parameters. [5] extended this idea to a more general family of distributions, the ϵ-skew model, and then discussed the moment estimation of the ϵ-skew model and the asymptotic normality of these estimates. Here we study this model from a Bayesian view.

The Fisher information matrix of the ϵ-skew model is given by

$$I^{\epsilon}(\mu,\sigma,\gamma)=\begin{pmatrix} \frac{2\alpha_1}{\sigma^2(1-\gamma^2)} & 0 & -\frac{2\alpha_3}{\sigma(1-\gamma^2)} \\ 0 & \frac{\alpha_2}{\sigma^2} & 0 \\ -\frac{2\alpha_3}{\sigma(1-\gamma^2)} & 0 & \frac{\alpha_2+1}{1-\gamma^2} \end{pmatrix}.$$

Based on the Fisher information matrix, we can obtain the RPPI and the propriety properties of

the posterior distributions according to Theorem 1 and Lemma 2

$$\begin{aligned}
\pi_1^{\epsilon}(\mu,\sigma,\gamma) &= \pi(\mu)\pi(\sigma,\gamma \mid \mu) \propto \frac{\pi(\mu)}{\sigma\sqrt{1-\gamma^2}};\\
\pi_2^{\epsilon}(\mu,\sigma,\gamma) &= \pi(\sigma)\pi(\mu,\gamma \mid \sigma) \propto \frac{1}{\sigma^d(1-\gamma^2)};\\
\pi_3^{\epsilon}(\mu,\sigma,\gamma) &= \pi(\gamma)\pi(\mu,\sigma \mid \gamma) \propto \frac{1}{\sigma^2};\\
\pi_4^{\epsilon}(\mu,\sigma,\gamma) &= \pi(\mu,\sigma)\pi(\gamma \mid \mu,\sigma) \propto \frac{\pi(\mu)}{\sigma^d\sqrt{1-\gamma^2}}.
\end{aligned}$$

Corollary 2 *Consider sampling from (2.3) with $f(\cdot)$ being a scale mixture of normals and $\{a(\gamma), b(\gamma)\}$ as in the ϵ-skew model. Then the posterior distribution is proper under prior $\pi_1^{\epsilon}(\mu,\sigma,\gamma)$ or $\pi_4^{\epsilon}(\mu,\sigma,\gamma)$ if $d = 1$, and the posterior distribution is improper under prior $\pi_2^{\epsilon}(\mu,\sigma,\gamma)$.*

Note that the propriety of the posterior distribution under prior $\pi_3^{\epsilon}(\mu,\sigma,\gamma)$ depends on the choice of $f(\cdot)$. We can see that the posterior distribution is proper provided that $f(|x|)$ is decreasing in $|x|$. For example, $f(\cdot)$ can be a pdf of normal distribution, the Laplace distribution or any other scale mixture of normals.

2.3.3 The inverse scale factors model

The inverse scale factors model corresponds to (2.3) by specifying $a(\gamma) = \gamma$ and $b(\gamma) = 1/\gamma$ with $\gamma > 0$. This model was firstly proposed by [18] through transforming an unimodal symmetric distribution. Bayesian inference for a regression analysis under the skewed-t distribution obtained from this class is also considered there. We study this model through the RPPI.

The Fisher information matrix of the inverse scale factors model is given by

$$I^{IS}(\mu,\sigma,\gamma) = \begin{pmatrix}
\frac{2\alpha_1}{\sigma^2} & 0 & \frac{4\alpha_3}{\sigma(1+\gamma^2)}\\
0 & \frac{\alpha_2}{\sigma^2} & \frac{\alpha_2(\gamma^2-1)}{\sigma(\gamma+\gamma^3)}\\
\frac{4\alpha_3}{\sigma(1+\gamma^2)} & \frac{\alpha_2(\gamma^2-1)}{\sigma(\gamma+\gamma^3)} & \frac{\alpha_2}{\gamma^2}+\frac{4}{(1+\gamma^2)^2}
\end{pmatrix},$$

which yield the following RPPI

$$\begin{aligned}
\pi_1^{IS}(\mu,\sigma,\gamma) &= \pi(\mu)\pi(\sigma,\gamma \mid \mu) \propto \frac{\pi(\mu)}{\sigma(1+\gamma^2)};\\
\pi_2^{IS}(\mu,\sigma,\gamma) &= \pi(\sigma)\pi(\mu,\gamma \mid \sigma) \propto \frac{1}{\sigma^d\gamma(\gamma^2+1)}\sqrt{\frac{(\gamma^2+1)^2}{2}+\gamma^2\left(2-\frac{4}{\pi}\right)};\\
\pi_3^{IS}(\mu,\sigma,\gamma) &= \pi(\gamma)\pi(\mu,\sigma \mid \gamma) \propto \frac{1}{\sigma^2};\\
\pi_4^{IS}(\mu,\sigma,\gamma) &= \pi(\mu,\sigma)\pi(\gamma \mid \mu,\sigma) \propto \frac{\pi(\mu)}{\sigma^d}\sqrt{\frac{\alpha_2}{\gamma^2}+\frac{4}{(\gamma^2+1)^2}}.
\end{aligned}$$

Corollary 3 *Consider sampling from (2.3) with $f(\cdot)$ being a scale mixture of normals and $\{a(\gamma), b(\gamma)\}$ as in the inverse scale factors model, then the posterior distribution is proper under prior $\pi_1^{IS}(\mu,\sigma,\gamma)$, and the posterior distribution is improper under prior $\pi_3^{IS}(\mu,\sigma,\gamma)$ or $\pi_4^{IS}(\mu,\sigma,\gamma)$.*

The propriety of the posterior distribution under prior $\pi_2^{IS}(\mu,\sigma,\gamma)$ heavily relies on the choice of $f(\cdot)$. For example, the posterior distribution is improper when $f(\cdot)$ is the standard normal density.

2.3.4 Inference on the asymmetric Laplace disrtribution

Now we focus on the posterior simulation when the ALD is considered. The RPPI are closely related to several independence Jeffreys priors, in order to distinguish them from the independence Jeffreys priors and highlight the subjective property that the RPPI possess, we advocate a different marginal prior for μ given by $\pi(\mu) \propto \exp\{-(\phi_1(\mu-\phi_2)^2)/2\}$ with $\phi_1 \geq 0$ and $\phi_2 \in \mathbb{R}$, due to the conjugacy property. In this chapter, we employ an efficient Gibbs sampler for the posterior simulation. As an illustration, we only consider the case for $\pi_1^{AL}(\mu,\sigma,\gamma)$. We observe from Theorem 1 that

$$\pi_1^{AL}(\mu,\sigma,\gamma) \propto \frac{1}{\sigma\sqrt{\gamma(1-\gamma)}}\exp\left\{-\frac{\phi_1(\mu-\phi_2)^2}{2}\right\}. \tag{2.11}$$

Note that a mixture representation of the ALD could help us develop an efficient algorithm for the posterior simulation. Let z be a random variable following the ALD(μ,σ,γ). From [32], the

pdf of z could be reexpressed as

$$z \mid \mu, \sigma, \gamma, v \sim \mathrm{N}\big(\mu + av, \ \ b\sigma v\big), \tag{2.12}$$

$$v \mid \mu, \sigma, \gamma \sim \exp\left(\frac{1}{\sigma}\right), \tag{2.13}$$

with $a = (1-2\gamma)/(\gamma(1-\gamma))$, $b = 2/(\gamma(1-\gamma))$, and $\exp(1/\sigma)$ stands for the exponential distribution with rate $1/\sigma$. If we multiply the conditional density of z and v in (2.12) and (2.13), then the integration of the joint density with respect to v gives us the density of ALD in (2.8).

The ALD has various mixture forms due to different reparameteraztions. For example, [56] used three parameters (μ^*, σ^*, p) to describe the ALD, a one-to-one transformation $\mu = \mu^*$, $\sigma = \sigma^*/2$, and $\gamma = p^*$ gives us the ALD in (2.8). The ALD in [56] can also be written as a scale mixture of normals with the scale mixing parameter following an exponential distribution. [57] discussed four-parameter ALD$(\mu^{**}, \sigma^{**}, p_1, p_2)$. If we let $\mu = \mu^{**}$, $\sigma = \sigma^{**}$, $\gamma = p_1$, and $p_1 + p_2 = 1$, we have the three-parameter ALD(μ, σ, γ) in (2.8). The four-parameter ALD$(\mu^{**}, \sigma^{**}, p_1, p_2)$ can be written as a scale mixture of uniform with the scale mixing parameter following a Gamma distribution.

Based on the mixture representation (2.12) and (2.13), the complete likelihood function of $\mathbf{y}$ becomes

$$\begin{aligned} L(\mathbf{y} \mid \mu, \sigma, \gamma, \mathbf{v}) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi b\sigma v_i}} \cdot \exp\left\{-\frac{(y_i - \mu - av_i)^2}{2b\sigma v_i}\right\} \\ &\propto \sigma^{-\frac{n}{2}} \gamma^{\frac{n}{2}} (1-\gamma)^{\frac{n}{2}} \exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \mu - av_i)^2}{2b\sigma v_i}\right\} \cdot \prod_{i=1}^{n} v_i^{-\frac{1}{2}}, \end{aligned} \tag{2.14}$$

where $\mathbf{v} = (v_1, v_2, \cdots, v_n)$, and $v_k \sim \exp(1/\sigma)$, $k = 1, 2, \cdots, n$. Note that the likelihood function in (2.14) involves the latent variable $\mathbf{v}$, which can be viewed as a usual parameter in Bayesian analysis.

Combining the complete likelihood function in (2.14) with the prior of $\mathbf{v}$ in (2.13), and the prior of (μ, σ, γ) in (2.11), the joint posterior distribution of $(\mu, \sigma, \gamma, \mathbf{v})$ becomes

$$\begin{aligned} \pi(\mu, \sigma, \gamma, \mathbf{v} \mid \mathbf{y}) &\propto L(\mathbf{y} \mid \mu, \sigma, \gamma, \mathbf{v}) \pi(\mathbf{v} \mid \mu, \sigma, \gamma) \pi_1^{AL}(\mu, \sigma, \gamma) \\ &\propto \frac{\sigma^{-\frac{3n}{2}-1}}{\gamma^{\frac{1-n}{2}}(1-\gamma)^{\frac{1-n}{2}}} \cdot \exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \mu - av_i)^2}{2b\sigma v_i} - \frac{\phi_1(\mu - \phi_2)^2}{2}\right\} \\ &\times \exp\left\{-\sum_{i=1}^{n} \frac{v_i}{\sigma}\right\} \prod_{i=1}^{n} v_i^{-\frac{1}{2}}. \end{aligned}$$

This yields the following full conditional posterior distributions

$$\begin{aligned}
v_k \mid \mu,\sigma,\gamma,\mathbf{y} &\propto \exp\left\{-\frac{(y_k-\mu-av_k)^2}{2b\sigma v_k}\right\}\cdot\exp\left\{-\frac{v_k}{\sigma}\right\}\cdot v_k^{-\frac{1}{2}}, \quad k=1,\ldots,n;\\
\mu \mid \sigma,\gamma,\mathbf{v},\mathbf{y} &\propto \exp\left\{-\sum_{i=1}^{n}\frac{(y_i-\mu-av_i)^2}{2b\sigma v_i}-\frac{\phi_1(\mu-\phi_2)^2}{2}\right\};\\
\sigma \mid \mu,\gamma,\mathbf{v},\mathbf{y} &\propto \sigma^{-\frac{3n}{2}-1}\exp\left\{-\sum_{i=1}^{n}\frac{(y_i-\mu-av_i)^2}{2b\sigma v_i}\right\}\cdot\exp\left\{-\sum_{i=1}^{n}\frac{v_i}{\sigma}\right\};\\
\gamma \mid \mu,\sigma,\mathbf{v},\mathbf{y} &\propto \gamma^{\frac{n-1}{2}}(1-\gamma)^{\frac{n-1}{2}}\cdot\exp\left\{-\sum_{i=1}^{n}\frac{(y_i-\mu-av_i)^2}{2b\sigma v_i}\right\}. \qquad (2.15)
\end{aligned}$$

Therefore, an efficient Gibbs sampler algorithm can be developed as follows.

(i) Simulate v_k from the inverse Gaussian distribution, $\mathrm{IG}(a_k,b_k)$ with

$$a_k=\sqrt{\frac{a^2+2b}{(y_k-\mu)^2}} \quad \text{and} \quad b_k=\frac{a^2+2b}{b\sigma},\ k=1,\cdots,n,$$

where the probability density function of $\mathrm{IG}(a,b)$ is given by

$$f(x\mid a,b)=\sqrt{\frac{b}{2\pi}}x^{-\frac{3}{2}}\exp\left\{-\frac{b(x-a)^2}{2a^2x}\right\},\ x>0.$$

(ii) Simulate μ from the normal distribution, $\mathrm{N}(\mu_0,\sigma_0^2)$ with

$$\mu_0=\frac{\sum_{i=1}^{n}\frac{y_i}{v_i}-na+\phi_1\phi_2 b\sigma}{\sum_{i=1}^{n}\frac{1}{v_i}+\phi_1 b\sigma} \text{ and } \sigma_0^2=\frac{b\sigma}{\sum_{i=1}^{n}\frac{1}{v_i}+\phi_1 b\sigma}.$$

(iii) Simulate σ from the inverse Gamma distribution, Inverse-Gamma$(3n/2,c_1)$ with

$$c_1=\sum_{i=1}^{n}v_i+\sum_{i=1}^{n}\frac{(y_i-\mu-av_i)^2}{2bv_i}.$$

(iv) Simulate γ from the conditional posterior distribution in (2.15).

Although the full conditional posterior distribution of γ is not of standard form, we could employ the Metropolis-Hastings (M-H) method ([13]) for the posterior simulation from (2.15). The proposed Metropolis-within-Gibbs algorithm is commonly used in many literature, such as [56]. The difference is that [56] used the estimates as the mean and variance of the proposal density in the

M-H algorithm, whereas we use a random walk proposal M-H algorithm. Our simulation studies in the next section show that the proposed sampling algorithm is quite efficient in terms of mixing and convergence.

2.4 Simulation study and real data analysis

In this section, we carry out both Monte Carlo simulations and real data analysis to compare the performance of the proposed method with that of the MLEs. In the Metropolis step, we use a normal proposal to generate the samples.

2.4.1 Simulation study

The data in simulation studies are generated from the ALD in (2.8). We generate datasets for each of sample size $n = \{50, 100, 250, 500\}$, and $\mu = 0$, $\sigma = 1$, and $\gamma = \{0.2, 0.3, 0.5, 0.7, 0.8\}$, respectively. The distribution in (2.8) with various values of γ covers the normal Laplace distribution ($\gamma = 0.5$) and some extreme situations ($\gamma = 0.2$ or 0.8). Moreover, in order to emphasize the case in which we have prior information of μ, we let $\phi_1 = 1$, and $\phi_2 = 0$ in (2.11). In other words, the subjective marginal prior of μ is the standard normal distribution.

As stated by [45], a good acceptance rate in the Metropolis step is between 0.15 and 0.5 because high acceptance rates indicate that the sampler is not moving around the parameter space reasonably well, while low acceptance rates may indicate a slow mixing of the chain. Thus we check the acceptance rate for each dataset and only use the ones with good acceptance rates. We record 1,000 samples for each sample size and each choice of γ.

Using a Markov chain Monte Carlo algorithm, a sample of size 40,000 was recorded from the posterior distribution after a burn-in period of 50,000 draws with a thinning of 10 draws. Based on the run length control diagnostic in [43], there is no evidence of lack of convergence. The Bayesian estimates were calculated by taking the average of the 1,000 repetitions. We here report the posterior mean and the posterior median, although other estimates may be used when an appropriate loss function is considered.

Table 2.2 compares the mean squared error (MSE) of Bayesian estimates and the one of the maximum likelihood estimates (MLE) for all three parameters μ, σ, and γ when $n = 100$. Note that the proposed Bayesian estimates outperform the MLEs in terms of the MSEs, especially when

γ is far away from 0.5. Similar results also were observed for $n = 50, 250$ and 500. As expected, the differences between these estimators become smaller as the sample size increases.

	Posterior mean			Posterior median			MLE		
Parameters	μ	σ	γ	μ	σ	γ	μ	σ	γ
$\gamma = 0.2$	0.1251	0.0319	0.0016	0.1273	0.0318	0.0016	0.2825	0.0452	0.0026
$\gamma = 0.3$	0.1009	0.0186	0.0021	0.1031	0.0186	0.0022	0.1681	0.0215	0.0030
$\gamma = 0.5$	0.0794	0.0107	0.0025	0.0805	0.0106	0.0026	0.1010	0.0108	0.0029
$\gamma = 0.7$	0.0936	0.0177	0.0020	0.0945	0.0176	0.0020	0.1517	0.0207	0.0028
$\gamma = 0.8$	0.1204	0.0304	0.0015	0.1213	0.0301	0.0015	0.2674	0.0418	0.0025

Table 2.2: MSEs of each parameter when $n = 100$.

Table 2.3 provides the Bayesian estimates and MLEs along with their corresponding standard deviations. We observe that the MLE of μ overestimates the location parameter when γ is less than 0.5, and underestimates μ when γ is greater than 0.5, especially when γ is close to the endpoints of the interval $(0, 1)$, whereas the Bayesian approach offers better results, not only for μ, but also for σ and γ. Moreover, the standard deviations for the Bayesian estimates are much smaller than the ones of the MLEs. Figure 2.3 - 2.5 provide a more straight forward comparison of two approaches on point estimation and the corresponding standard deviations.

Table 2.4 displays the acceptance rate for the Metropolis step when sample size is 100. Note that all rates are located in the interval [0.15, 0.5] which indicates that the proposed method is efficient. In addition, Table 2.5 gives the frequentist coverage probabilities of 95% credible intervals for the three parameters. Note that the coverage probabilities are very close to 0.95 for all the three parameters. Therefore, the Bayesian estimates have good frequentist properties. The phenomenon when sample size is not 100 is quite similar, thus, the results where the sample size is different are

	Posterior mean			Posterior median			MLE		
Parameters	μ	σ	γ	μ	σ	γ	μ	σ	γ
$\gamma = 0.2$	0.0470	1.0319	0.2086	0.0386	1.0202	0.2061	0.1068	0.9952	0.2056
	(0.3507)	(0.1759)	(0.0393)	(0.3550)	(0.1772)	(0.0399)	(0.5209)	(0.2128)	(0.0515)
$\gamma = 0.3$	0.0372	1.0035	0.3043	0.0344	0.9965	0.3027	0.0802	0.9795	0.3037
	(0.3156)	(0.1367)	(0.0467)	(0.3195)	(0.1364)	(0.0472)	(0.4023)	(0.1454)	(0.0547)
$\gamma = 0.5$	0.0057	1.0051	0.5011	0.0048	0.9985	0.5011	0.0049	0.9842	0.5011
	(0.2819)	(0.1038)	(0.0505)	(0.2839)	(0.1031)	(0.0510)	(0.3180)	(0.1030)	(0.0543)
$\gamma = 0.7$	-0.0138	1.0054	0.6952	-0.0101	0.9982	0.6968	-0.0529	0.9806	0.6961
	(0.3058)	(0.1332)	(0.0449)	(0.3075)	(0.1330)	(0.0454)	(0.3861)	(0.1429)	(0.0533)
$\gamma = 0.8$	-0.0574	1.0305	0.7903	-0.0502	1.0192	0.7927	-0.1342	1.0009	0.7916
	(0.3424)	(0.1719)	(0.0380)	(0.3448)	(0.1726)	(0.0384)	(0.4996)	(0.2046)	(0.0493)

Table 2.3: Bayesian estimates and MLEs with corresponding standard deviations (sd) in parenthesis when sample size is 100.

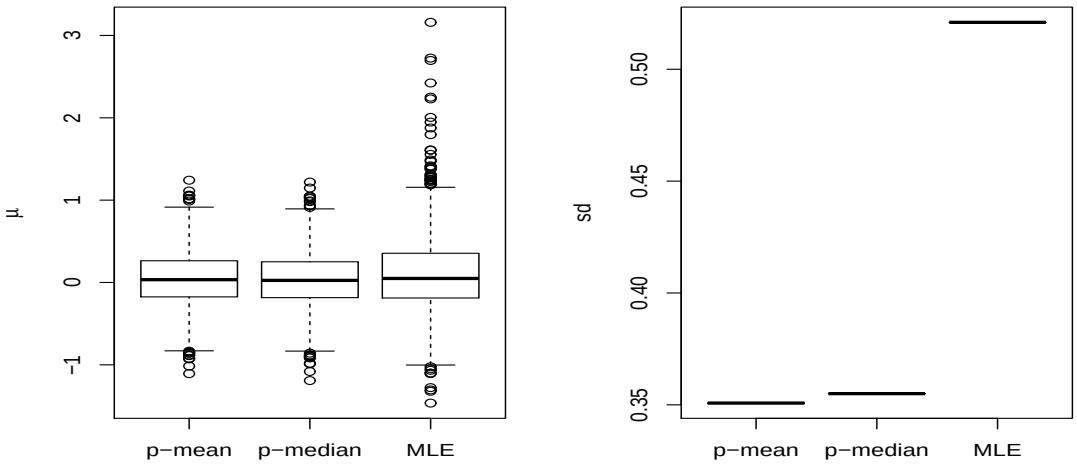

Figure 2.3: Boxplots summarizing the Bayesian estimates and MLE of μ, and the corresponding standard deviations when $\gamma = 0.2$ and sample size is 100.

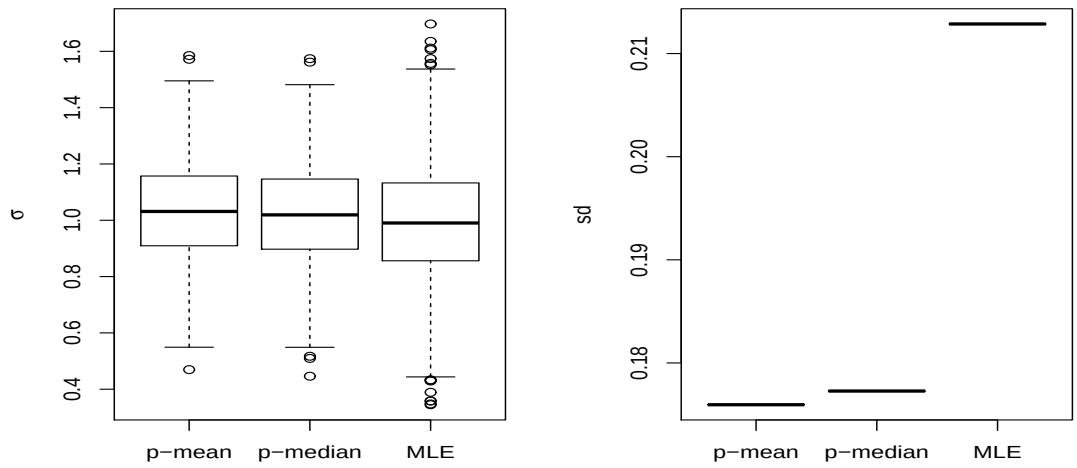

Figure 2.4: Boxplots summarizing the Bayesian estimates and MLE of σ, and the corresponding standard deviations when $\gamma = 0.2$ and sample size is 100.

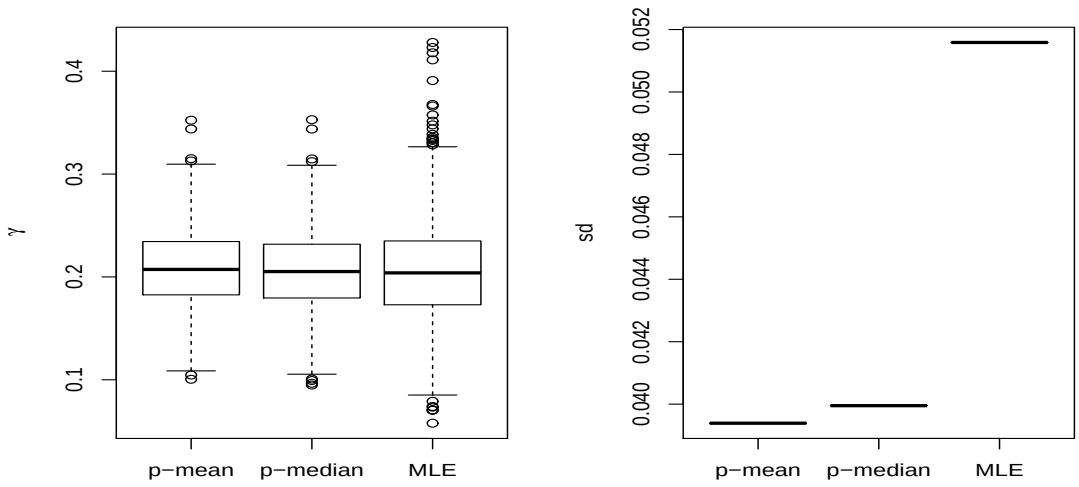

Figure 2.5: Boxplots summarizing the Bayesian estimates and MLE of γ, and the corresponding standard deviations of them when $\gamma = 0.2$ and sample size is 100.

omitted here for simplicity.

Parameters	acceptance rate
$\gamma = 0.2$	0.1672
$\gamma = 0.3$	0.2718
$\gamma = 0.5$	0.3771
$\gamma = 0.7$	0.2659
$\gamma = 0.8$	0.1734

Table 2.4: Acceptance rates for the Metropolis step when n=100.

2.4.2 Real data analysis

We consider the fibre strength data. The samples are the experimental data of the strength of $n = 63$ glass of fibre of length 1.5 cm, from the National Physical Laboratory in England. The data set was provided by [50]. We fit the data by applying the asymmetric Laplace distribution.

The MLEs of (μ, σ, γ) are given by (1.6600, 0.0943, 0.6774). As stated in Corollary 1, the posterior distribution is proper when a normal prior for μ is considered. Thus we choose the normal distribution $\mathrm{N}(1.6, 1)$ as our subjective prior for μ whose prior mean is close to the MLE of μ. According to Theorem 1, the RPPI of (μ, σ, γ) is $\pi(\mu, \sigma, \gamma) \propto 1/(\sigma\sqrt{\gamma(1-\gamma)}) \exp\{-(\mu - 1.6)^2/2\}$.

	n=100		
Parameters	μ	σ	γ
$\gamma = 0.2$	0.965	0.961	0.964
$\gamma = 0.3$	0.969	0.953	0.960
$\gamma = 0.5$	0.971	0.951	0.958
$\gamma = 0.7$	0.962	0.950	0.956
$\gamma = 0.8$	0.972	0.956	0.960

Table 2.5: Frequentist converge probability of 95% CI.

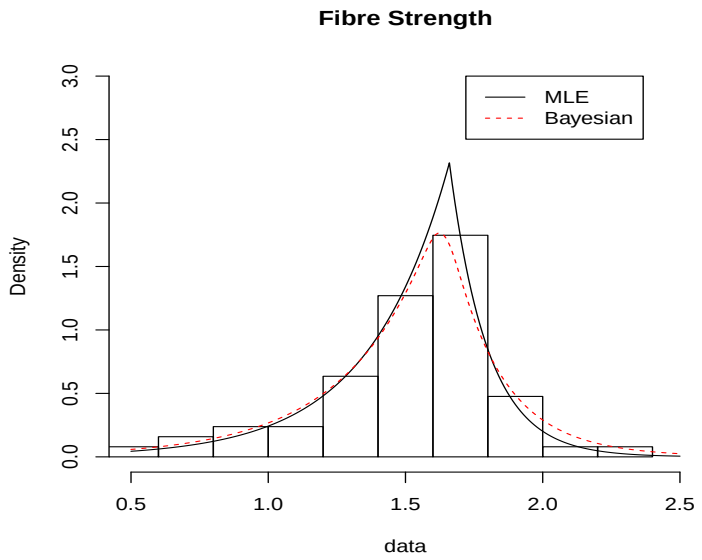

Figure 2.6: Fitted curves and data histogram: Bayesian predictive density curve (Bayesian), and MLEs based curve (MLE).

The acceptance rate for Metropolis sampling is 0.40432 showing that the result is reasonable. The improvement of fitting is demonstrated in Figure 2.6, where the histogram of the data with the 0.2 cm bin-size is overlaid with Bayesian predictive density curve and fitted MLE curve. We observe that the predictive density curve reflects the data more precisely, especially at the mode of the data.

2.5 Concluding remarks

In this chapter, we have extensively studied Bayesian analysis for a special case of the two-piece location-scale models, the ALD. We have also proposed similar results other two sub-models, the inverse scale factors model, and the ϵ-skew model. We have derived the reference priors with

partial information proposed in [52] and have shown that two of them lead to proper posterior distributions for the ALD. The RPPI are different from the independence Jeffreys prior in that they could include subjective prior information for some parameters. Specifically, we focus our discussion on the mostly widely used distribution, the ALD. Numerical results have shown that the proposed Bayesian estimates outperform the MLEs, especially when symmetry is not appropriate.

Since the seminal work of [27], quantile regression has gained increasing popularity in different areas of study as a comprehensive approach to the statistical analysis of linear response models. Due to the close relationship between the ALD and quantile regression discussed in [61], researchers conduct Bayesian analysis by imposing the ALD on the error terms in the classical linear regression model. Therefore, we plan to extend the proposed Bayesian approach to deal with the quantile regression. This work is currently under investigation and will be reported elsewhere.

2.6 Proofs

Proof of Theorem 1 : We provide the proof for part (a) only because the proofs for other cases are quite similar. Let $I_{22}(\sigma,\gamma)$ denote the corresponding Fisher information submatrix of (σ,γ). It can be seen from Remark 2 that

$$I_{22}(\sigma,\gamma)=\begin{pmatrix} \frac{\alpha_2}{\sigma^2} & \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] \\ \frac{\alpha_2}{\sigma}\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right] & \frac{\alpha_2+1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2 \end{pmatrix}.$$

Through Lemma 1, the conditional reference prior for (σ,γ) given μ is given by

$$\begin{aligned}\pi(\sigma,\gamma\mid\mu) &\propto |I_{22}(\sigma,\gamma)| \\ &\propto \frac{1}{\sigma}\sqrt{\frac{1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2}.\end{aligned}$$

This yields the RPPI of (μ,σ,γ)

$$\pi(\mu,\sigma,\gamma)\propto\frac{\pi(\mu)}{\sigma}\sqrt{\frac{1}{a(\gamma)+b(\gamma)}\left[\frac{b'(\gamma)^2}{b(\gamma)}+\frac{a'(\gamma)^2}{a(\gamma)}\right]-\left[\frac{a'(\gamma)+b'(\gamma)}{a(\gamma)+b(\gamma)}\right]^2}.$$

□

Proof of Theorem 2: Given data $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, and the prior $\pi(\mu, \sigma, \gamma)$ in (2.9), the posterior distribution of (μ, σ, γ) becomes

$$\pi(\mu, \sigma, \gamma \mid \mathbf{y}) \propto c(\mu, \sigma, \gamma) \frac{\gamma^s (1-\gamma)^t}{\sigma^h} \frac{\gamma^n (1-\gamma)^n}{\sigma^n} \prod_{i=1}^{n} \exp\left\{ -\frac{y_i - \mu}{\sigma} (\gamma - I(y_i \le \mu)) \right\}.$$

Note that the above posterior distribution is proper if and only if

$$\int_0^1 \int_{-\infty}^{\infty} \int_0^{\infty} c(\mu, \sigma, \gamma) \frac{\gamma^{n+s} (1-\gamma)^{n+t}}{\sigma^{n+h}} \prod_{i=1}^{n} \exp\left\{ -\frac{y_i - \mu}{\sigma} (\gamma - I(y_i \le \mu)) \right\} d\sigma d\mu d\gamma < \infty.$$

(a) Since $c(\mu, \sigma, \gamma)$ is a positive and bounded function, there exists a positive number B such that $c(\mu, \sigma, \gamma) < B$. Therefore, we obtain an upper bound for the left side of the above inequality

$$B \cdot \int_0^1 \int_{-\infty}^{\infty} \int_0^{\infty} \frac{\gamma^{n+s} (1-\gamma)^{n+t}}{\sigma^{n+h}} \prod_{i=1}^{n} \exp\left\{ -\frac{y_i - \mu}{\sigma} (\gamma - I(y_i \le \mu)) \right\} d\sigma d\mu d\gamma.$$

After integrating out σ, the above formula is finite if and only if

$$\int_0^1 \gamma^{n+s} (1-\gamma)^{n+t} \int_{-\infty}^{\infty} \left[\frac{1}{\sum (y_i - \mu)(\gamma - I(y_i \le \mu))} \right]^{n+h-1} d\mu d\gamma < \infty.$$

Assume $(y_{(1)}, y_{(2)}, \cdots, y_{(n)})$ is the order statistics of $\mathbf{y}$. For each interval $(y_{(i)}, y_{(i+1)})$, we have

$$\begin{aligned} & \int_{y_{(i)}}^{y_{(i+1)}} \left[\frac{1}{\sum_{k=1}^{n} (y_k - \mu)(\gamma - I(y_k \le \mu))} \right]^{n+h-1} d\mu \\ \le & \int_{y_{(i)}}^{y_{(i+1)}} \left[\frac{1}{\sum_{k=i+1}^{n} (y_{(k)} - \mu)\gamma} \right]^{n+h-1} d\mu \\ = & \frac{1}{\gamma^{n+h-1}} A_i, \end{aligned}$$

where A_i is a finite constant that depends on data only, for $i = 1, 2, \cdots, n-1$. Similar results

hold for the intervals $(-\infty, y_{(1)})$ and $(y_{(n)}, \infty)$,

$$\int_{-\infty}^{y_{(1)}} \left[\frac{1}{\sum_{k=1}^{n}(y_k - \mu)(\gamma - I(y_k \le \mu))} \right]^{n+h-1} d\mu = \frac{1}{\gamma^{n+h-1}} A_0,$$

$$\int_{y_{(n)}}^{\infty} \left[\frac{1}{\sum_{k=1}^{n}(y_k - \mu)(\gamma - I(y_k \le \mu))} \right]^{n+h-1} d\mu = \frac{1}{(1-\gamma)^{n+h-1}} A_n,$$

where $A_0 = \frac{[\sum_{k=1}^{n} y_k - n y_{(1)}]^{-(n+h)+2}}{(-(n+h)+2)(-n)}$ and $A_n = \frac{[n y_{(n)} - \sum_{k=1}^{n} y_k]^{-(n+h)+2}}{(-(n+h)+2)(-n)}$.

Combining the above results, we have

$$\begin{aligned} & \int_0^1 \int_{-\infty}^{\infty} \int_0^{\infty} \pi(\mu, \sigma, \gamma \mid \mathbf{y}) d\sigma d\mu d\gamma \\ < \; & B \int_0^1 \gamma^{n+s} (1-\gamma)^{n+t} \left[\frac{1}{\gamma^{n+h-1}} \sum_{i=0}^{n-1} A_i + \frac{1}{(1-\gamma)^{n+h-1}} A_n \right] d\gamma \\ = \; & \sum_{i=0}^{n-1} A_i B \int_0^1 \gamma^{s-h+1} (1-\gamma)^{n+t} d\gamma + A_n B \int_0^1 \gamma^{n+s} (1-\gamma)^{t-h+1} d\gamma. \end{aligned}$$

Note that the first term is finite if and only if $s - h + 1 > -1$ and $t + n > -1$, and the second term is finite if and only if $t - h + 1 > -1$ and $s + n > -1$. Therefore, we obtain a sufficient condition for the posterior distribution to be proper,

$$h < \min\{s, t\} + 2 \text{ and } \; n > \max\{-h + 1, -\min\{s, t\} - 1\}.$$

(b) If $c(\mu, \sigma, \gamma) \propto c_1(\mu)$, then the prior for (μ, σ, γ) becomes

$$\pi(\mu, \sigma, \gamma) \propto \frac{\gamma^s (1-\gamma)^t}{\sigma^h} c_1(\mu).$$

The necessary part is just the direct application of Lemma 2. Therefore, the condition (2.10) is necessary and sufficient for the posterior distribution to be proper. □

Chapter 3

Variable Selection in Quantile Regression

3.1 Quantile regression

Consider a simple decision theoretic problem: a point estimate is required for a random variable X with distribution function $F(\cdot)$. Given the loss function is

$$\rho_\gamma(\mu) = \mu(\gamma - I(\mu < 0)), \tag{3.1}$$

where γ is a number between 0 and 1, we want to find the $\hat{x}$ that minimizes the expected loss which is given by

$$\mathrm{E}[\rho_\gamma(X - \hat{x})] = (\gamma - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \gamma \int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{3.2}$$

Differentiate both sides with respect to x, we have

$$0 = (1 - \gamma)\int_{-\infty}^{\hat{x}} dF(x) - \gamma \int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \gamma.$$

Thus any element of $\{x : F(x) = \gamma\}$ minimizes the expected loss. If the solution is unique, then $\hat{x} = F^{-1}(\gamma)$, otherwise, we have an interval of x that satisfies the above equation, we may choose

the smallest one. Thus the γth quantile of X is defined by ([27])

$$F^{-1}(\gamma) = \inf\{x : F(x) \geq \gamma\}.$$

If $F(\cdot)$ is replaced by the empirical distribution function

$$F_n(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n},$$

where $(x_1, x_2, \cdots, x_n)$ is a random sample. We still choose $\hat{x}$ to minimize the expect loss (3.2)

$$\int \rho_\gamma(x - \hat{x}) dF_n(x) = \frac{\sum_{i=1}^{n} \rho_\gamma(x_i - \hat{x})}{n},$$

the resulting estimate is the γth sample quantile. Therefore, the problem of finding γth sample quantile becomes a minimization problem

$$\min_x \sum_{i=1}^{n} \rho_\gamma(x_i - x).$$

[27] employed this idea and proposed the quantile regression. Consider the linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{3.3}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the response vector, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, for $i = 1, 2, \cdots, n$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent error terms whose distribution is unknown, but is restricted to have the γth quantile equal to zero. Thus the γth regression quantile is defined as any solution $\boldsymbol{\beta}(\gamma)$ to the quantile minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{3.4}$$

Compared with the linear mean regression, quantile regression has two advantages. First, it provides richer information in the effects of the predictors on the different quantiles of the response variable than the one under the regular mean regression. Second, it is very insensitive to heteroscedasticity and outliers, thus quantile regression can accommodate non-normal errors, which

are commonly encountered in many practical applications. These two appealing features of quantile regression result in its broader application in a wide range of disciplines, see [23], [28], [55].

Variable selection plays an important role in the model building process to obtain a better interpretation and to improve the precision of model fit. The problem of variable selection is equivalent to identifying an appropriate subset of important variables via the regression coefficients. Over the years, numerous procedures have been developed for the variable selection in the quantile regression models. In Section 3.2, we review some frequentist variable selection methods in the quantile regression. Bayesian variable selection methods in the quantile regression are reviewed in Section 3.3. In Section 3.4, we discuss a drawback in the existing methods and the way how we handle it.

3.2 Frequentist variable selection in the quantile regression

Frequentists usually adopt regularization methods for the variable selection in the quantile regression by automatically setting several coefficient estimates to zeros, such as the Lasso ([37]), SCAD ([58]), adaptive Lasso([58]), to name just a few.

3.2.1 The Lasso

Consider the usual linear regression situation, the ordinary least squares (OLS) estimates are obtained by minimizing the residual squared error. The OLS estimates have two drawbacks. First, it often have low bias but large variance which would decrease prediction accuracy. Second, with a large number of predictors, we desire to choose a small subset which provides the strongest effects rather than estimating the whole model. Two techniques, ridge regression and subset selection, have been proposed to improve the OLS estimates. However, both two techniques have some drawbacks, too. The ridge regression increases the stability of the model by shrinking the coefficients, but it does not set any coefficients to 0 and thus could not provide an easy interpretable model. Subset selection provides interpretable models, but it is sensitive to the data since it is a discrete process, thus small changes in the data can result in quite different models and the prediction accuracy will be reduced.

Motivated by those facts, [53] proposed the Lasso method which not only shrinks the coefficients and set others to 0, but also remains the good properties of ridge regression and subset

selection. Consider the linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{3.5}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the response vector, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, for $i = 1, 2, \cdots, n$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent identically distributed normal errors with mean 0 and unknown variance σ^2. The Lasso estimates are defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t, \tag{3.6}$$

for some non-negative number t, and (3.6) is equivalent to the minimization problem

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{3.7}$$

where $\lambda \geq 0$.

We consider the check function (3.1) as our loss function in the quantile regression model (3.3), the Lasso estimates of the quantile regression are given by ([37])

$$\min_{\boldsymbol{\beta}} \quad \left\{ \sum_{i=1}^{n} \rho_\gamma (y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta}) \right\} \tag{3.8}$$

$$\text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s, \tag{3.9}$$

where s is the regularization parameter. We refer to it as L_1-norm quantile regression (L_1-norm QR). The above minimization problem could be rewritten as

$$\begin{aligned} \min_{\beta_0, \boldsymbol{\beta}} \quad & \gamma \sum_{i=1}^{n} \xi_i + (1-\gamma) \sum_{i=1}^{n} \zeta_i, \\ \text{subject to} \quad & \sum_{j=1}^{p} |\beta_j| \leq s, \\ & -\zeta_i \leq y_i - f(\mathbf{x}_i) \leq \xi_i, \\ & \zeta_i, \xi_i \geq 0, \quad i = 1, \ldots, n, \end{aligned}$$

where $f(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$. [37] derived the Lagrangian primal function to compute the solution path $\{\boldsymbol{\beta}(s), 0 \ \leq s \leq \infty\}$

$$\begin{aligned} L_p : \quad & \gamma \sum_{i=1}^{n} \xi_i + (1-\gamma) \sum_{i=1}^{n} \zeta_i + \lambda^\star \left(\sum_{j=1}^{p} |\beta_j| - s \right) + \sum_{i=1}^{n} \alpha_i (y_i - f(\mathbf{x}_i) - \xi_i) \\ & - \sum_{i=1}^{n} \delta_i (y_i - f(\mathbf{x}_i) + \zeta_i) - \sum_{i=1}^{n} \kappa_i \xi_i - \sum_{i=1}^{n} \eta_i \zeta_i, \end{aligned}$$

where $\lambda^\star, \alpha_i, \delta_i, \kappa_i$ and η_i are non-negative Lagrangian multipliers. Let $\theta_i = \alpha_i - \delta_i$, and we define the elbow set as

$$\boldsymbol{\Phi} = \{i : y_i - f(\mathbf{x}_i) = 0, -(1-\gamma) \leq \theta_i \leq \gamma\}.$$

As s increases, an event is defined to be either a residual $y_i - f(\mathbf{x}_i)$ changes from nonzero to zero or a coefficient β_j changes from nonzero to zero. Thus points in $\boldsymbol{\Phi}$ stay in the elbow set unless an event happens. Therefore, nonzero β_j's satisfy:

$$y_i - (\beta_0 + \sum_{j \in \boldsymbol{\Psi}} \beta_j x_{ij}) = 0 \quad \text{for} \quad i \in \boldsymbol{\Phi},$$

where $\boldsymbol{\Psi} = \{j : \beta_j \neq 0\}$. The idea of the algorithm proposed by [37] is: we start with $s = 0$ and increase it, and keep track of the location of all data points relative to the elbow set and also of the magnitude of the fitted coefficients along the way. As s increases, if a point passes through $\boldsymbol{\Phi}$, the corresponding θ_i must change from γ to $-(1-\gamma)$ or vice versa, thus points in $\boldsymbol{\Phi}$ must linger in the elbow set. Since all points in the elbow set satisfy $y_i - f(\mathbf{x}_i) = 0$, a path for $\boldsymbol{\beta}$ can be established.

The algorithm focuses on the set of points $\boldsymbol{\Phi}$ and the set of nonzero coefficients $\boldsymbol{\Psi}$. Let $\beta_0^l, \boldsymbol{\beta}^l$ and s^l be the parameter values, $\boldsymbol{\Psi}^l$ be the set of nonzero coefficients, $f^l(\cdot)$ be the function immediately after the lth event. As shown by [37],

$$\beta_0 = \beta_0^l + (s - s^l) v_0, \tag{3.10}$$

$$\beta_j = \beta_j^l + (s - s^l) v_j, \quad \forall j \in \boldsymbol{\Psi}^l, \tag{3.11}$$

$$f(\mathbf{x}) = (s - s^l) \left(v_0 + \sum_{j \in \boldsymbol{\Psi}^l} v_j x_j \right) + f^l(\mathbf{x}), \tag{3.12}$$

where $v_0 = (\beta_0 - \beta_0^l)/(s - s^l)$ and $v_j = (\beta_j - \beta_j^l)/(s - s^l)$. Note that from equations (3.10) and (3.11), β_0 and β_j proceed linearly in s for $s^l < s < s^{l+1}$, and equations (3.11) and (3.12) provide us a way to compute s^{l+1}. We compute the rate of change of the loss function to update sets $\mathbf{\Phi}$ and $\mathbf{\Psi}$ when the lth event occurs:

$$\begin{aligned}\frac{\triangle \text{loss}}{\triangle s} &= \frac{\sum_{i=1}^n \rho_\gamma(y_i - f(\mathbf{x}_i)) - \sum_{i=1}^n \rho_\gamma(y_i - f^l(\mathbf{x}_i))}{s - s^l} \\ &= (1-\gamma)\sum_{i\in\mathcal{L}}\left(v_0 + \sum_{j\in\mathbf{\Psi}} v_j x_{ij}\right) - \gamma\sum_{i\in\mathcal{R}}\left(v_0 + \sum_{j\in\mathbf{\Psi}} v_j x_{ij}\right),\end{aligned}$$

where $\mathcal{L} = \{i : y_i - f(\mathbf{x}_i) < 0, \theta_i = -(1-\gamma)\}$ and $\mathcal{R} = \{i : y_i - f(\mathbf{x}_i) > 0, \theta_i = \gamma\}$. By the definition of an event, there will be $|\mathbf{\Psi}|$ variables with nonzero coefficients and $|\mathbf{\Psi}| + 1$ points in the elbow set, we need to either remove a point in $\mathbf{\Phi}$ from $\mathbf{\Phi}$, or add a variable not in $\mathbf{\Psi}$ into $\mathbf{\Psi}$. We choose the update that corresponds to the smallest $\triangle\text{loss}/\triangle s$, and terminate the algorithm when all $\triangle\text{loss}/\triangle s$ are non-negative.

3.2.2 Smoothly clipped absolute deviation

In the framework of regularization, many different types of penalties have been introduced to achieve variable selection. A good penalty should have three desirable properties ([17]), i) the penalty functions have to singular at the origin to produce sparse solutions; ii) they have to be bounded by a constant to produce nearly unbiased estimates for large coefficients; and iii) they have to ensure the stability of model selection. The Lasso method we discussed in Section 3.2.1 uses the L_1 penalty, it creates large bias for coefficients ([17]) and thus L_1 penalty may not be a good penalty. As an alternative way, [17] proposed an approach by choosing new penalty functions who are symmetric and convex on $(0, \infty)$, the smoothly clipped absolute deviation (SCAD) function is one of those penalty functions.

Later on, [58] studied the penalized quantile regression with the SCAD penalty. The SCAD penalty is defined in terms of its first derivative and is symmetric around the origin. For $\theta > 0$, its first derivative is

$$p_\lambda'(\theta) = \lambda\left\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\right\}, \tag{3.13}$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. Consider the model (3.3) and the loss function (3.1), the SCAD penalized quantile regression solves the minimization problem ([58])

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|). \tag{3.14}$$

Note that the first derivative of the SCAD penalty function (3.13) can be viewed as a sum of two functions: one constant term and the other with a decreasing function on the range $(0, \infty)$. Therefore, the SCAD penalty function can be decomposed as the difference of two convex functions

$$p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta), \tag{3.15}$$

where both $p_{\lambda,1}(\cdot)$ and $p_{\lambda,2}(\cdot)$ are convex functions with derivatives given by

$$\begin{aligned} p'_{\lambda,1}(\theta) &= \lambda, \\ p'_{\lambda,2}(\theta) &= \lambda \left(1 - \frac{(a\lambda - \theta)_+}{(a-1)\lambda}\right) I(\theta > \lambda), \quad \text{for} \quad \theta > 0. \end{aligned}$$

Based on this decomposition, the Difference Convex Algorithm (DCA) ([3]) can be used to solve the minimization problem (3.14). The DCA minimizes a non-convex objective function by solving a sequence of convex minimization problems. At each iteration, it approximates the second convex function by a linear function. Then the objective function at each step is convex and easier to optimize than the original one. For the quantile regression, the resulting optimization at each iteration is a linear programming which indicates that the algorithm is very efficient.

According to the decomposition (3.15), the objective function in (3.14) can be decomposed as $Q_{vex}(\boldsymbol{\beta}) + Q_{cav}(\boldsymbol{\beta})$, where

$$\begin{aligned} Q_{vex}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + n \sum_{j=1}^{p} p_{\lambda_n,1}(|\beta_j|), \\ Q_{cav}(\boldsymbol{\beta}) &= -n \sum_{j=1}^{p} p_{\lambda_n,2}(|\beta_j|). \end{aligned}$$

Let $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \beta_2^{(t)}, \cdots, \beta_p^{(t)})^T$ be the solution at step t. As proposed by [58], the algorithm that minimizes $Q_{vex}(\boldsymbol{\beta}) + Q_{cav}(\boldsymbol{\beta})$ is as follows

1. Initialize $\boldsymbol{\beta}^{(0)}$.

2. Repeat $\boldsymbol{\beta}^{(t+1)} = \text{argmin}_{\boldsymbol{\beta}} \left(Q_{vex}(\boldsymbol{\beta}) + < Q'_{cav}(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)} > \right)$ until convergence.

In the $(t+1)$th iteration, the DCA approximates the objective function in the algorithm by a linear function and solves the minimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_{\lambda_n,1}(|\beta_j|) \\ - \quad & n\sum_{j=1}^{p} p'_{\lambda_n,2}(|\beta_j^{(t)}|)\text{sign}(\beta_j^{(t)})(\beta_j - \beta_j^{(t)}). \end{aligned}$$

The above minimization problem could be rewritten as the following linear programming problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}(\gamma\xi_i + (1-\gamma)\zeta_i) + n\lambda_n\sum_{j=1}^{p} v_j - n\sum_{j=1}^{p} p'_{\lambda_n,2}(|\beta_j^{(t)}|)\text{sign}(\beta_j^{(t)})(\beta_j - \beta_j^{(t)}) \\ \text{subject to} \quad & \xi_i \geq 0,\ \zeta_i \geq 0,\ \xi_i - \zeta_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}, \quad i = 1, \cdots, n, \\ & v_j \geq |\beta_j|, \quad j = 1, \cdots, p, \end{aligned}$$

which can be easily solved by many softwares.

In addition, the oracle properties of the SCAD have been proved by [58]. Rewrite the model (3.3) as

$$y_i = \mathbf{x}_{i1}^T\boldsymbol{\beta}_1 + \mathbf{x}_{i2}^T\boldsymbol{\beta}_2 + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.16}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, $\mathbf{x}_{i1} \in \mathbb{R}^s$, $\mathbf{x}_{i2} \in \mathbb{R}^{p-s}$, and the errors $\{\varepsilon_i\}$ are independent and identically distributed with γth quantile zero and a continuous, positive density in a neighborhood of zero. Assume the true regression coefficients are $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ with each component being nonzero, $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20} = 0$, and there exists a positive definite matrix Σ such that $\lim_{n\to\infty}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T)/n = \Sigma$.

Remark 3 *(Theorems 1 and 2, [58]) Consider a sample $\{(\mathbf{x}_i, y_i), i = 1, 2, \cdots, n\}$ from the model (3.16). If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n \to \infty$, then with probability tending to one*

1. *There exists a local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}_1^T}, \hat{\boldsymbol{\beta}_2^T})^T$ such that $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-1/2})$, where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$.*

2. *$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.*

3. $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{L} N(\mathbf{0}, \gamma(1-\gamma)\Sigma_{11}^{-1}/f^2(0))$*, where* Σ_{11} *is the top left s-by-s submatrix of* Σ*.*

3.2.3 The adaptive Lasso

It has been shown that the Lasso is a very efficient method in the variable selection problems. The Lasso minimizes the penalized likelihood function by adding a L_1 penalty to the likelihood function. The Lasso shrinks the coefficients towards 0 as the penalty parameter increases, and some coefficients are shrunk exactly to 0 when the penalty parameter is large enough. It often increase the prediction accuracy by reducing the variance of the estimators. However, there are solid arguments against the Lasso oracle properties. [17] showed that the Lasso shrinkage produces biased estimates for the large coefficients, and thus it would be suboptimal in terms of estimation risk.

Whether the Lasso has the oracle properties is an important question demanding a definite answer due to its wide applications. [66] discussed this question in detail and showed that the Lasso will be inconsistent in some scenarios. In order to enjoy the oracle properties, [66] proposed the adaptive Lasso by putting different weights for different regression coefficients and proved that the adaptive Lasso enjoys the oracle properties.

For the quantile regression, consider the model (3.3), the adaptive Lasso penalized quantile regression ([58]) minimizes

$$\sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_{\gamma,j}| \tag{3.17}$$

with respect to $\boldsymbol{\beta}_\gamma$, where the weights are set to be $\hat{w}_j = 1/|\hat{\beta_{\gamma,j}}|^\tau$, $j = 1, 2, \cdots, p$, for some chosen $\tau > 0$, and $\hat{\boldsymbol{\beta}}_\gamma = (\hat{\beta_{\gamma,1}}, \hat{\beta_{\gamma,2}}, \cdots, \hat{\beta_{\gamma,p}})^T$ is the root-n consistent estimator of $\boldsymbol{\beta}$ which is given by

$$\hat{\boldsymbol{\beta}}_\gamma = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T\boldsymbol{\beta}).$$

As shown by [58], the minimization problem (3.17) can be casted into the linear programming

problem

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}(\gamma\xi_i + (1-\gamma)\zeta_i) + n\lambda_n \sum_{j=1}^{p} \hat{w}_j \eta_j \\ \text{subject to} \quad & \xi_i \geq 0,\ \zeta_i \geq 0,\ \xi_i - \zeta_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta},\ \ i = 1, \cdots, n, \\ & \eta_j \geq |\beta_j|,\ \ j = 1, \cdots, p. \end{aligned}$$

Similar to the SCAD penalized quantile regression, the above problem can be easily solved by many softwares. [58] also proved the oracle properties of the adaptive Lasso.

Remark 4 *(Theorem 3, [58]) Consider a sample* $\{(\mathbf{x}_i, y_i), i = 1, 2, \cdots, n\}$ *from the model (3.16). If* $\sqrt{n}\lambda_n \to 0$ *and* $n^{(\tau+1)/2}\lambda_n \to \infty$ *as* $n \to \infty$*, and let* $\hat{\boldsymbol{\beta}}_{AL} = (\hat{\boldsymbol{\beta}}_{AL,1}^T, \hat{\boldsymbol{\beta}}_{AL,2}^T)^T$ *be the solution of (3.17), then we have*

1. $\hat{\boldsymbol{\beta}}_{AL,2} = \mathbf{0}$.
2. $\sqrt{n}(\hat{\boldsymbol{\beta}}_{AL,1} - \boldsymbol{\beta}_{10}) \xrightarrow{L} N(\mathbf{0}, \gamma(1-\gamma)\sum_{11}^{-1} / f^2(0))$, *where* $\sum_{11}$ *is the top left s-by-s submatrix of* $\sum$.

3.3 Bayesian variable selection in the quantile regression

Bayesian analysis to the quantile regression begins with specifying a likelihood, which is often obtained from the ALD, due to the relationship between the ALD and the quantile regression firstly studied by [29]. This interesting finding motivates [61] to propose Bayesian quantile regression by adopting the ALD as the error distribution in the linear regression model, and they employed a random-walk Metropolis-Hastings algorithm for the posterior sampling. Thereafter, many researchers also started to perform quantile regression from the Bayesian perspective. [54] studied Bayesian quantile regression in a similar treatment as done by [61]. Later on, [60] explored the use of the ALD and proposed stochastic search variable selection (SSVS) for quantile regression models. [1] improved SSVS by introducing an informative prior, the power prior, for the regression coefficients. The resulting method is called the informative stochastic search variable selection (ISSVS). As [41] proposed Bayesian Lasso method based on the combined use of the Lasso regularization method and Bayesian approach in the linear regression model, Bayesian Lasso method and its variants have then been extensively studied in the literature, such as the Bayesian Lasso ([36]),

Bayesian adaptive Lasso ([2], [35]), to name just a few.

3.3.1 Stochastic search variable selection

A crucial problem in building regression model is the selection of predictors to be included. Considering the model (3.3), the problem is to find and fit the best model from the 2^p possible submodels. Numerous procedures based on the comparison of all 2^p possible submodels have been proposed, such as AIC, BIC. Unfortunately, it has been shown that the computational requirements for these procedures could be huge when p is large. In order to solve the computational issue, [20] proposed stochastic search variable selection (SSVS) algorithm. SSVS is based on embedding the entire regression setup in a hierarchical Bayesian normal mixture model, where latent variables are used to identify subset selection. A Gibbs sampler algorithm is used to sample from the multinomial posterior distribution on the set of possible subset choices. Thus those subsets with higher probability can be identified by their more frequent appearance in the Gibbs sampler. Therefore, the drawback of computing posterior probabilities for all 2^p subsets has been avoid.

[60] explored the use of SSVS algorithm and applied it to the quantile regression. Rewrite the model (3.3) as

$$y_i = \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.18}$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_p)^T \in \boldsymbol{\Omega}$ is a model index that $\tau_j = 1$ denotes the jth predictor is included in the model, and $\tau_j = 0$ means the jth predictor is excluded from the model, for $j = 1, 2, \cdots, p$, $\mathbf{x}_{\boldsymbol{\tau},i} = \{x_{ij} : \tau_j = 1\}$ and $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ are the corresponding vector of predictors and vector of coefficients in model $\boldsymbol{\tau}$, respectively, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ is the error vector as defined in the model (3.3).

We follow the work of [61] by using the asymmetric Laplace distribution (ALD) as the error distribution in the model (3.18). In particular, ε_j follows ALD$(0, 1, \gamma)$, for $j = 1, 2, \cdots, p$. Therefore, the density of y_i is given by

$$f(y_i \mid \boldsymbol{\beta}_{\boldsymbol{\tau}}, \boldsymbol{\tau}) = \gamma(1-\gamma) \exp\left\{-\rho_\gamma(y_i - \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}})\right\}.$$

As shown by [60], y_i can be written as a mixture of normals. In particular, let w_i be an

exponential random variable with rate $\gamma(1-\gamma)$. Then we have

$$y_i \mid \boldsymbol{\beta}_{\boldsymbol{\tau}}, \boldsymbol{\tau}, w_i \sim \mathrm{N}\left((1-2\gamma)w_i + \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}}, 2w_i\right), \quad i = 1, 2, \cdots .n. \tag{3.19}$$

Therefore, the likelihood function of $\mathbf{y}$ is given by

$$L(\mathbf{y} \mid \boldsymbol{\beta}_{\boldsymbol{\tau}}, \mathbf{w}, \boldsymbol{\tau}) \propto \left(\prod_{i=1}^{n} w_i^{-1/2}\right) \exp\left\{-\frac{1}{4}\sum_{i=1}^{n} \frac{(y_i - (1-2\gamma)w_i - \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}})^2}{w_i}\right\}, \tag{3.20}$$

where $\mathbf{w} = (w_1, w_2, \cdots, w_n)$.

The choice of priors for the unknown parameters is essential for the Bayesian variable selection. We first choose Bernoulli priors for the index $\boldsymbol{\tau}$

$$\pi(\boldsymbol{\tau}) = \prod_{j=1}^{p} \pi_0^{\tau_j}(1-\pi_0)^{1-\tau_j},$$

where π_0 is the prior probability of including a randomly selected predictor, and a convenient hyperprior for π_0 is Beta(a_0, b_0), where $a_0, b_0 > 0$.

We can embed all the submodels $\boldsymbol{\tau} \in \boldsymbol{\Omega}$ within the full model by letting $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ denote the coefficients on the p predictors in the full model, with $\beta_j = 0$ for all j such that $\tau_j = 0$. Therefore, we can simultaneously induce a prior for $\boldsymbol{\tau}$ and $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ by choosing a prior for $\boldsymbol{\beta}$ as

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\tau}) = \prod_{j=1}^{p} \{(1-\tau_j)\delta_0 + \tau_j N(\mathbf{0}, \lambda_j^{-1})\},$$

where δ_0 indicates a degenerate distribution with all its mass at 0, and λ_j follows Gamma$(1/2, 1/2)$ inducing that a heavy-tailed Cauchy prior marginally for the coefficients on the predictors selected in the model.

Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_p)$. Bayesian quantile regression can be written as a Bayesian hierar-

chical model

$$
\begin{aligned}
y_i &= \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}} + (1-2\gamma)w_i + \sqrt{2w_i} z_i, \\
\mathbf{w} &\sim \prod_{i=1}^{n} \gamma(1-\gamma)\exp\{-\gamma(1-\gamma)w_i\}, \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\}, \\
\boldsymbol{\beta}_{\boldsymbol{\tau}} \mid \boldsymbol{\tau}, \boldsymbol{\lambda} &\sim \prod_{j=1}^{p} \{(1-\tau_j)\delta_0 + \tau_j N(\mathbf{0}, \lambda_j^{-1})\}, \\
\boldsymbol{\tau} \mid \pi_0 &\sim \prod_{j=1}^{p} \pi_0^{\tau_j}(1-\pi_0)^{1-\tau_j}, \\
\boldsymbol{\lambda} &\sim \prod_{j=1}^{p} \lambda_j^{-\frac{1}{2}} \exp\left\{-\frac{\lambda_j}{2}\right\}, \\
\pi_0 &\sim \pi_0^{a_0-1}(1-\pi_0)^{b_0-1},
\end{aligned}
$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_n)$.

A Gibbs sampler algorithm proposed by [60] to sample from the joint posterior distribution is given below:

1. Set initial values for π_0, sample $\mathbf{w}$ and $\boldsymbol{\lambda}_{\boldsymbol{\tau}} = (\lambda_j : \tau_j = 1)$ from their priors, respectively.

2. Update the indicator τ_j marginalizing out $\boldsymbol{\beta}_{\boldsymbol{\tau}}$. Note that marginalizing out $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ gives

$$
\pi(\boldsymbol{\tau} \mid \mathbf{w}, \boldsymbol{\lambda}, \mathbf{y}) \propto g(\boldsymbol{\tau}) \equiv \left(\prod_{j,\tau_j=1} \lambda_j^{-\frac{1}{2}} \right) |\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^T \tilde{\mathbf{X}}_{\boldsymbol{\tau}}|^{-\frac{1}{2}} \cdot \exp\left\{ -\frac{1}{2} ||\tilde{\mathbf{u}} - \tilde{\mathbf{X}}_{\boldsymbol{\tau}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}||^2 \right\},
$$

where

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}} &= (\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^T \tilde{\mathbf{X}}_{\boldsymbol{\tau}})^{-1} \tilde{\mathbf{X}}_{\boldsymbol{\tau}}^T \tilde{\mathbf{u}}, \\
\tilde{\mathbf{X}}_{\boldsymbol{\tau}} &= \left[\sqrt{\frac{1}{2}} \mathbf{X}_{\boldsymbol{\tau}}^T \mathbf{W}^{\frac{1}{2}} \quad \text{diag}\{\sqrt{\lambda_j} : \tau_j = 1\} \right]^T, \\
\mathbf{X}_{\boldsymbol{\tau}} &= (x_{\boldsymbol{\tau},1}, x_{\boldsymbol{\tau},2}, \cdots, x_{\boldsymbol{\tau},3})^T, \\
\tilde{\mathbf{u}} &= \left[\sqrt{\frac{1}{2}} (\mathbf{y} - (1-2\gamma)\mathbf{w})^T \mathbf{W}^{\frac{1}{2}} \quad \mathbf{0} \right]^T, \\
\mathbf{W} &= \text{diag}\{w_i^{-1}, i = 1, 2, \cdots, n\}.
\end{aligned}
$$

Simulate τ_j from the Bernoulli distribution, Ber(π_1) with

$$\pi_1 = \frac{\pi_0 \; g(\tau_j = 1, \boldsymbol{\tau}_{-j})}{\pi_0 \; g(\tau_j = 1, \boldsymbol{\tau}_{-j}) + (1 - \pi_0) \; g(\tau_j = 0, \boldsymbol{\tau}_{-j})}, \quad j = 1, 2, \cdots, p,$$

where $\boldsymbol{\tau}_{-j} = \{\tau_k, k \neq j\}$.

3. Simulate $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ from the normal distribution, N($\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}, (\tilde{\mathbf{X}}_{\boldsymbol{\tau}}^T \tilde{\mathbf{X}}_{\boldsymbol{\tau}})^{-1}$).

4. Simulate w_i from the inverse Gaussian distribution, IG($a_i, 1/2$) with

$$a_i = \frac{1}{|y_i - \mathbf{x}_{\boldsymbol{\tau},i}^T \boldsymbol{\beta}_{\boldsymbol{\tau}}|}, \quad i = 1, 2, \cdots, n,$$

where the density of IG(a, b) is given by

$$f(x \mid a, b) = \sqrt{\frac{b}{2\pi}} x^{-\frac{3}{2}} \exp\left\{-\frac{b(x-a)^2}{2a^2 x}\right\}, \quad x > 0.$$

5. Simulate λ_j from the exponential distribution, Exp($(\beta_j^2 + 1)/2$).

6. Simulate π_0 from a Beta distribution, Beta($p_{\boldsymbol{\tau}} + a_0, p - p_{\boldsymbol{\tau}} + b_0$), where $p_{\boldsymbol{\tau}} = \sum_{j=1}^{p} \tau_j$.

3.3.2 Informative stochastic search variable selection

The SSVS algorithm ([20]) is an efficient method in Bayesian variable selection. [60] considered variable selection in the quantile regression through SSVS and obtained a nice result. This approach answers a difficulty question in the quantile regression, that is, how to identify promising subsets of predictors. However, SSVS has the inflexible disadvantage of relying on priors that are independent of the values of quantiles, that is, for different quantiles, the same prior is used. Since the characterization of data changes for different quantiles, it is more reasonable to consider different priors according to different quantiles. For example, the parameter values from a 90% quantile regression should be different than the ones from a 70% quantile regression, thus the priors used for modeling these two quantiles should be different.

With regard to prior selection, incorporating historical information in the process has a lot of advantages. First, informative priors can be used when choosing vague priors for the Bayesian regression formulation may cause instability in the posterior estimates for the Gibbs sampler ([24]).

Second, a prior with historical data can result in consistent estimates ([16]). Lastly, Bayesian analysis based on priors derived from the historical data has advantages in terms of power and estimation accuracy for decisions with small sizes ([15]).

One method to incorporate historical data into the prior selection is choosing a power prior ([24]), which is constructed by raising the likelihood function that based on the historical data to a power parameter which is between 0 and 1. The power parameter represents the proportion of the historical data needed. [1] adopted the power prior for variable selection and estimation in Bayesian quantile regression. A new procedure, the Bayesian informative stochastic search variable selection (ISSVS), is thus proposed.

Let $D_0 = ((y_{01}, \mathbf{x}_{01}), (y_{02}, \mathbf{x}_{02}), \cdots, (y_{0n_0}, \mathbf{x}_{0n_0}))$ denote the historical data with sample size n_0, and $D = ((y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \cdots, (y_n, \mathbf{x}_n))$ be the current data. Consider the model (3.18) and the same mixture representations in (3.19). Thus the posterior density of $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ based on one sample is given by

$$f(\boldsymbol{\beta}_{\boldsymbol{\tau}} \mid y_k, w_k) \propto (w_k)^{-\frac{1}{2}} \exp\left\{-\frac{(y_k - (1-2\gamma)w_k - \mathbf{x}_k^T \boldsymbol{\beta}_{\boldsymbol{\tau}})^2}{4w_k}\right\}.$$

Let $\mathbf{w}_0 = (w_{01}, \cdots, w_{0n_0})$ denote the vector of latent variables for $\mathbf{y}_0 = (y_{01}, \cdots, y_{0n_0})$. We define a joint power prior distribution for $(\boldsymbol{\beta}_{\boldsymbol{\tau}}, \boldsymbol{\tau}, \mathbf{w}_0)$

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\tau}}, \boldsymbol{\tau}, \mathbf{w}_0 \mid D_0, a_0) \propto \left(\prod_{k=1}^{n_0} [f(\boldsymbol{\beta}_{\boldsymbol{\tau}} \mid y_{0k}, w_{0k})]^{a_0}\right) \pi_0(\boldsymbol{\beta}_{\boldsymbol{\tau}}) \pi(\boldsymbol{\tau}),$$

where $0 \leq a_0 \leq 1$ is a fixed parameter which indicates the influence of the historical information on the current analysis, $\pi_0(\boldsymbol{\beta}_{\boldsymbol{\tau}})$ denotes the initial prior for $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ before any historical information is collected, and $\pi(\boldsymbol{\tau})$ is the prior for $\boldsymbol{\tau}$.

Similar to the SSVS in Section 3.3.1, we choose the initial prior for $\boldsymbol{\beta}$ as

$$\pi_0(\boldsymbol{\beta} \mid \boldsymbol{\tau}) = \prod_{j=1}^{p} \{(1-\tau_j)\delta_0 + \tau_j N(\mathbf{0}, \lambda_j^{-1})\},$$

where δ_0 indicates a degenerate distribution with all its mass at 0, and λ_j follows Gamma$(1/2, 1/2)$ inducing that a heavy-tailed Cauchy prior marginally for the coefficients on the predictors selected in the model.

For $j = 1, 2, \cdots, p$, the prior of τ_j is chosen to be an independent Bernoulli prior, Ber(π_1)

$$\pi(\boldsymbol{\tau}) = \prod_{j=1}^{p} \pi_1^{\tau_j}(1-\pi_1)^{1-\tau_j},$$

where $0 \leq \pi_1 \leq 1$ is the prior probability of including a randomly selected predictor, the role of π_1 is important since the small value of π_1 will indirectly prevent the number of selected covariates in the model. We treat π_1 as an unknown parameter and put a Beta prior, Beta(a, b) on it, where $a, b > 0$.

Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_p)$. It is easy to derive a prior probability of model $\boldsymbol{\tau}$ of the form

$$\begin{aligned}
\pi(\boldsymbol{\tau} \mid \boldsymbol{\lambda}, \pi_1, \mathbf{w}_0, D_0, a_0) &= \int \pi(\boldsymbol{\beta}_{\boldsymbol{\tau}}, \boldsymbol{\tau} \mid \boldsymbol{\lambda}, \pi_1, \mathbf{w}_0, D_0, a_0) d\boldsymbol{\beta}_{\boldsymbol{\tau}} \qquad (3.21)\\
&= \left(\prod_{j:\tau_j=1} \lambda_j^{-\frac{1}{2}} \right) |a_0 \mathbf{X}_{0\boldsymbol{\tau}}^T \mathbf{W}_0 \mathbf{X}_{0\boldsymbol{\tau}} + \boldsymbol{\Lambda}|^{-\frac{1}{2}}\\
&\times \prod_{j=1}^{p_{\boldsymbol{\tau}}} \pi_1^{\tau_j}(1-\pi_1)^{1-\tau_j} \exp\left\{ -\frac{a_0}{4} \mathbf{u}_0^T \mathbf{W}_0 \mathbf{u}_0 \right\}\\
&\times \exp\left\{ -\frac{1}{4} \mathbf{u}_0^T \mathbf{W}_0 \mathbf{X}_{0\boldsymbol{\tau}} |a_0 \mathbf{X}_{0\boldsymbol{\tau}}^T \mathbf{W}_0 \mathbf{X}_{0\boldsymbol{\tau}} + \boldsymbol{\Lambda}|^{-1} \mathbf{X}_{0\boldsymbol{\tau}}^T \mathbf{W}_0 \mathbf{u}_0 \right\},
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{u}_0 &= (u_{01}, u_{02}, \cdots, u_{0n_0}), \quad u_{0k} = y_{0k} - (1-2\gamma) w_{0k}, \quad k = 1, 2, \cdots, n_0,\\
\mathbf{X}_{0\boldsymbol{\tau}} &= (x_{01,\boldsymbol{\tau}}, x_{02,\boldsymbol{\tau}}, \cdots, x_{0n_0,\boldsymbol{\tau}})^T,\\
\mathbf{W}_0 &= \mathrm{diag}(w_{01}^{-1}, w_{02}^{-1}, \cdots, w_{0n_0}^{-1}),\\
\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_p),
\end{aligned}$$

and $p_{\boldsymbol{\tau}} = \sum_{j=1}^{p} \tau_j$ is the size of the $\boldsymbol{\tau}$th subset model. Note that the prior of $\boldsymbol{\tau}$ in (3.21) depends on the quantile γ, thus it changes automatically when we change the quantile.

Considering the current data D, the likelihood function of $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ is the same as the one in Section 3.3.1. Therefore, Bayesian quantile regression model can be written as a

hierarchical model

$$
\begin{aligned}
y_i &= \mathbf{x}_{i,\boldsymbol{\tau}}^T \boldsymbol{\beta}_{\boldsymbol{\tau}} + (1-2\gamma)w_i + \sqrt{2w_i}z_i, \\
\mathbf{w} &\sim \prod_{i=1}^{n} \gamma(1-\gamma)\exp\{-\gamma(1-\gamma)w_i\}, \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z_i^2}{2}\right\}, \\
y_{0k} &= \mathbf{x}_{0k,\boldsymbol{\tau}}^T \boldsymbol{\beta}_{\boldsymbol{\tau}} + (1-2\gamma)w_{0k} + \sqrt{2w_{0k}}z_{0k}, \\
\mathbf{w}_0 &\sim \prod_{k=1}^{n_0} \gamma(1-\gamma)\exp\{-\gamma(1-\gamma)w_{0k}\}, \\
\mathbf{z}_0 &\sim \prod_{k=1}^{n_0} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{z_{0i}^2}{2}\right\}, \\
\boldsymbol{\beta}_{\boldsymbol{\tau}} \mid \boldsymbol{\tau}, \boldsymbol{\lambda} &\sim \prod_{j=1}^{p_{\boldsymbol{\tau}}} \{(1-\tau_j)\delta_0 + \tau_j N(\mathbf{0}, \lambda_j^{-1})\}, \\
\boldsymbol{\tau} \mid \pi_1 &\sim \prod_{j=1}^{p_{\boldsymbol{\tau}}} \pi_1^{\tau_j}(1-\pi_1)^{1-\tau_j}, \\
\boldsymbol{\lambda} &\sim \prod_{j=1}^{p_{\boldsymbol{\tau}}} \lambda_j^{-\frac{1}{2}}\exp\left\{-\frac{\lambda_j}{2}\right\}, \\
\pi_1 &\sim \pi_1^{a-1}(1-\pi_1)^{b-1},
\end{aligned}
$$

where $\mathbf{z} = (z_1, z_2, \cdots, z_n)$ and $\mathbf{z}_0 = (z_{01}, z_{02}, \cdots, z_{0n_0})$.

[1] proposed a Gibbs sampler algorithm for the posterior sampling based on the above hierarchical model

1. Simulate $\boldsymbol{\beta}_{\boldsymbol{\tau}}$ from the multivariate normal distribution, $\mathrm{N}(\mu_{\boldsymbol{\beta}_{\boldsymbol{\tau}}}, \Sigma_{\boldsymbol{\beta}_{\boldsymbol{\tau}}})$ with

$$
\mu_{\boldsymbol{\beta}_{\boldsymbol{\tau}}} = \frac{1}{2}\Sigma_{\boldsymbol{\beta}_{\boldsymbol{\tau}}}(\mathbf{X}_{\boldsymbol{\tau}}^T\mathbf{W}\mathbf{u} + a_0\mathbf{X}_{0\boldsymbol{\tau}}^T\mathbf{W}_0\mathbf{u}_0) \quad \text{and} \quad \Sigma_{\boldsymbol{\beta}_{\boldsymbol{\tau}}} = \left(\frac{1}{2}\mathbf{X}_{\boldsymbol{\tau}}^T\mathbf{W}\mathbf{X}_{\boldsymbol{\tau}} + \frac{a_0}{2}\mathbf{X}_{0\boldsymbol{\tau}}^T\mathbf{W}_0\mathbf{X}_{0\boldsymbol{\tau}} + \boldsymbol{\Lambda}\right)^{-1},
$$

where

$$
\begin{aligned}
\mathbf{u} &= (u_1, u_2, \cdots, u_n), \quad u_i = y_i - (1-2\gamma)w_i, \\
\mathbf{X}_{\boldsymbol{\tau}} &= (x_{1,\boldsymbol{\tau}}, x_{2,\boldsymbol{\tau}}, \cdots, x_{n,\boldsymbol{\tau}})^T, \\
\mathbf{W} &= \mathrm{diag}(w_1^{-1}, w_2^{-1}, \cdots, w_n^{-1}).
\end{aligned}
$$

2. Simulate w_i from a generalized inverse Gaussian distribution, GIG$(1/2, a_{1i}, b_{1i})$ with

$$a_{1i} = \sqrt{\frac{(y_i - \mathbf{x}_{i,\boldsymbol{\tau}}^T \boldsymbol{\beta}_{\boldsymbol{\tau}})^2}{2}} \quad \text{and} \quad b_{1i} = \frac{1}{\sqrt{2}}, \quad i = 1, 2, \cdots, n,$$

and the density of GIG(p, a, b) is given by

$$f(x \mid p, a, b) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left\{-\frac{1}{2}\left(ax + \frac{b}{x}\right)\right\},$$

where $K_p(\cdot)$ is the modified Bessel function of the second kind.

3. Simulate w_{0k} from a generalized inverse Gaussian distribution, GIG$((-a_0+2)/2, a_{2k}, b_{2k})$ with

$$a_{2k} = \sqrt{\frac{a_0(y_{0k} - \mathbf{x}_{0k,\boldsymbol{\tau}}^T \boldsymbol{\beta}_{\boldsymbol{\tau}})^2}{2}} \quad \text{and} \quad b_{2k} = \sqrt{\frac{a_0(1-2\gamma)^2 + 4\gamma(1-\gamma)}{2}}, \quad k = 1, 2, \cdots, n_0.$$

4. Simulate λ_j from an exponential distribution, Exp$((\beta_j^2 + 1)/2)$.

5. Simulate τ_j from the Bernoulli distribution, Ber(π_2) with

$$\pi_2 = \left(1 + \frac{(1-\pi_1)\ g(\tau_j = 0, \boldsymbol{\tau}_{-j})}{\pi_1\ g(\tau_j = 1, \boldsymbol{\tau}_{-j})}\right)^{-1},$$

where

$$\begin{aligned} g(\boldsymbol{\tau}) &= \pi(\boldsymbol{\tau} \mid \boldsymbol{\lambda}, \pi_1, \mathbf{w}_0, D_0, a_0)\ |\mathbf{X}_{\boldsymbol{\tau}}^T \mathbf{W} \mathbf{X}_{\boldsymbol{\tau}}|^{-\frac{1}{2}} \\ &\times \exp\left\{-\frac{1}{4}\mathbf{u}^T \mathbf{W} \mathbf{u} - \frac{1}{4}\mathbf{u}^T \mathbf{W} \mathbf{X}_{\boldsymbol{\tau}} (\mathbf{X}_{\boldsymbol{\tau}}^T \mathbf{W} \mathbf{X}_{\boldsymbol{\tau}})^{-1} \mathbf{X}_{\boldsymbol{\tau}}^T \mathbf{W} \mathbf{u}\right\}, \end{aligned}$$

$\boldsymbol{\tau}_{-j} = \{\tau_t, t \neq j\}$, and $\pi(\boldsymbol{\tau} \mid \boldsymbol{\lambda}, \pi_1, \mathbf{w}_0, D_0, a_0)$ is given by (3.21).

6. Simulate π_1 from a Beta distribution, Beta$(p_{\boldsymbol{\tau}} + a, p - p_{\boldsymbol{\tau}} + b)$.

3.3.3 The Bayesian Lasso

For the linear regression model (3.5), the Lasso ([53]) estimates often are viewed as L_1 penalized least squares estimates. They achieve

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{3.22}$$

where $\lambda \geq 0$. [53] suggested that Lasso estimates can be interpreted as the posterior modes when the regression parameters have independent and identical Laplace priors. Motivated by this connection, [41] considered a fully Bayesian analysis using a conditional Laplace prior for the regression coefficients

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} \exp\left\{ -\frac{\lambda|\beta_j|}{\sqrt{\sigma^2}} \right\}.$$

Conditional on σ^2 is important since it ensures a unimodel full posterior ([41]). Based on the mixture representations of the Laplace distribution, [41] established a hierarchical representation of the full model which can be used to conduct an efficient Gibbs sampler algorithm.

Motivated by the work of [41], [36] considered the variable selection in the quantile regression with the L_1 penalty. Consider the model (3.3), due to the close relationship between the ALD and the quantile regression proposed by [61], an asymmetric Laplace error distribution ALD$(0, \tau, \gamma)$ is considered for the error ε_i's. Recently, [34] proved that the ALD can be written as a scale mixture of normals with the scale mixing parameter following an exponential distribution.

Remark 5 *(Lemma 1, [36]) Suppose that v is a standard exponential random variable and z is a standard normal random variable. For $\gamma \in (0, 1)$, denote*

$$\xi_1 = \frac{1 - 2\gamma}{\gamma(1-\gamma)} \quad \textit{and} \quad \xi_2 = \sqrt{\frac{2}{\gamma(1-\gamma)}}.$$

It follows that the variable $\varepsilon = \xi_1 v + \xi_2 \sqrt{v} z$ follows the ALD$(0, 1, \gamma)$.

Let $\mathbf{v} = (v_1, v_2, \cdots, v_n)$ and $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, and consider the ALD$(0, \tau, \gamma)$ as the error

distribution, then the model (3.3) can be rewritten as the following hierarchical model

$$\begin{aligned} y_i &= \mathbf{x}_i^T\boldsymbol{\beta} + \xi_1 v_i + \tau^{-\frac{1}{2}}\xi_2\sqrt{v_i}z_i, \\ \mathbf{v} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp\{-\tau v_i\}, \\ \mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z_i^2\right\}. \end{aligned} \tag{3.23}$$

The quantile regression with the Lasso penalty solves

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Similar to [41], we put a Laplace prior on β_j

$$\pi(\boldsymbol{\beta}) = \left(\frac{\tau\lambda}{2}\right)^p \exp\left\{-\tau\lambda \sum_{j=1}^{p} |\beta_j|\right\}.$$

As shown by [4], the prior of $\boldsymbol{\beta}$ can be written as a mixture of normals. Let $\eta = \tau\lambda$,

$$\pi(\boldsymbol{\beta} \mid \eta) = \prod_{j=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\} \frac{\eta^2}{2} \exp\left\{-\frac{\eta^2 s_j}{2}\right\} ds_j. \tag{3.24}$$

Denote $\mathbf{s} = (s_1, s_2, \cdots, s_p)$. We put a Gamma prior, Gamma(a, b) on τ and a Gamma prior, Gamma(c, d) on η^2 due to the conjugacy, where a, b, c, and d are known positive constant. Therefore, Bayesian Lasso quantile regression is a hierarchical model given by

$$\begin{aligned} y_i &= \mathbf{x}_i^T\boldsymbol{\beta} + \xi_1 v_i + \tau^{-\frac{1}{2}}\xi_2\sqrt{v_i}z_i, \\ \mathbf{v} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp\{-\tau v_i\}, \\ \mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z_i^2\right\}, \\ \boldsymbol{\beta} \mid \mathbf{s} &\sim \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\}, \\ \mathbf{s} \mid \eta^2 &\sim \prod_{j=1}^{p} \frac{\eta^2}{2} \exp\left\{-\frac{\eta^2 s_j}{2}\right\}, \\ \tau, \eta^2 &\sim \tau^{a-1}\exp\{-b\tau\} \cdot (\eta^2)^{c-1}\exp\{-d\eta^2\}, \end{aligned}$$

which yields a Gibbs sampler algorithm ([36]) below

1. Simulate β_j from a normal distribution, $\mathrm{N}(\mu_j, \sigma_j^2)$ with

$$\mu_j = \frac{\sigma_j^2 \tau \sum_{i=1}^n \frac{y_{ij} x_{ij}}{v_i}}{\xi_2^2} \quad \text{and} \quad \sigma_j^{-2} = \frac{\tau}{\xi_2^2} \sum_{i=1}^n \frac{x_{ij}^2}{v_i} + \frac{1}{s_j}, \quad j = 1, 2, \cdots, p,$$

   where $y_{ij} = y_i - \xi_1 v_i - \sum_{k=1, k \neq j}^p x_{ik} \beta_k$.

2. Simulate v_i from a generalized inverse Gaussian, $\mathrm{GIG}(1/2, a_{1i}, b_{1i})$ with

$$a_{1i} = \frac{\tau \xi_1^2}{\xi_2^2} + 2\tau \quad \text{and} \quad b_{1i} = \frac{\tau (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2}{\xi_2^2}, \quad i = 1, 2, \cdots, n.$$

3. Simulate s_j from a generalized inverse Gaussian, $\mathrm{GIG}(1/2, \eta^2, \beta_j^2)$.

4. Simulate τ from a Gamma distribution, $\mathrm{Gamma}(a + 3n/2, b_\tau)$ with

$$b_\tau = b + \sum_{i=1}^n \left(\frac{(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \xi_1 v_i)^2}{2 \xi_2^2 v_i} + v_i \right).$$

5. Simulate η^2 from a Gamma distribution, $\mathrm{Gamma}(p + c, b_\eta)$ with

$$b_\eta = d + \sum_{j=1}^p \frac{s_j}{2}.$$

3.3.4 The Bayesian elastic net

Due to the nature of L_1 penalty, the Lasso does both continuous shrinkage and automatic variable selection simultaneously, therefore, it has been shown successful in many situations. However, the Lasso still have two drawbacks. First, when the number of regression coefficients is greater than the sample size of the data, the Lasso selects at most n (n is the sample size) variables because the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Second, if there is a group of variables among which the pairwise correlations are very high, then the Lasso tends to select only one variable from the group and does not care which one is selected. Those two drawbacks make the Lasso an inappropriate method in some situations.

Motivated by those facts, [67] proposed a new regularization method which is called the elastic net. Similar to the Lasso, the elastic net does continuous shrinkage and automatic variable

selection simultaneously, it also can select groups of correlated variables. Consider the regular linear model (3.5), the elastic net estimates are defined as ([67])

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\},$$

where both λ_1 and λ_2 are non-negative numbers. Note that the elastic net penalty is a combination of the Lasso (L_1) penalty and the ridge (L_2) penalty. When $\lambda_1 = 0$, the elastic net becomes the ridge regression, and it degenerates to the Lasso regression when $\lambda_2 = 0$.

Consider the model (3.3) and the check function as the loss function, [36] discussed the quantile regression with the elastic net penalty, which solve the minimization problem

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Similar to Section 3.3.3, we consider an asymmetric Laplace error distribution $\text{ALD}(0, \tau, \gamma)$, then $\mathbf{y}$ has the same mixture representations as in (3.23). Let $\eta_1 = \tau\lambda_1$ and $\eta_2 = \tau\lambda_2$, we set the prior of β_j as

$$\pi(\beta_j \mid \eta_1, \eta_2) = C(\eta_1, \eta_2) \frac{\eta_1}{2} \exp\{-\eta_1 |\beta_j| - \eta_2 \beta_j^2\},$$

where $C(\eta_1, \eta_2)$ is a normalizing constant which is given by

$$C(\eta_1, \eta_2) = \Gamma^{-1} \left(\frac{1}{2}, \frac{\eta_1^2}{4\eta_2} \right) \left(\frac{\eta_1^2}{4\eta_2} \right)^{-\frac{1}{2}} \exp\left\{ -\frac{\eta_1^2}{4\eta_2} \right\},$$

where $\Gamma(\cdot, \cdot)$ is the upper incomplete Gamma function.

Therefore, we consider $\tilde{\eta_1} = \eta_1^2/(4\eta_2)$ and put Gamma priors on τ, $\tilde{\eta_1}$ and η_2 due to the conjugacy. The Bayesian quantile regression with the elastic net penalty is determinated by the

following Bayesian hierarchical model

$$\begin{aligned}
y_i &= \mathbf{x}_i^T\boldsymbol{\beta} + \xi_1 v_i + \tau^{-\frac{1}{2}}\xi_2\sqrt{v_i}z_i, \\
\mathbf{v} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp\{-\tau v_i\}, \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\}, \\
\boldsymbol{\beta} \mid \mathbf{t}, \eta_2 &\sim \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi(t_j-1)/(2\eta_2 t_j)}} \exp\left\{-\frac{1}{2}\left(\frac{t_j-1}{2\eta_2 t_j}\right)^{-1}\beta_j^2\right\}, \\
\mathbf{t} \mid \tilde{\eta_1} &\sim \prod_{j=1}^{p} \Gamma^{-1}\left(\frac{1}{2}, \tilde{\eta_1}\right) t_j^{-\frac{1}{2}} \tilde{\eta_1}^{-\frac{1}{2}} \exp\{-\tilde{\eta_1} t_j\} I(t_j > 1), \\
\tau, \tilde{\eta_1}, \eta_2 &\sim \tau^{a-1}\exp\{-b\tau\} \cdot \tilde{\eta_1}^{c_1-1}\exp\{-d_1\tilde{\eta_1}\} \cdot \eta_2^{c_2-1}\exp\{-d_2\eta_2\},
\end{aligned}$$

where $\mathbf{t} = (t_1, t_2, \cdots, t_p)$, and $a, b, c_1, c_2, d_1, d_2 \geq 0$.

A Gibbs sampler algorithm was proposed by ([36]):

1. Simulate β_j from a normal distribution, $\mathrm{N}(\mu_j, \sigma_j^2)$ with

$$\mu_j = \frac{\sigma_j^2\tau\sum_{i=1}^{n}\frac{y_{ij}x_{ij}}{v_i}}{\xi_2^2} \quad \text{and} \quad \sigma_j^{-2} = \frac{\tau}{\xi_2^2}\sum_{i=1}^{n}\frac{x_{ij}^2}{v_i} + \frac{2\eta_2 t_j}{t_j-1}, \quad j = 1, 2, \cdots, p,$$

   where $y_{ij} = y_i - \xi_1 v_i - \sum_{k=1, k\neq j}^{p} x_{ik}\beta_k$.

2. Simulate v_i from a generalized inverse Gaussian distribution, $\mathrm{GIG}(1/2, a_i, b_i)$ with

$$a_i = \frac{\tau\xi_1^2}{\xi_2^2} + 2\tau \quad \text{and} \quad b_i = \frac{\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{\xi_2^2}, \quad i = 1, 2, \cdots, n.$$

3. Simulate $t_j - 1$ from a generalized inverse Gaussian distribution, $\mathrm{GIG}(1/2, 2\tilde{\eta_1}, 2\eta_2\beta_j^2)$.

4. Simulate τ from a Gamma distribution, $\mathrm{Gamma}(a + 3n/2, b_\tau)$ with

$$b_\tau = b + \sum_{i=1}^{n}\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1 v_i)^2}{2\xi_2^2 v_i}.$$

5. Simulate η_2 from a Gamma distribution, $\mathrm{Gamma}(p/2 + c_2, b_{\eta_2})$ with

$$b_{\eta_2} = d_2 + \sum_{j=1}^{p}\frac{t_j\beta_j^2}{t_j - 1}.$$

6. Simulate $\tilde{\eta_1}$ from its full conditional distribution through Metropolis-Hastings algorithm

$$\pi(\tilde{\eta_1} \mid \boldsymbol{\beta}, \mathbf{v}, \mathbf{t}, \tau, \eta_2, \mathbf{y}) \propto \Gamma^{-p}\left(\frac{1}{2}, \tilde{\eta_1}\right) \tilde{\eta_1}^{\frac{p}{2}+c_1-1} \exp\left\{-\tilde{\eta_1}\left(d_1 + \sum_{j=1}^{p} t_j\right)\right\}.$$

3.3.5 The Bayesian group Lasso

In many regression problems we are interested in finding explanatory factors in predicting the response variables, where each explanatory factor may be represented by a group of input variables. In such cases, variable selection corresponds to the selection of groups of the variables. Although the Lasso enjoys great computational advantage and excellent performance, it is designed to select individual variables, not for general factor selection.

Consider the model (3.5), and suppose the predictors are grouped into G groups and $\boldsymbol{\beta}_g$ is the coefficient vector of the gth group and $\mathbf{x}_{ig}$ is the corresponding predictors, $g = 1, 2, \cdots, G$. Then $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \cdots, \boldsymbol{\beta}_G^T)^T$ and $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \mathbf{x}_{i2}^T, \cdots, \mathbf{x}_{iG}^T)^T$, and the model (3.5) can be written as

$$y_i = \sum_{g=1}^{G} \mathbf{x}_{ig}^T \boldsymbol{\beta}_g + \varepsilon_i, \quad i = 1, 2, \cdots, n.$$

When the Lasso is applied to the above model, it tends to make selection based on the strength of individual variables rather than the strength of group variables, often resulting selecting more factors than necessary ([64]). Another drawback of using the Lasso is that the result heavily relies on how the factors are orthonormalized, we may obtain a very different set of factors if any factor is reparameterized. Motivated by those facts, [64] considered the extension of the Lasso and proposed the group Lasso which can be used for factor selection. The group Lasso estimates are defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_{\mathbf{K}_g} \right\},$$

where $\lambda \geq 0$, $||\boldsymbol{\beta}_g||_{\mathbf{K}_g} = (\boldsymbol{\beta}_g^T \mathbf{K}_g \boldsymbol{\beta}_g)^{-1/2}$ for some known positive definite matrix $\mathbf{K}_g$, $g = 1, 2, \cdots, G$. Note that the group Lasso penalty is intermediate between the L_1-normal penalty and the L_2-normal penalty, and it reduces to the Lasso when $G = p$.

The group Lasso has also been considered for the quantile regression. Consider the model (3.3) and the same partition of predictors as above, the group Lasso regularized quantile regression

solves the following minimization problem

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} ||\boldsymbol{\beta}_g||_{\mathbf{K}_g} \right\}.$$

Similar to Section 3.3.3, we consider an asymmetric Laplace error distribution ALD$(0, \tau, \gamma)$, then $\mathbf{y}$ has the same mixture representations as in (3.23). Let $\eta = \tau\lambda$, d_g be the dimension of the vector $\boldsymbol{\beta}_g$. We put a Laplace prior on $\boldsymbol{\beta}_g$

$$\pi(\boldsymbol{\beta}_g \mid \eta) = C_{d_g} \sqrt{\det(\mathbf{K}_g)} \eta^{d_g} \exp\{-\eta ||\boldsymbol{\beta}_g||_{\mathbf{K}_g}\},$$

where $C_{d_g} = 2^{-(d_g+1)/2} (2\pi)^{-(d_g-1)/2} / \Gamma((d_g+1)/2)$, and $\Gamma(\cdot)$ is the Gamma function.

Therefore, we can write the prior of $\boldsymbol{\beta}_g$ as a mixture of normals that similar to (3.24). We put Gamma priors on τ and η^2 due to conjugacy, then Bayesian group Lasso quantile regression is a Bayesian hierarchical model given by

$$\begin{aligned}
y_i &= \mathbf{x}_i^T \boldsymbol{\beta} + \xi_1 v_i + \tau^{-\frac{1}{2}} \xi_2 \sqrt{v_i} z_i, \\
\mathbf{v} \mid \tau &\sim \prod_{i=1}^{n} \tau \exp\{-\tau v_i\}, \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\}, \\
\boldsymbol{\beta} \mid \mathbf{s} &\sim \prod_{g=1}^{G} (\det(s_g \mathbf{K}_g^{-1}))^{-\frac{1}{2}} \exp\left\{-\frac{1}{2 s_g} \boldsymbol{\beta}_g^T \mathbf{K}_g \boldsymbol{\beta}_g\right\}, \\
\mathbf{s} \mid \eta^2 &\sim \prod_{g=1}^{G} \left(\frac{\eta^2}{2}\right)^{\frac{d_g+1}{2}} s_g^{\frac{d_g-1}{2}} \exp\left\{-\frac{\eta^2}{2} s_g\right\}, \\
\tau, \eta^2 &\sim \tau^{a-1} \exp\{-b\tau\} \cdot (\eta^2)^{c-1} \exp\{-d\eta^2\},
\end{aligned}$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_G)$, and $a, b, c, d \geq 0$.

Based on the above hierarchical model, a Gibbs sampler algorithm was proposed by ([36]):

1. Simulate β_g from a multivatiate normal distribution, N$(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ with

$$\boldsymbol{\mu}_g = \frac{\boldsymbol{\Sigma}_g \tau \sum_{i=1}^{n} \frac{y_{ig} \mathbf{x}_{ig}}{v_i}}{\xi_2^2} \quad \text{and} \quad \boldsymbol{\Sigma}_g^{-1} = \frac{\tau}{\xi_2^2} \sum_{i=1}^{n} \frac{\mathbf{x}_{ig} \mathbf{x}_{ig}^T}{v_i} + \frac{1}{s_g} \mathbf{K}_g, \quad g = 1, 2, \cdots, G,$$

   where $y_{ig} = y_i - \xi_1 v_i - \sum_{k=1, k \neq g}^{G} \mathbf{x}_{ik}^T \boldsymbol{\beta}_k$.

2. Simulate v_i from a generalized inverse Gaussian distribution, GIG$(1/2, a_{1i}, b_{1i})$ with

$$a_{1i} = \frac{\tau\xi_1^2}{\xi_2^2} + 2\tau \quad \text{and} \quad b_{1i} = \frac{\tau(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{\xi_2^2}, \quad i = 1, 2, \cdots, n.$$

3. Simulate s_g from a generalized inverse Gaussian distribution, GIG$(1/2, a_{2g}, b_{2g})$ with

$$a_{2g} = \eta^2 \quad \text{and} \quad b_{2g} = \frac{\boldsymbol{\beta}_g^T\mathbf{K}_g\boldsymbol{\beta}_g}{s_g}.$$

4. Simulate τ from a Gamma distribution, Gamma$(a + 3n/2, b_\tau)$ with

$$b_\tau = b + \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \xi_1 v_i)^2}{2\xi_2^2 v_i}.$$

5. Simulate η^2 from a Gamma distribution, Gamma$((p + G)/2 + c, b_\eta)$ with

$$b_\eta = d + \sum_{g=1}^{G} \frac{s_g}{2}.$$

3.3.6 The Bayesian adaptive Lasso

With regards to the Lasso regression, [66] proved that the adaptive Lasso regression enjoys the oracle properties reported by [17] that the Lasso does not have. Then the adaptive Lasso receives much attention as an extension of the Lasso. [35] discussed the Bayesian adaptive Lasso by putting different Laplace priors for different regression coefficients. Similar to [41], [35] also established a hierarchical Bayesian model based on the mixture representation of the Laplace priors.

[2] employed this idea to the quantile regression. Consider the model

$$y_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{3.25}$$

where $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is the response vector, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent error terms whose distribution is unknown, but is restricted to have the γth quantile equal to zero. Note that this model is different from the model (3.3) in which it involves the intercept β_0.

As shown in Section 3.3.3, $\mathbf{y}$ can be written as a mixture of normals if we consider the asymmetric Laplace error distribution ALD$(0, 1/\sigma, \gamma)$. In particular, let $\theta = (1-2\gamma)/(\gamma-\gamma^2)$ and $\phi^2 = 2/(\gamma-\gamma^2)$, we have

$$y_i = \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \theta z_i + \phi\xi_i\sqrt{\sigma^{-1}z_i},$$

where z_i follows an exponential distribution with rate σ and ξ_i folllows a standard normal distribution, for $i = 1, 2, \cdots, n$. We put a Laplace prior on each β_j

$$\pi(\beta_j \mid \sigma, \lambda_j) = \frac{\sigma^{1/2}}{2\lambda_j}\exp\left\{-\frac{\sigma^{1/2}|\beta_j|}{\lambda_j}\right\}.$$

Note that the penalty coefficient for β_j is $\sigma^{1/2}/\lambda_j$, thus different penalty parameters are put on the different regression coefficients.

The Laplace prior of $\boldsymbol{\beta}$ can be written as a mixture of normals as shown in (3.24). Let $\boldsymbol{\lambda} = (\lambda_1^2, \lambda_2^2, \cdots, \lambda_p^2)$. We put an inverse Gamma prior on each λ_j^2 and a Gamma prior on σ.

$$\pi(\boldsymbol{\lambda}, \sigma \mid \delta, \tau) = \prod_{j=1}^{p}\frac{\tau^\delta}{\Gamma(\delta)}(\lambda_j^2)^{-1-\delta}\exp\left\{-\frac{\tau}{\lambda_j^2}\right\}\cdot\sigma^{a-1}\exp\{-b\sigma\},$$

where $a, b \geq 0$, and δ and τ are two positive hyperparameters. Since smaller τ and larger δ lead to bigger penalization, it is important to treat τ and δ as unkown parameters to avoid enforcing specific values that affect the estimates of the regression coefficients ([59]). Let $\mathbf{z} = (z_1, z_2, \cdots, z_n)$ and $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)$. Therefore, a Bayesian hierarchical model for Bayesian adaptive Lasso

quantile regression is given by

$$\begin{aligned}
y_i &= \beta_0 + \mathbf{x}_i^T\boldsymbol{\beta} + \theta z_i + \phi\xi_i\sqrt{\sigma^{-1}z_i}, \\
\mathbf{z} \mid \sigma &\sim \prod_{i=1}^{n} \sigma \exp\{-\sigma z_i\}, \\
\boldsymbol{\xi} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\xi_i^2}{2}\right\}, \\
\beta_0 &\sim 1, \\
\boldsymbol{\beta} \mid \mathbf{s} &\sim \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\}, \\
\mathbf{s} \mid \sigma, \boldsymbol{\lambda} &\sim \prod_{j=1}^{p} \frac{\sigma}{2\lambda_j^2} \exp\left\{-\frac{\sigma s_j}{2\lambda_j^2}\right\}, \\
\boldsymbol{\lambda} \mid \delta, \tau &\sim \prod_{j=1}^{p} \frac{\tau^\delta}{\Gamma(\delta)} (\lambda_j^2)^{-1-\delta} \exp\left\{-\frac{\tau}{\lambda_j^2}\right\}, \\
\sigma &\sim \sigma^{a-1} \exp\{-b\sigma\}, \\
\tau, \delta &\sim \frac{1}{\tau}.
\end{aligned}$$

[2] proposed a Gibbs sampler algorithm:

1. Simulate β_0 from a normal distribution, $\mathrm{N}(\hat{\beta}_0, \sigma^2_{\beta_0})$ with

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta} - \theta z_i), \quad \text{and} \quad \sigma^2_{\beta_0} = \frac{\sigma\phi^2}{n^2}\sum_{i=1}^{n} z_i.$$

2. Simulate z_i^{-1} from the inverse Gaussian distribution, $\mathrm{IG}(a_i, b_i)$ with

$$a_i = \sqrt{\frac{\theta^2 + 2\phi^2}{(y_i - \beta_0 - \mathbf{x}_i^T\boldsymbol{\beta})^2}} \quad \text{and} \quad b_i = \frac{\sigma(\theta^2 + 2\phi^2)}{\phi^2}, \quad i = 1, 2, \cdots, n.$$

3. Simulate β_j from a normal distribution, $\mathrm{N}(\hat{\beta}_j, \sigma^2_{\beta_j})$ with

$$\hat{\beta}_j = \frac{\sigma\sigma^2_{\beta_j}}{\phi^2}\sum_{i=1}^{n}\frac{x_{ij}}{z_i}(y_i - \beta_0 - \sum_{k\neq j} x_{ik}\beta_k - \theta z_i) \quad \text{and} \quad \sigma^{-2}_{\beta_j} = \frac{\sigma}{\phi^2}\sum_{i=1}^{n}\frac{x_{ij}^2}{z_i} + \frac{1}{s_j}, \quad j = 1, 2, \cdots, p.$$

4. Simulate s_j from the inverse Gaussian distribution, $\text{IG}(c_j, d_j)$ with

$$c_j = \sqrt{\frac{\beta_j^2 \lambda_j^2}{\sigma}} \quad \text{and} \quad d_j = \beta_j^2.$$

5. Simulate σ from a Gamma distribution, $\text{Gamma}(a_\sigma, b_\sigma)$ with

$$a_\sigma = \frac{3n}{2} + p + a \quad \text{and} \quad b_\sigma = \sum_{i=1}^{n} \left(\frac{(y_i - \beta_0 - \mathbf{x}_i^T \boldsymbol{\beta} - \theta z_i)^2}{2\phi^2 z_i} + z_i \right) + \sum_{j=1}^{p} \frac{s_j}{2\lambda_j^2} + b.$$

6. Simulate λ_j^2 from an inverse Gamma distribution, $\text{Inverse-Gamma}(1 + \delta, \sigma s_j/2 + \tau)$.

7. Simulate τ from a Gamma distribution, $\text{Gamma}(p\delta, \sum_{j=1}^{p} \lambda_j^{-2})$.

8. Simulate δ from its full conditional posterior distribution through Methopolis-Hastings algorithm

$$\pi(\delta \mid \boldsymbol{\lambda}, \tau) \propto \frac{\tau^{p\delta}}{(\Gamma(\delta))^p} \prod_{j=1}^{p} \frac{1}{\lambda_j^{2\delta}}.$$

3.4 Discussion

We have reviewed many variable selection methods in the quantile regression, both frequentist and Bayesian. Note that all those methods consider quantile at some fixed value. However, if our purpose is, among all the quantile regression models, to identify which one fits the data best, then the traditional quantile regression may not be appropriate. For example, given a range of quantile, $(0.1, 0.2, \cdots, 0.9)$, we could fit 9 different regression models according to each quantile value, we are interested in which one is the most probable one to exact the most information from the data. That is, which model could reflect the inner relationship of the data and which quantile would be the most likely one. In such cases, those questions can be easily answered if we consider the quantile as an unknown parameter and estimate it from the data.

On the other hand, consider the quantile regression model $y = x_1\beta_1 + x_2\beta_2 + \varepsilon$, with the γth quantile of y being $x_1\beta_1 + x_2\beta_2$. Given β_1, we may be interested in finding the representative quantile of the unobservable distribution. The traditional quantile regression focus on the graph of β_1 as a function of γ, we are interested in the graph of the quantile γ as a function of β_1. That

means, given β_1, the estimated quantile γ is the most probable one for the data. Those situations are often encountered in economic problems. For example, for the U.S. wage data, the effect of policy variables on distributional outcomes are of fundamental interest. Of particular interest is the estimation of the quantile treatment effects, that is, the effect of some policy variables on the different quantiles of a conditional response variable. In such case, regarding the quantile as an unknown parameter could help us complement the quantile treatment effects analysis by estimating it at the most probable quantile value.

Therefore, we desire to consider the variable selection problem in the quantile regression with unknown quantile. An important question is how to choose the prior for the quantile parameter γ. Due to the lack of information on γ, objective priors seem more reasonable, such as the Jeffreys prior and the reference priors. However, one potential issue in using such priors is computational complexity. Both Jeffreys and reference priors are derived from the Fisher information matrix of the model. Bayesian analysis to the quantile regression often begins with specifying a likelihood which is obtained from the ALD, due to the relationship between the ALD and quantile regression firstly studied by [29]. A mixture representation of the ALD is often used to obtain an efficient Gibbs sampler algorithm. Thus the Fisher information matrix has dimension (n+2)×(n+2), it is quite difficult to calculate the determinant or the inverse of this matrix when the sample size n is large. Therefore, as an alternative way, we consider several other objective priors.

Given that the quantile γ is a number between 0 and 1, some priors such as Beta prior, logistic normal prior, and uniform prior can be considered. If we do have some information about γ, for example, if γ is more likely to concentrate on a known number μ, a Beta distribution, Beta(α, β) with the distribution mean $\alpha/(\alpha + \beta)$ equal to μ can be considered; if γ is more likely to near 0, then we can choose Beta(α, β) with $\alpha < \beta$ as our prior. The idea here is quite similar to the reference priors with partical information that we discussed in Chapter 2. If we do not have any information about γ, a uniform prior within interval $(0, 1)$ can be a good choice since we consider every possible value of γ with equal chance. Of particular note is that, no matter what prior is used, the full conditional posterior distribution of γ often do not have a standard form due to the mixture representation of the ALD. Thus the Metropolis-Hasting algorithm might be considered to sample γ. We propose a Metropolis within Gibbs algorithm to deal with the variable selection problem in the next Chapter.

Chapter 4

Variable Selection in Bayesian Maximum Entropy Quantile Regression

4.1 Introduction

Since the seminar work of [27], quantile regression has gained increasing popularity due to its two main advantages. First, it provides richer information in the effects of the predictors on the different quantiles of the response variable than the one under the regular mean regression. Second, it is very insensitive to heteroscedasticity and outliers, thus quantile regression can accommodate non-normal errors, which are commonly encountered in many practical applications. These two appealing features of quantile regression result in its broader application in a wide range of disciplines, such as economics ([23]), survival analysis ([28]), biology ([21]), and microarray study ([55]).

Although many frequentist methods for the quantile regression have been developed (see [27], [29]), a Bayesian approach that enables exact inference even when the sample size is small or moderate has been proposed by [61]. Most existing method consider quantile regression at a fixed value, that is, give the data, we are always asked to fit a 90% quantile regression model or a 75% quantile regression model. However, if our purpose is, among all the quantile regression models, to identify which one fits the data best, then the traditional quantile regression may not be appropriate.

For example, given a range of quantile, $(0.1, 0.2, \cdots, 0.9)$, we could fit 9 different regression models according to each quantile value, we are interested in which one is the most probable one to exact the most information from the data. That is, which model could reflect the inner relationship of the data and which quantile would be the most likely one. In such cases, those questions can be easily answered if we consider the quantile as an unknown parameter and estimate it from the data.

On the other hand, consider the quantile regression model $y = x_1\beta_1 + x_2\beta_2 + \varepsilon$, with the γth quantile of y being $x_1\beta_1 + x_2\beta_2$. Given β_1, we may be interested in finding the representative quantile of the unobservable distribution. The traditional quantile regression focus on the graph of β_1 as a function of γ, we are interested in the graph of the quantile γ as a function of β_1. That means, given β_1, the estimated quantile γ is the most probable one for the data. Those situations are often encountered in economic problems. For example, for the U.S. wage data, the effect of policy variables on distributional outcomes are of fundamental interest. Of particular interest is the estimation of the quantile treatment effects, that is, the effect of some policy variables on the different quantiles of a conditional response variable. In such case, regarding the quantile as an unknown parameter could help us complement the quantile treatment effects analysis by estimating it at the most probable quantile value. Therefore, those facts motivate us to consider quantile as an unknown parameter.

In the quantile regression, normally, it is not necessary to specify the distribution of the error term as it is allowed to take any form. In order to jointly estimate the quantile parameter and regression coefficients, we obtain the error distribution by maximizing Shannon's entropy measure subject to two moment constraints, we refer it to the maximum entropy quantile regression (see [9]). The resulting error distribution, as shown by [9], is the asymmetric Laplace distribution (ALD). The application of the ALD in the quantile regression has become very popular, see, for example, [61], and [1], among others.

When the model contains many predictors, variable selection plays an important role in the model building process to obtain a better interpretation and to improve the precision of model fit. The problem of variable selection is equivalent to identifying an appropriate subset of important variables via the regression coefficients. Over the years, numerous procedures have been developed for variable selection in quantile regression models ranging from the frequentist approaches to the Bayesian procedures. Frequentists usually adopt regularization methods for variable selection in quantile regression by automatically setting several coefficient estimates to zeros, such as the least

absolute deviation (LAD)-Lasso ([55]), SCAD ([58]), adaptive sup-norm regularization ([7]), to name just a few.

Bayesian analysis to the quantile regression begins with specifying a likelihood, which is often obtained from the ALD, due to the relationship between the ALD and quantile regression firstly studied by [29]. This interesting finding motivates [61] to propose Bayesian quantile regression by adopting the ALD as the error distribution in the linear regression model, and they employed a random-walk Metropolis-Hastings algorithm for the posterior sampling. Thereafter, many researchers also started to perform quantile regression from the Bayesian perspective. [54] studied Bayesian quantile regression in a similar treatment as done by [61]. Later on, [60] explored the use of the ALD and proposed stochastic search variable selection (SSVS) for quantile regression models. [1] improved SSVS by introducing an informative prior, the power prior, for the regression coefficients. The resulting method is called the informative stochastic search variable selection (ISSVS). As [41] proposed Bayesian Lasso method based on the combined use of the Lasso regularization method and Bayesian approach in the linear regression model, Bayesian Lasso method and its variants have then been extensively studied in the literature, such as the Bayesian Lasso ([36]), Bayesian adaptive Lasso ([2], [35]), to name just a few.

In this chapter, we study the maximum entropy quantile regression from a Bayesian perspective. Specifically, we consider Bayesian adaptive Lasso on the maximum entropy quantile regression (for short, BMEQR). For prior specification of the unknown model parameters, we consider the commonly used inverse Gamma prior for the scale parameter and a class of regular priors for the penalty parameters as in [36]. In addition, we adopt a flat prior for the quantile parameter whose value is between 0 and 1. Due to the complexity of the joint posterior distribution of the unknown parameters, we propose a hierarchical model based on the mixture representation of the ALD. Therefore, an efficient sampling algorithm based on the combination of the Gibbs sampler and Metropolis-Hastings method is developed for the Bayesian variable selection in the maximum entropy quantile regression. Our simulation studies show that the proposed BMEQR method outperforms the Bayesian Lasso quantile regression (BLQR) ([36]) and the Bayesian adaptive Lasso quantile regression (BALQR) ([2]).

The rest of this chapter is organized as follows. In Section 4.2, we briefly describe the maximum entropy quantile regression model. In Section 4.3, we present the Bayesian adaptive Lasso for the maximum entropy quantile regression along with other two methods, BALQR and

BLQR. In Section 4.4, we carry out simulation studies to examine the performance of the proposed method. A real data example is analyzed in Section 4.5. Finally, some concluding remarks are provided in Section 4.6.

4.2 Maximum entropy quantile regression

Suppose that we have a sample $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)$, where y_i is the response variable, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, for $i = 1, 2, \cdots, n$. The linear quantile regression model is given by

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{4.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent error terms whose distribution is unknown, but is restricted to have the γth quantile equal to zero. The regression coefficients $\boldsymbol{\beta}$ can be estimated as the solution to the following minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}), \tag{4.2}$$

where $\rho_\gamma(t)$ is the check function given by

$$\rho_\gamma(t) = t\{\gamma - I(t < 0)\}. \tag{4.3}$$

Note that the check function is not differentiable at zero, thus we can not derive explicit solutions to the minimization problem (4.2). Therefore, many linear programming methods have been proposed to estimate the quantile regression coefficients ([30]).

If we consider the ALD as the error distribution, the minimization problem (4.2) is equivalent to maximizing the likelihood function, which means we can use the parametric methods to analyze the quantile regression. In addition, the ALD can also be considered as a density that maximizes the Shannon's entropy subject to two moment constraints. The Shannon's entropy for a one dimension random variable Z with probability density function $f(\cdot)$ is given by

$$f_{ME}(z) = \arg\max_f \left\{ -\int f(z) \log f(z) dz \right\}. \tag{4.4}$$

The principle of maximum entropy states that one should consider the one with the largest entropy among all distributions that satisfy certain constraints, because this distribution is believed not to incorporate any extraneous information other than that specified by relevant constraints ([25]).

It is well known that normal distribution maximizes the Shannon's entropy among all continuous distributions with given mean, variance and $(-\infty, \infty)$ support ([44]). Among all continuous random variables with support on $(0, \infty)$ and given mean, the exponential distribution provides the largest entropy ([22]). In addition, the Laplace distribution maximizes the entropy for all continuous distributions with given first absolute moment and $(-\infty, \infty)$ support. In this chapter, we are interested in maximizing the Shannon's entropy with given mean and first absolute moment among all continuous distributions. That is, we desire to maximize the entropy (4.4) subject to

$$E[|Z|] = r_1, \tag{4.5}$$

$$E[Z] = r_2, \tag{4.6}$$

and the normalization constraint, $\int f(z)dz = 1$, where r_1 and r_2 are known constants and $|r_2| < r_1$. It has been shown by [32] that, among all continuous distributions, the one that maximizing the entropy (4.4) under conditions (4.5) and (4.6) is the ALD$(0, \sigma, \gamma)$, whose density is given by

$$f(y \mid \sigma, \gamma) = \frac{\gamma(1-\gamma)}{\sigma} \exp\left\{-\frac{\rho_\gamma(y)}{\sigma}\right\}, \ -\infty < y < \infty,$$

where $\sigma > 0$ is the scale parameter and $\gamma \in (0, 1)$ is the skewness parameter which are respectively given by

$$\begin{aligned} \sigma &= \frac{1}{2}\sqrt{r_1^2 - r_2^2}, \\ \gamma &= \frac{\sqrt{r_1 - r_2}}{\sqrt{r_1 - r_2} + \sqrt{r_1 + r_2}}. \end{aligned}$$

It is easy to check that the mode is 0. Note that γ is equal to 0.5 if r_2 equals 0, then the ALD becomes the standard Laplace distribution. Figure 4.1 provides the graph of γ as a function of r_2 when r_1 is equal to 10.

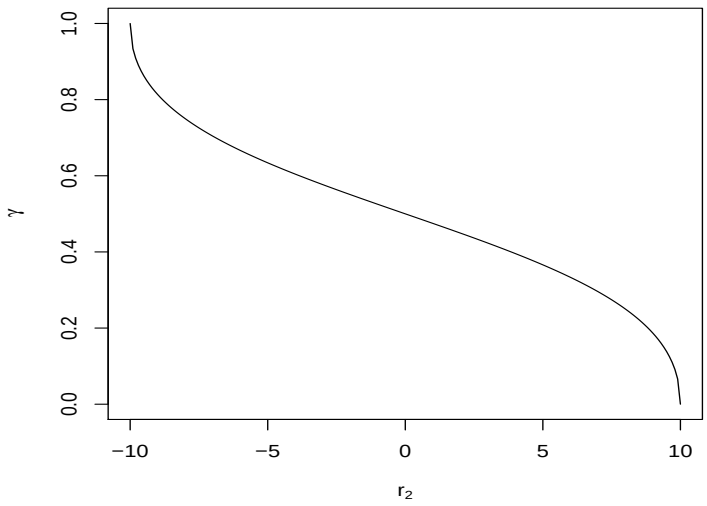

Figure 4.1: Graph of $\gamma(r_2)$ when $r_1 = 10$.

Note that when r_2 is less than 0, γ is greater than 0.5 and the condition (4.6) indicates the mean is less than the mode, so the curve of the ALD is skewed to the right. Similarly, γ is less than 0.5 when r_2 is greater than 0, and the mean is greater than the mode which indicates left skewness in the ALD.

4.3 Bayesian adaptive Lasso on the maximum entropy quantile regression

4.3.1 Bayesian adaptive Lasso

When the model contains many predictors, variable selection plays an important role in the model building process to obtain a better interpretation and to improve the precision of model fit. The problem of variable selection is equivalent to identifying an appropriate subset of important variables via the regression coefficients. Over the years, numerous procedures have been developed for the variable selection in the quantile regression, such as the least absolute deviation (LAD)-Lasso ([55]), SCAD ([58]), adaptive sup-norm regularization ([7]), stochastic search variable selection ([60]), and so on. Of particular note is that, the Lasso which firstly proposed by [53] for the linear regression

models, has been shown is very effective in quantile regression in improving the prediction accuracy, see [37], [58]. The Bayesian Lasso, introduced to regression models by [41], has been studied in the literature. For example, [36] and [2] considered variable selection in quantile regression models by adopting Bayesian Lasso and Bayesian adaptive Lasso, respectively.

In particular, the Lasso estimates are defined by ([53])

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{4.7}$$

where $\lambda \geq 0$. Note that the ordinary least squares (OLS) estimates are obtained by minimizing the first term in (4.7). Compared with OLS estimates, Lasso has several advantages. First of all, with a large number of predictors, Lasso could be used to determine a smaller subset that exhibits strongest effects. Besides, the OLS estimates often have low bias but large variance, Lasso reduces the variance of the predicted values by sacrificing a little bias and hence improve the overall prediction accuracy.

Consider the quantile regression model (4.1), as stated by ([37]), the regression coefficients of Lasso regularized quantile regression can be obtained by

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right\}, \tag{4.8}$$

where λ is a nonnegative regularization parameter. The second term in (4.8) is the Lasso penalty which is crucial for the success of the Lasso method. The Lasso shrinks quantile regression coefficients towards zeros as λ increases.

From a Bayesian point of view, it is essential to derive priors for the regression coefficients and parameters. [36] exploited a Laplace prior on β_j, $\pi(\beta_j|\sigma, \lambda) = \lambda/2\sigma \exp\{-\lambda|\beta_j|/\sigma\}$, for $j = 1, 2, \cdots, p$, and assumed that the error terms ε_i's follow the ALD. [2] extended this idea by placing different penalty parameters on the different regression coefficients, that is, the prior on β_j is $\pi(\beta_j|\sigma, \lambda_j) = \lambda_j/2\sigma \exp\{-\lambda_j|\beta_j|/\sigma\}$. This leads to the adaptive Lasso quantile regression with the regression coefficients obtained by

$$\arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}. \tag{4.9}$$

As discussed before, we consider the quantile level γ as an unknown parameter instead of

fixing it. Since there is no information about γ, we place a flat prior on γ, $\pi(\gamma) = 1$ for $\gamma \in (0, 1)$. We adopt the Laplace priors in [2] for the regression coefficients $\boldsymbol{\beta}$. Specifically,

$$\pi(\beta_j \mid \sigma, \lambda_j) = \frac{\lambda_j}{2\sigma} \exp\left\{-\frac{\lambda_j|\beta_j|}{\sigma}\right\},\ j = 1, 2, \cdots, p.$$

Let $\eta_j = \lambda_j/\sigma$. By adopting the mixture representation of the Laplace distribution proposed by [4], we have

$$\pi(\beta_j \mid \eta_j) = \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\} \frac{\eta_j^2}{2} \exp\left\{-\frac{\eta_j^2}{2} s_j\right\} ds_j.$$

This motivates us to consider a Gamma prior on η_j^2. Moreover, we put an inverse Gamma prior on σ due to the conjugacy. Therefore, the prior on $(\sigma, \boldsymbol{\eta})$ is

$$\pi(\sigma, \boldsymbol{\eta}) = \frac{b^a}{\Gamma(a)} \sigma^{-a-1} \exp\left\{-\frac{b}{\sigma}\right\} \prod_{j=1}^{p} \frac{d^c}{\Gamma(c)} (\eta_j^2)^{c-1} \exp\{-d\eta_j^2\}, \tag{4.10}$$

where $a, b, c, d \geq 0$, and $\boldsymbol{\eta} = (\eta_1^2, \eta_2^2, \cdots, \eta_p^2)$. Note that the priors on σ and $\boldsymbol{\eta}$ become the noninformative priors if we set $a = b = c = d = 0$.

4.3.2 Maximum entropy quantile regression with Bayesian adaptive Lasso

Assume that the error terms ε_i's in (4.1) are i.i.d random variables from the ALD$(0, \sigma, \gamma)$, thus the response variable y_i follows ALD$(\mathbf{x}_i\boldsymbol{\beta}, \sigma, \gamma)$. Denote $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)^T$, the likelihood function of $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ has the form

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma, \gamma) = \frac{\gamma^n(1-\gamma)^n}{\sigma^n} \exp\left\{-\frac{\sum_{i=1}^n \rho_\gamma(y_i - \mathbf{x}_i^T\boldsymbol{\beta})}{\sigma}\right\}. \tag{4.11}$$

Note that the minimization problem in (4.2) is equivalent to maximizing the likelihood function in (4.11) with respect to $\boldsymbol{\beta}$. Such relationship has been exploited in the literature, for example, [61] showed that Bayesian approach based on this relationship is efficient and useful. [21] used the quantile regression to analyze longitudinal data by applying the ALD to the response variable. The estimates of the regression coefficients based on this approach are more efficient than the ones from other methods. A Gibbs sampler method for Bayesian analysis of quantile regression models with

the ALD was proposed by [34].

A good property of the ALD is that it has various mixture representations. For example, the ALD can be written as a scale mixture of normals with the scale mixing parameter following an exponential distribution ([56]); [57] discussed the ALD by writing it as a scale mixture of uniform with the scale mixing parameter following a Gamma distribution. Here, we consider writing the ALD as a mixture of normals, in particular, we have the following lemma

Lemma 3 *Suppose that U is an asymmetric Laplace random variable, and V is an exponential random variable with rate $1/\sigma$. Let*

$$\phi_1 = \frac{1-2\gamma}{\gamma(1-\gamma)} \quad \textit{and} \quad \phi_2 = \frac{2}{\gamma(1-\gamma)}.$$

It follows that

$$U \mid V = v \sim \mathrm{N}\big(\mu + \phi_1 v, \phi_2 \sigma v\big).$$

Note that, based on Lemma 3, the response variable y_i follows a normal distribution with mean $\mathbf{x}_i^T\boldsymbol{\beta} + \phi_1 v_i$, and variance $\phi_2\sigma v_i$, given that $v_i \sim \exp(1/\sigma)$. Denote $\mathbf{v} = (v_1, v_2, \cdots, v_n)$, and view $\mathbf{v}$ as an unknown parameter. The likelihood function of $\mathbf{y}$ becomes

$$f(\mathbf{y} \mid \boldsymbol{\beta}, \sigma, \gamma, \mathbf{v}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\phi_2\sigma v_i}} \exp\left\{-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \phi_1 v_i)^2}{2\phi_2\sigma v_i}\right\}. \tag{4.12}$$

The integration of (4.12) with respect to $\mathbf{v}$ after multiplying the prior of $\mathbf{v}$ provides the same expression in (4.11). The equation in (4.12) simplifies the original likelihood function, and provides an easy way to construct a Gibbs sampler algorithm for the posterior simulation.

Let $\mathbf{s} = (s_1, s_2, \cdots, s_p)$, with the combination of the likelihood function in (4.12) and the specified priors, we obtain the posterior distribution for $(\boldsymbol{\beta}, \sigma, \gamma, \boldsymbol{\eta}, \mathbf{s}, \mathbf{v})$

$$\begin{aligned}
\pi(\boldsymbol{\beta},\sigma,\gamma,\boldsymbol{\eta},\mathbf{s},\mathbf{v}\mid\mathbf{y}) &\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\phi_2\sigma v_i}}\exp\left\{-\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\phi_1 v_i)^2}{2\phi_2\sigma v_i}\right\}\\
&\times \prod_{j=1}^{p}\frac{1}{\sqrt{2\pi s_j}}\exp\left\{-\frac{\beta_j^2}{2s_j}\right\}\frac{\eta_j^2}{2}\exp\left\{-\frac{\eta_j^2}{2}s_j\right\}\\
&\times \sigma^{-a-1}\exp\left\{-\frac{b}{\sigma}\right\}\prod_{j=1}^{p}(\eta_j^2)^{c-1}\exp\{-d\eta_j^2\}\prod_{i=1}^{n}\frac{1}{\sigma}\exp\left\{-\frac{v_i}{\sigma}\right\}. \quad (4.13)
\end{aligned}$$

This yields the following full conditional distributions

$$\begin{aligned}
v_k\mid\boldsymbol{\beta},\sigma,\gamma,\boldsymbol{\eta},\mathbf{s},\mathbf{y} &\propto \frac{1}{\sqrt{v_k}}\exp\left\{-\frac{(y_k-\mathbf{x}_k^T\boldsymbol{\beta}-\phi_1 v_k)^2}{2\phi_2\sigma v_k}\right\}\exp\left\{-\frac{v_k}{\sigma}\right\},\quad k=1,2,\cdots,n;\\
\beta_j\mid\sigma,\gamma,\boldsymbol{\eta},\mathbf{s},\mathbf{v},\mathbf{y} &\propto \prod_{i=1}^{n}\exp\left\{-\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\phi_1 v_i)^2}{2\phi_2\sigma v_i}\right\}\exp\left\{-\frac{\beta_j^2}{2s_j}\right\},\quad j=1,2,\cdots,p;\\
s_j\mid\boldsymbol{\beta},\sigma,\gamma,\boldsymbol{\eta},\mathbf{v},\mathbf{y} &\propto \frac{1}{\sqrt{s_j}}\exp\left\{-\frac{\beta_j^2}{2s_j}\right\}\exp\left\{-\frac{\eta_j^2}{2}s_j\right\};\\
\eta_j^2\mid\boldsymbol{\beta},\sigma,\gamma,\mathbf{s},\mathbf{v},\mathbf{y} &\propto \frac{\eta_j^2}{2}\exp\left\{-\frac{\eta_j^2}{2}s_j\right\}(\eta_j^2)^{c-1}\exp\{-d\eta_j^2\};\\
\sigma\mid\boldsymbol{\beta},\gamma,\boldsymbol{\eta},\mathbf{s},\mathbf{v},\mathbf{y} &\propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\phi_2\sigma v_i}}\exp\left\{-\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\phi_1 v_i)^2}{2\phi_2\sigma v_i}-\frac{v_i}{\sigma}\right\}\frac{1}{\sigma^{n+a+1}}\exp\left\{-\frac{b}{\sigma}\right\};\\
\gamma\mid\boldsymbol{\beta},\sigma,\boldsymbol{\eta},\mathbf{s},\mathbf{v},\mathbf{y} &\propto \gamma^{\frac{n}{2}}(1-\gamma)^{\frac{n}{2}}\exp\left\{-\sum_{i=1}^{n}\frac{(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\phi_1 v_i)^2}{2\phi_2\sigma v_i}\right\}.
\end{aligned}$$

It's thus easy to construct the following efficient Gibbs sampler algorithm for the posterior simulation:

(i) Simulate v_k^{-1} from the inverse Gaussian distribution, $\mathrm{IG}(a_k,b_k)$ with

$$a_k=\sqrt{\frac{\phi_1^2+2\phi_2}{(y_k-\mathbf{x}_k^T\boldsymbol{\beta})^2}}\quad\text{and}\quad b_k=\frac{\phi_1^2+2\phi_2}{\phi_2\sigma},$$

and the density of $\mathrm{IG}(a,b)$ is given by

$$f(x\mid a,b)=\sqrt{\frac{b}{2\pi}}x^{-\frac{3}{2}}\exp\left\{-\frac{b(x-a)^2}{2a^2x}\right\},\ x>0.$$

(ii) Simulate β_j from the normal distribution, $\mathrm{N}(\mu_j, \sigma_j^2)$ with

$$\mu_j = \sum_{i=1}^{n} \frac{(y_i - \phi_1 v_i - \sum_{t \neq j} x_{it}\beta_t)x_{ij}}{\phi_2 \sigma v_i}\sigma_j^2 \quad \text{and} \quad \sigma_j^2 = \left(\frac{1}{s_j} + \sum_{i=1}^{n} \frac{x_{ij}^2}{\phi_2 \sigma v_i}\right)^{-1}.$$

(iii) Simulate s_j^{-1} from the inverse Gaussian distribution, $\mathrm{IG}(c_j, d_j)$ with

$$c_j = \frac{\sqrt{\eta_j^2}}{|\beta_j|} \quad \text{and} \quad d_j = \eta_j^2.$$

(iv) Simulate η_j^2 from the Gamma distribution, $\mathrm{Gamma}(c+1, s_j/2 + d)$.

(v) Simulate σ from the inverse Gamma distribution, $\text{Inverse-Gamma}(\hat{\alpha}, \hat{\theta})$ with

$$\hat{\alpha} = \frac{3}{2}n + a \quad \text{and} \quad \hat{\theta} = b + \sum_{i=1}^{n}\left(\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \phi_1 v_i)^2}{2\phi_2 v_i} + v_i\right).$$

(vi) Simulate γ from its full conditional posterior distribution

$$\pi(\gamma \mid \boldsymbol{\beta}, \sigma, \mathbf{v}, \mathbf{y}) \propto \gamma^{\frac{n}{2}}(1-\gamma)^{\frac{n}{2}} \exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \phi_1 v_i)^2}{2\phi_2 \sigma v_i}\right\}. \tag{4.14}$$

Although the full conditional posterior distribution of γ is not of standard form, we can employ the Metropolis-Hastings method ([38]) for the posterior simulation from (4.14). Our simulation study in the next section shows that the proposed Gibbs sampler algorithm is quite efficient.

4.4 Simulation study

In this section, we study the performance of the BMEQR with comparison to the BALQR and BLQR. The simulation setup is similar to the one in [37]. The data are generated from the model (4.1). We consider the following three different settings for $\boldsymbol{\beta}$:

- Simulation 1: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, which corresponds to the sparse case.
- Simulation 2: $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^T$, which corresponds to the dense case.
- Simulation 3: $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)^T$, which corresponds to the very sparse case.

In each simulation study, the rows of $\mathbf{X}$ follow a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \Sigma)$ with the (i, j)th element of Σ being $0.5^{|i-j|}$. As there are criticisms that placing a specific error distribution is departing from the semiparametric nature of the quantile regression because quantile regression treats the error distribution nonparametrically, we choose the following six different error distributions satisfy that the γth quantile of each distribution is 0:

- The first choice is the $\mathrm{ALD}(0, 1, \gamma)$.
- The second choice is the normal distribution, $\mathrm{N}(\mu, 1)$.
- The third choice is a mixture of two normal distributions, $0.1\mathrm{N}(\mu_1, 1)+0.9\mathrm{N}(\mu_2, 9)$, where $\mu_1 = \Phi^{-1}(1-\gamma)$, $\mu_2 = 3\mu_1$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.
- The fourth choice is the Laplace distribution, $\mathrm{Laplace}(\mu, 1)$.
- The fifth choice is a mixture of two Laplace distributions, $0.1\mathrm{Laplace}(\mu_1, 1)+0.9\mathrm{Laplace}(\mu_2, 9)$, where $\mu_1 = \Phi_2^{-1}(1-\gamma)$, $\mu_2 = 3\mu_1$, and $\Phi_2(\cdot)$ is the cumulative distribution function of Laplace random variable.
- The sixth choice is the non-central t distribution with degree freedom of 3, $t_3(\mu)$.

For each error choice, we set $\gamma \in \{0.5, 0.75, 0.95\}$. In order to illustrate that an inappropriate choice of γ would lead to a poor estimation of the model, we fix the quantile in BALQR and BLQR at a value different from the true one. Specially, we fix γ at 0.75 or 0.95 for both BALQR and BLQR if the true value in the model (4.1) is 0.5. In all the three methods, the hyperparameters (a, b, c, d) in prior (4.10) for σ and $\boldsymbol{\eta}$ are set to be 0.1. We generate a training set with 50 observations, and a testing set with 200 observations from each $\boldsymbol{\beta}$ setting, associated with each of the six error distributions and each quantile γ. Since the true model is known, we could calculate the mean absolute deviation (MAD), defined as

$$\mathrm{MAD} = \frac{1}{200}\sum_{i=1}^{200} |\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \mathbf{x}_i^T\boldsymbol{\beta}^{\mathrm{true}}|,$$

where $\hat{\boldsymbol{\beta}}$ is the posterior mean based on the training data set. We repeat each case 150 times, and compute the median of mean absolute deviations (MMAD), which is equal to the median of MADs over 150 simulations.

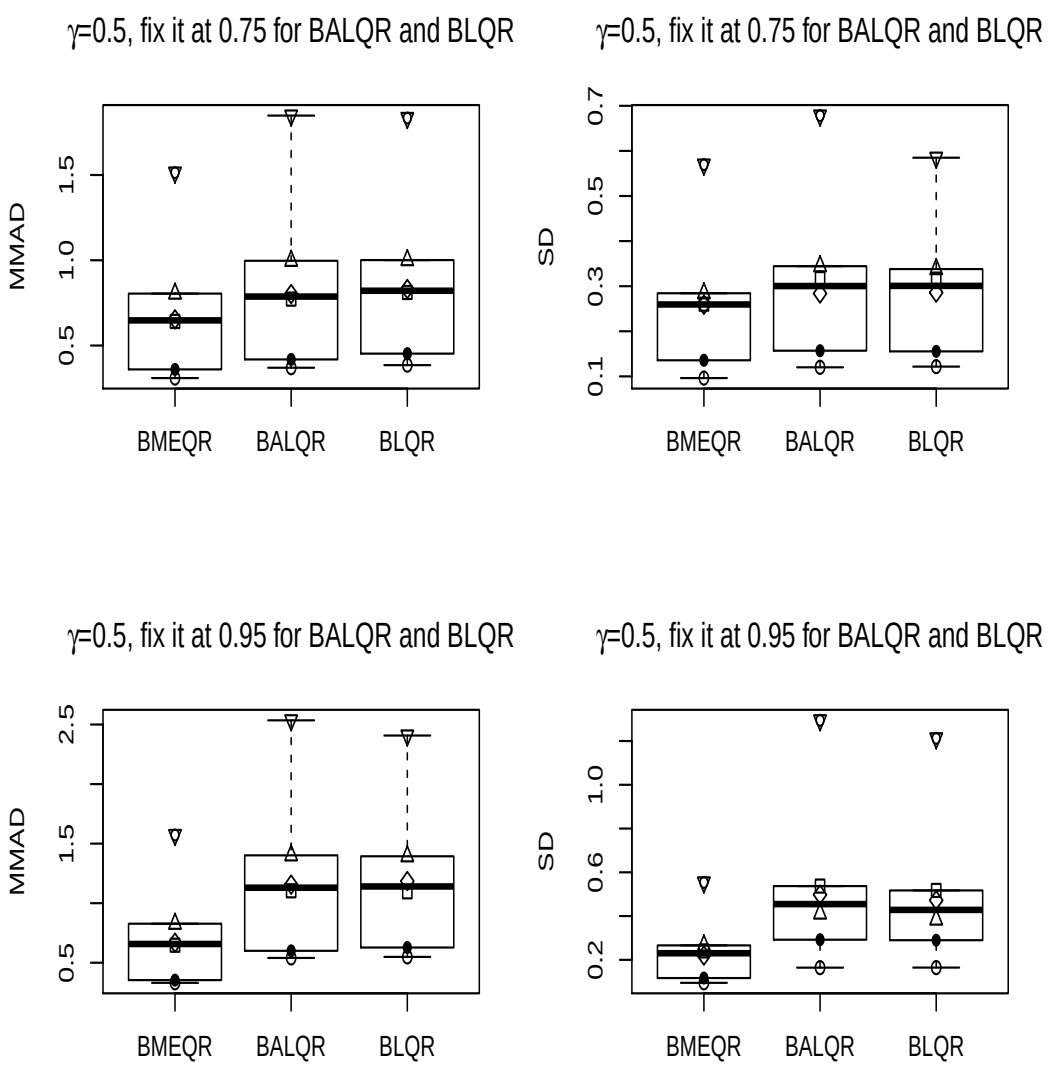

Figure 4.2: Boxplots summarizing the MMADs and the corresponding standard deviations under the three methods for the six error distributions in Simulation 1 when γ is 0.5. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$).

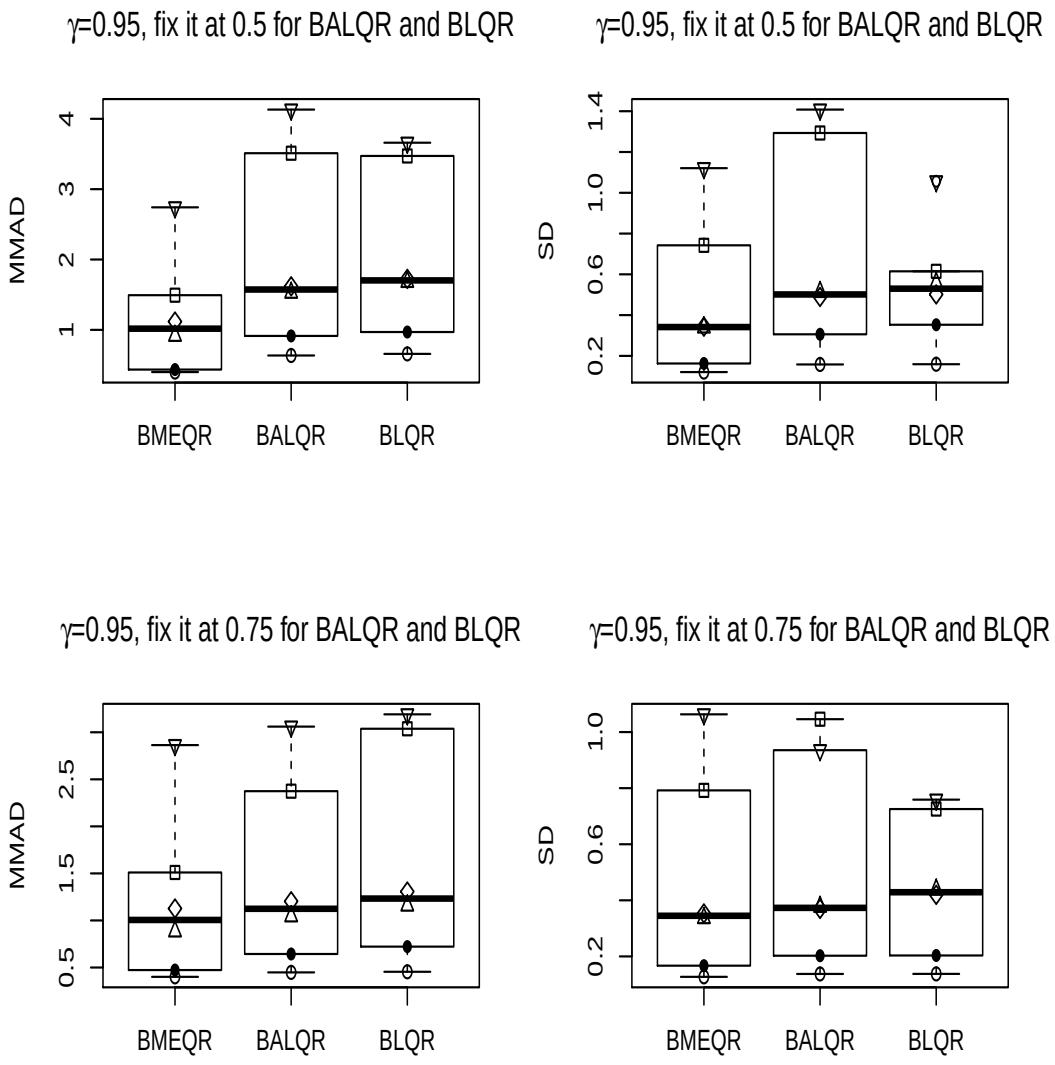

Figure 4.3: Boxplots summarizing the MMADs and the corresponding standard deviations under the three methods for the six error distributions in Simulation 3 when γ is 0.95. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$).

Simulation	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
1	$\boldsymbol{\beta}^{\text{true}}$	3.000	1.500	0.000	0.000	2.000	0.000	0.000	0.000
	BMEQR	2.990	1.436	0.048	0.008	1.934	-0.016	0.030	0.038
		(0.262)	(0.303)	(0.226)	(0.233)	(0.282)	(0.220)	(0.222)	(0.193)
	BALQR	3.064	1.304	0.061	0.011	1.919	-0.030	0.017	0.048
		(0.433)	(0.524)	(0.358)	(0.355)	(0.476)	(0.319)	(0.330)	(0.300)
	BLQR	2.982	1.327	0.082	0.030	1.852	-0.003	0.020	0.059
		(0.418)	(0.489)	(0.372)	(0.374)	(0.464)	(0.336)	(0.357)	(0.317)
2	$\boldsymbol{\beta}^{\text{true}}$	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850
	BMEQR	0.827	0.899	0.783	0.836	0.843	0.889	0.853	0.780
		(0.289)	(0.340)	(0.320)	(0.292)	(0.323)	(0.311)	(0.336)	(0.290)
	BALQR	0.735	0.903	0.738	0.754	0.731	0.843	0.836	0.708
		(0.402)	(0.479)	(0.471)	(0.443)	(0.486)	(0.445)	(0.489)	(0.442)
	BLQR	0.738	0.880	0.748	0.772	0.735	0.835	0.837	0.703
		(0.374)	(0.431)	(0.428)	(0.405)	(0.447)	(0.404)	(0.444)	(0.410)
3	$\boldsymbol{\beta}^{\text{true}}$	5.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	BMEQR	5.006	0.003	-0.012	-0.002	-0.002	-0.008	-0.004	-0.003
		(0.262)	(0.230)	(0.192)	(0.230)	(0.245)	(0.217)	(0.216)	(0.181)
	BALQR	5.020	-0.045	-0.003	0.006	0.003	0.014	-0.016	0.012
		(0.406)	(0.368)	(0.324)	(0.343)	(0.331)	(0.338)	(0.339)	(0.306)
	BLQR	4.894	0.012	0.003	0.012	-0.002	0.017	-0.012	0.022
		(0.419)	(0.386)	(0.339)	(0.364)	(0.346)	(0.354)	(0.355)	(0.317)

Table 4.1: The parameter estimates for the simulated data with normally distributed errors and the corresponding standard deviations in the parenthesis. The true value of γ is 0.95, whereas we set it to be 0.5 for BALQR and BLQR.

For simplicity, we only present the boxplot results in Simulations 1 and 3, because Simulation 2 provides some similar results. Figures 4.2 and 4.3 show, in terms of MMAD, the proposed method (BMEQR) generally performs better than the other two methods (BALQR and BLQR) for all the distributions under consideration, especially for the case in which the fixed γ in BALQR and BLQR is far away from the true value.

Secondly, Table 4.1 summaries the estimates of $\boldsymbol{\beta}$ under the three approaches. Again we only present the result for $\gamma = 0.95$ with normally distributed errors. We choose the posterior mean of $\boldsymbol{\beta}$ in each simulation and report its median of 150 simulations. Note that, the standard deviations of $\boldsymbol{\beta}$ in the proposed method are consistently smaller than those in BALQR and BLQR, which means the results of the BMEQR approach tend to be more accurate.

Moreover, Figures 4.4 and 4.5 display the posterior distributions of $\boldsymbol{\beta}$ for the three methods overplotted with the histogram of simulated $\boldsymbol{\beta}$ in BMEQR. Similar to Table 4.1, we only present the case from Simulation 1 for $\gamma = 0.95$ with normally distributed errors, and we fix it at 0.5 for

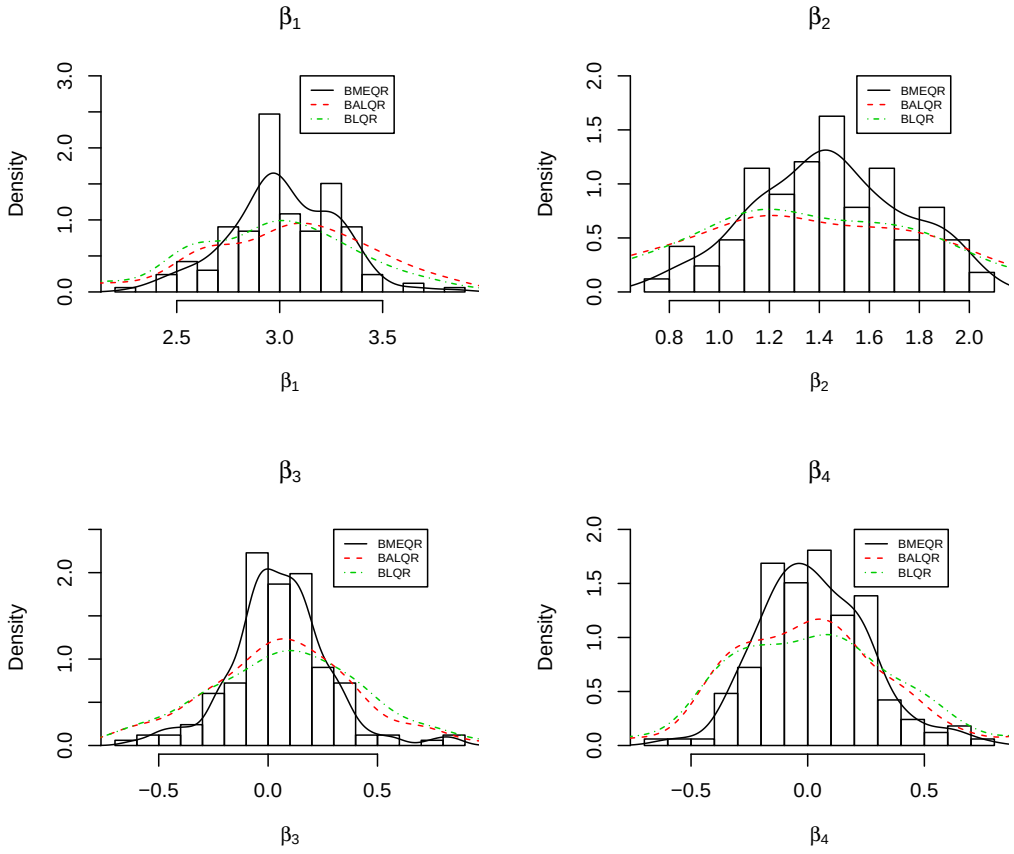

Figure 4.4: Posterior distributions of $(\beta_1, \beta_2, \beta_3, \beta_4)$ overplotted with the histogram of simulated values in BMEQR.

BALQR and BLQR. Note that the posterior distributions of $\boldsymbol{\beta}$ in BMEQR are more concentrated on the true values, which reinforces our conclusions from Table 4.1.

Figures 4.6, 4.7, and 4.8 show the estimated γ and its corresponding standard deviation in BMEQR using the six error distributions in each simulation study. Note that the estimates are close to the true values and that the standard deviations are rather small which indicates that the proposed method estimates the quantile parameter well.

4.5 Real data analysis

In this section, we compare the performance of the three methods BMEQR, BALQR, and BLQR using the prostate cancer data. This data was provided by [51] and analyzed by [53] and [63]

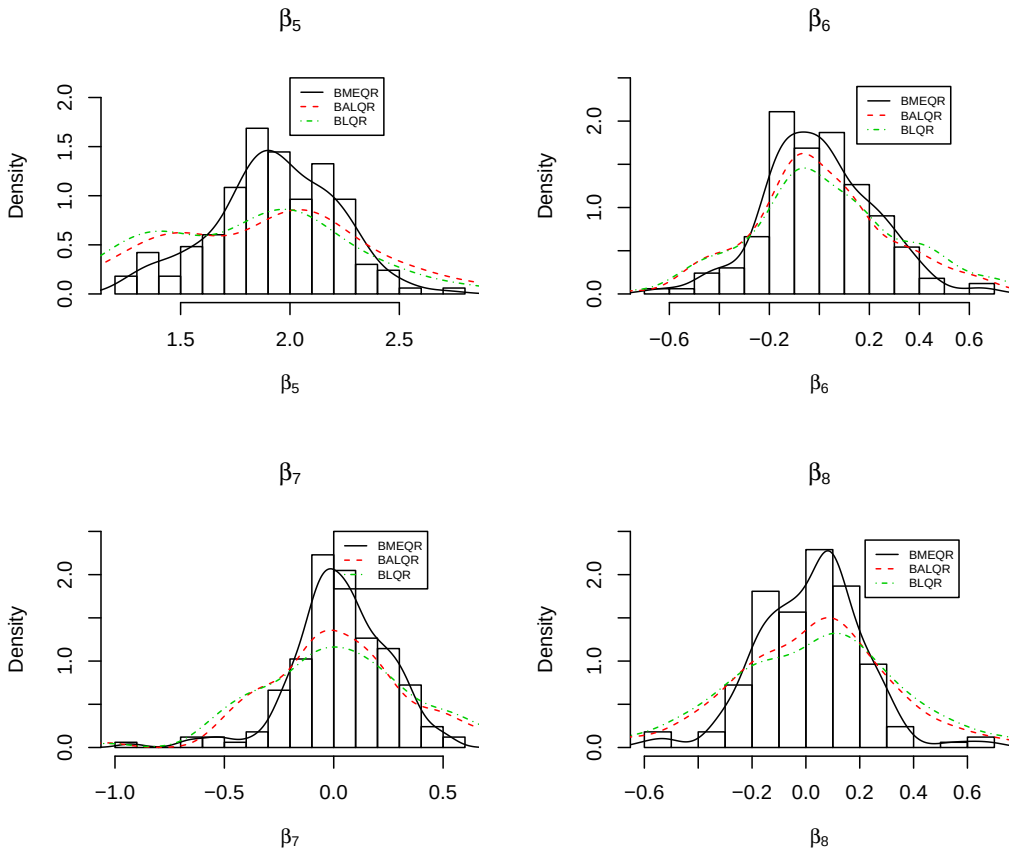

Figure 4.5: Posterior distributions of $(\beta_5, \beta_6, \beta_7, \beta_8)$ overplotted with the histogram of simulated values in BMEQR.

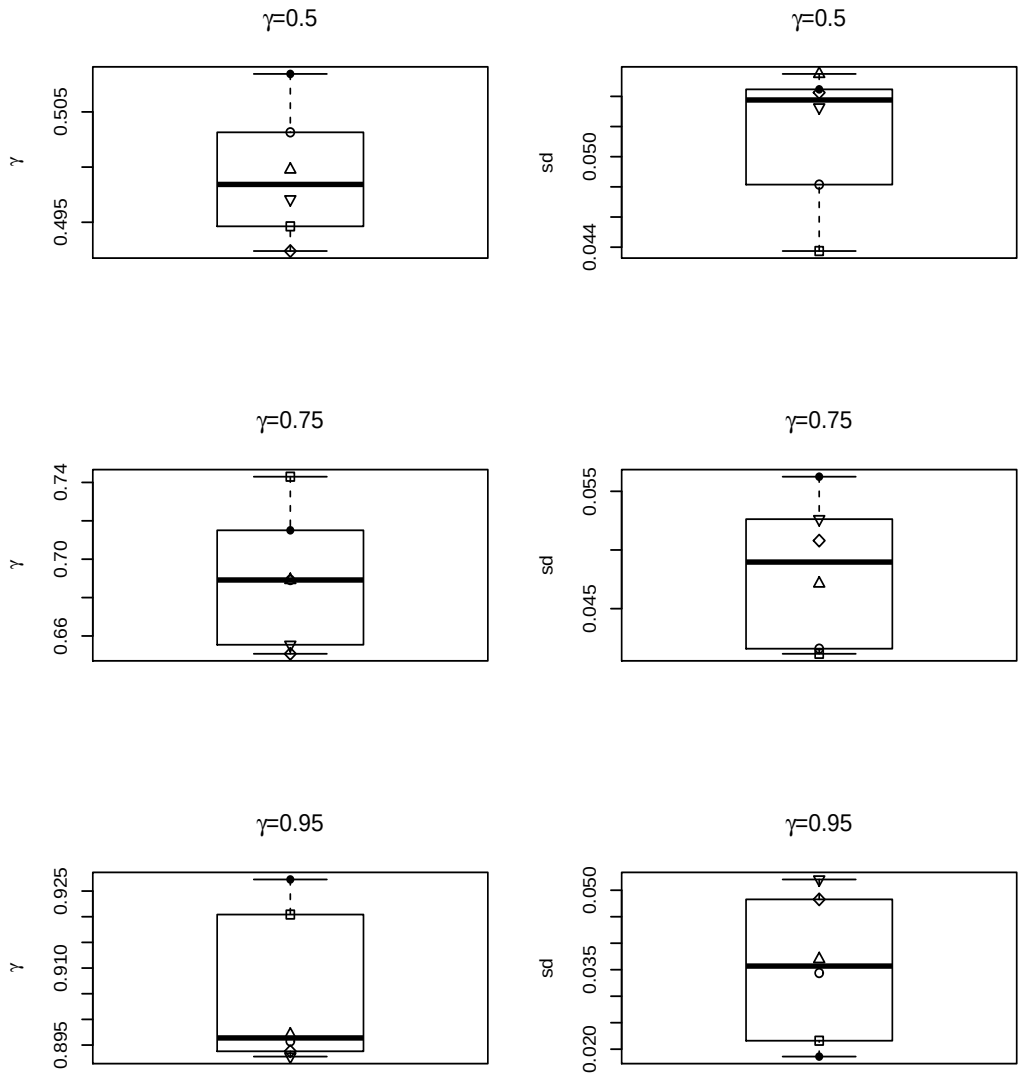

Figure 4.6: Boxplots summarizing simulated γ and the corresponding standard deviation for BMEQR using the six error distributions in Simulation 1. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$).

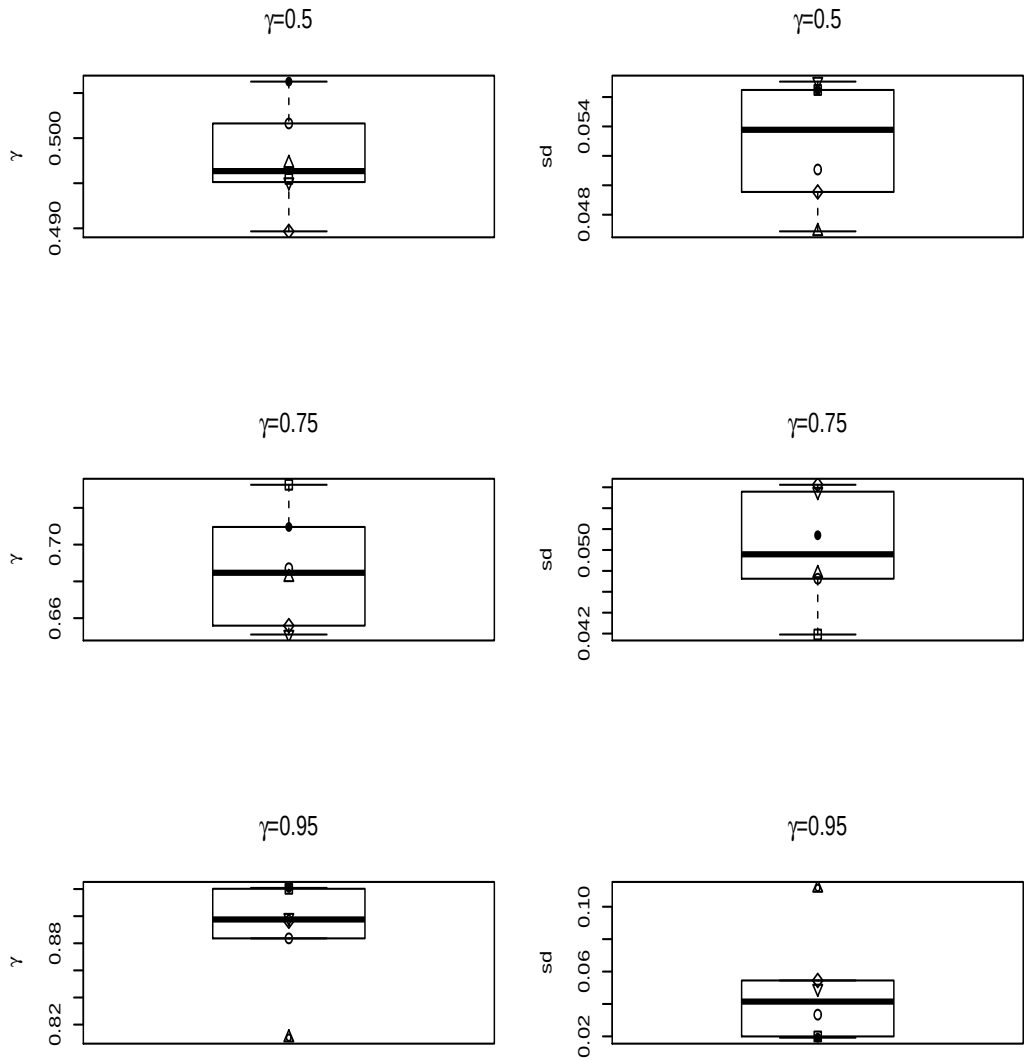

Figure 4.7: Boxplots summarizing simulated γ and the corresponding standard deviation for BMEQR using the six error distributions in Simulation 2. Overlaid are AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$).

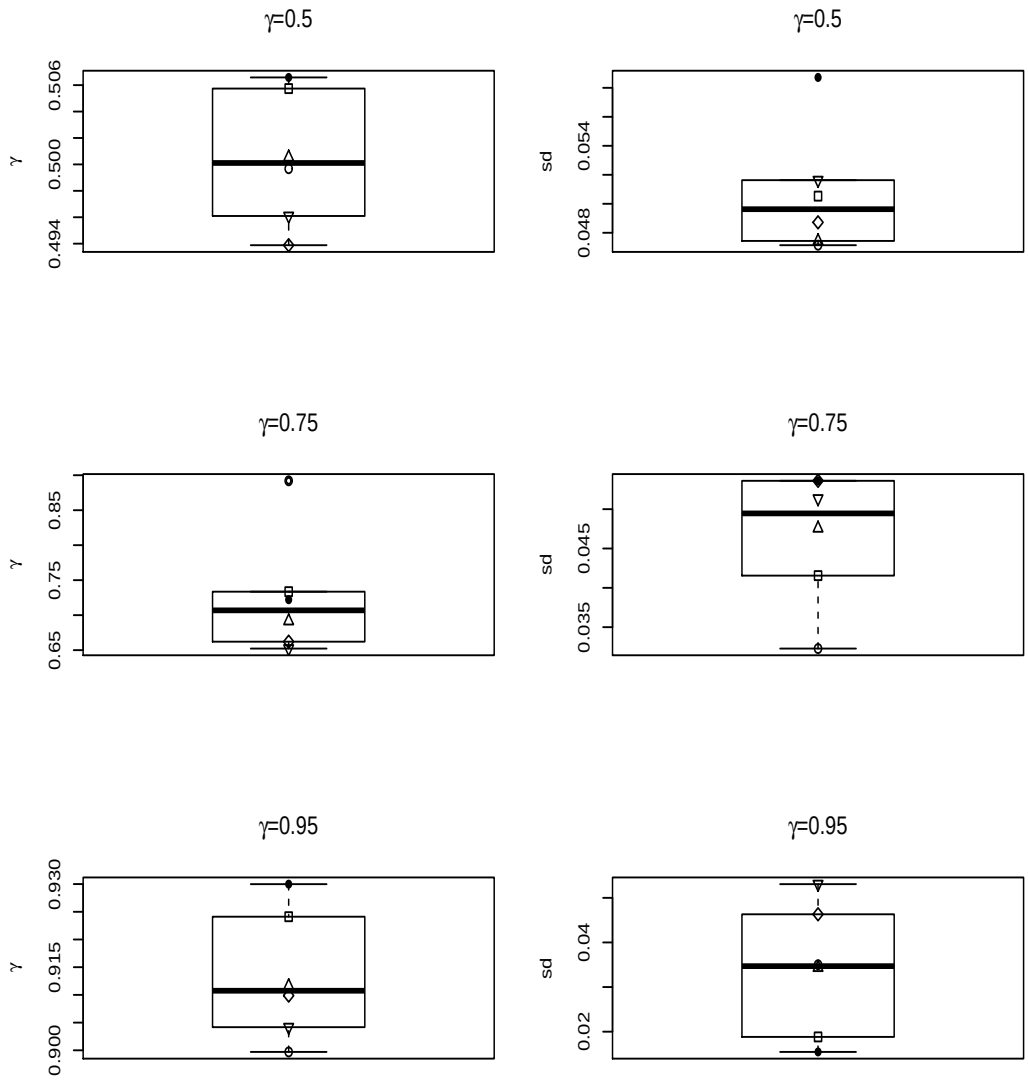

Figure 4.8: Boxplots summarizing simulated γ and corresponding standard deviation for BMEQR using the six error distributions in Simulation 3. Overlaid are the AL ($\square$), normal distribution ($\circ$), normal mixture ($\triangle$), Laplace ($\diamondsuit$), Laplace mixture (∇), and t distribution ($\bullet$).

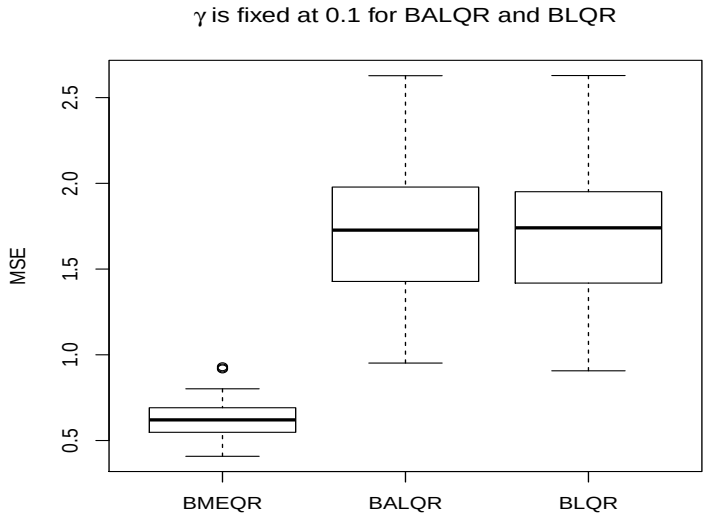

Figure 4.9: Boxplots summarizing the MSE of the three methods when $\gamma = 0.1$.

by using the linear regression model. The data consists of the medical records of 97 male patients who were about to receive a radical prostatectomy. The response variable is the level of prostate antigen (lpsa) and there are eight predictors, log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason) and percentage of Gleason scores 4 or 5 (pgg45).

We fit the data through the maximum entropy quantile regression model. For the BALQR and BLQR, we consider three choices of γ, 0.1, 0.3, and 0.5. The prior specifications are the same as those in Section 4.3, and we set $a = b = c = d = 0.1$ for the joint prior of $(\sigma, \boldsymbol{\eta})$. In each repetition, we randomly split all the 97 observations into training and testing data sets of size 40 and 57, respectively. We choose the mean squared error (MSE) of the testing data as our criterion. The performances over 40 repetitions of the three methods are presented in Figures 4.9, 4.10, and 4.11, respectively. Note that, the proposed method performs better than BALQR and BLQR when $\gamma = 0.1$ and 0.3, and that these three approaches have almost the same results when γ is chosen to be 0.5.

Figure 4.12 displays the estimated γ in BMEQR. We observe that the estimated value of γ is around 0.5 and this explains why the performances of the three methods behave similarly in Figure 4.11.

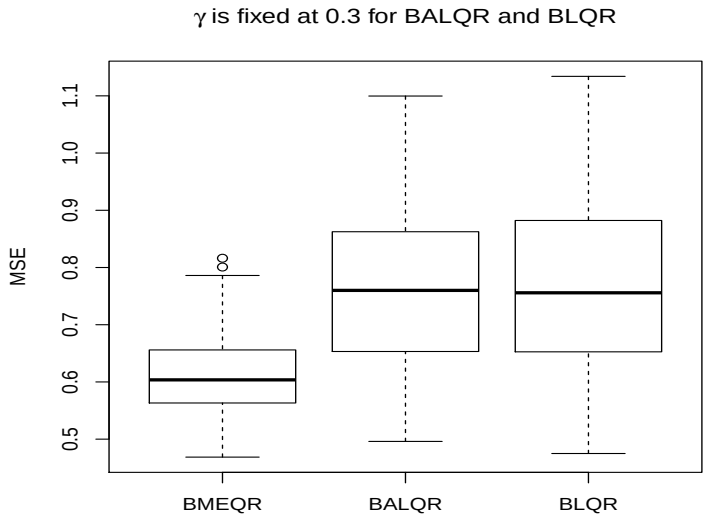

Figure 4.10: Boxplots summarizing the MSE of the three methods when $\gamma = 0.3$.

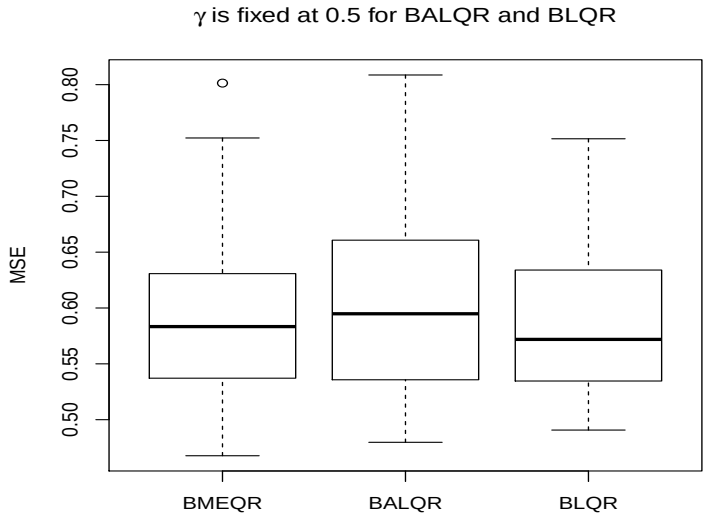

Figure 4.11: Boxplots summarizing the MSE of the three methods when $\gamma = 0.5$.

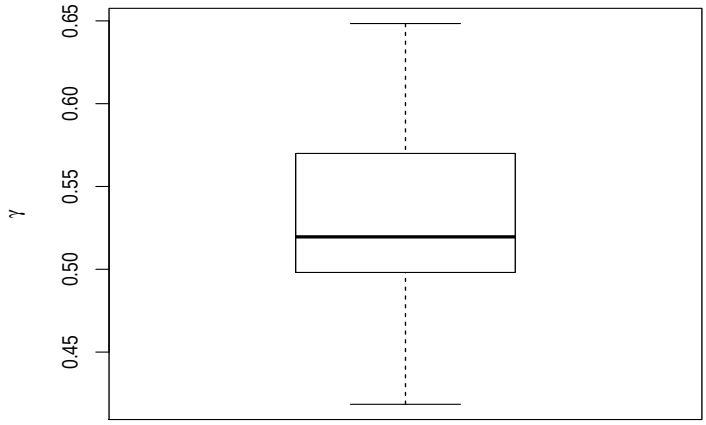

Figure 4.12: Boxplot summarizing the estimated γ in BMEQR.

4.6 Concluding Remarks

In this chapter, we consider Bayesian variable selection in the maximum entropy quantile regression model. Consider a quantile regression, $y = x\beta + \varepsilon$, given that the true model is the 0.95th quantile of y being $x\beta$. The researcher will obtain unsatisfied estimates if one considers the model at the 0.75th quantile. Motivated by this fact, we regard the quantile as an unknown parameter instead of fixing it and estimate it with other parameters jointly. We use the Bayesian adaptive Lasso to achieve variable selection. Since the lack of information for the quantile parameter, we place a flat prior on it. The estimated quantile in the proposed method is the most probable one and reflects the inner relationship of the data. We compare the MMADs and point estimates of the regression coefficients in our approach with the ones under BALQR and BLQR. Numerical evidence shows that the proposed method outperforms the other two methods.

The proposed approach provides an alternative interpretation of quantile regression. It estimates the quantile parameter from the data, which allows the data to speak for itself. In some practical problems, the correlations among the covariates are high, which may need grouped variable selection. It is possible to start with the proposed method. Moreover, it deserves mentioning that the proposed method could be extended to other types of quantile regression model, such as the

binary quantile regression model, and the probit quantile regression model, which are currently under investigation and will be reported elsewhere.

Chapter 5

Future Work

The proposed variable selection method in Chapter 4 provides an alternative interpretation of quantile regression. Instead of fixing quantile at a specified value, we consider it as a random variable and estimate it from the data, which allows the data to speak for itself. It deserves mentioning that the proposed method could be extended to other types of quantile regression model, such as the binary quantile regression model, and the logistic quantile regression model.

5.1 Bayesian binary quantile regression

The binary quantile regression model is given by:

$$\begin{aligned} y_i^* &= \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \\ y_i &= \begin{cases} 1, & y_i^* \geq 0 \\ 0, & y_i^* < 0 \end{cases}, \quad i = 1, 2, \cdots, n, \end{aligned}$$

where y_i^* is the latent unobserved response variable and y_i is the observed response variable, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent error terms whose distribution is unknown, but is restricted to have the γth quantile equal to zero.

The binary quantile regression model has the advantages of the quantile regression model. It is robustness and detailed insights in covariate effects, and overcomes issues related to overfitting

([8]). Thus it has received much attention and been studied by many literature, such as [31], [8].

We consider Bayesian adaptive Lasso on the binary quantile regression and regard quantile as an unknown parameter. The adaptive Lasso estimates of the binary quantile regression are given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \rho_\gamma(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j |\beta_j| \right\}, \tag{5.1}$$

where $\lambda_j > 0$ is the penalty coefficient, $j = 1, 2, \cdots, p$, and $\rho_\gamma(\cdot)$ is the check function that given by $\rho_\gamma(y) = y(\gamma - I(y < 0))$.

Due to the close relationship between the asymmetric Laplace distribution (ALD) and the quantile regression studied by [61], we employ an asymmetric Laplace error distribution $\text{ALD}(0, \sigma, \gamma)$ for ε_i, thus y_i^* follows $\text{ALD}(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma, \gamma)$. The density function of the ALD is given by

$$f(x \mid \mu, \sigma, \gamma) = \frac{\gamma(1-\gamma)}{\sigma} \exp\left\{ -\frac{\rho_\gamma(x-\mu)}{\sigma} \right\}, \quad -\infty < x < \infty.$$

We put the Laplace priors on the regression coefficients $\boldsymbol{\beta}$,

$$\pi(\beta_j \mid \sigma, \lambda_j) = \frac{\lambda_j}{2\sigma} \exp\left\{ -\frac{\lambda_j |\beta_j|}{\sigma} \right\}, \ j = 1, 2, \cdots, p.$$

Let $\eta_j = \lambda_j / \sigma$. We put an inverse Gamma prior on σ and a Gamma prior on η_j^2 due to the conjugacy

$$\pi(\sigma, \boldsymbol{\eta}) = \frac{b^a}{\Gamma(a)} \sigma^{-a-1} \exp\left\{ -\frac{b}{\sigma} \right\} \prod_{j=1}^{p} \frac{d^c}{\Gamma(c)} (\eta_j^2)^{c-1} \exp\{-d\eta_j^2\},$$

where $a, b, , c, d \geq 0$, and $\boldsymbol{\eta} = (\eta_1^2, \eta_2^2, \cdots, \eta_p^2)$.

Since there is no information about the quantile γ, a uniform prior within the interval $(0, 1)$ is considered as the prior for γ. As stated by [34], the ALD can be written as a scale mixture of normals with the scale mixing parameter following the exponential distribution. We use such mixture representations for the priors on $\boldsymbol{\beta}$ and the likelihood function of $\mathbf{y} = (y_1, y_2, \cdots, y_n)$.

Thus a Bayesian hierarchical model for the binary quantile regression is established. A Gibbs sampler algorithm can be derived from it.

5.2 Bayesian logistic quantile regression

The logistic quantile regression is often used to model data within a known range, when the traditional statistics methods, such as least-squares regression, mixed-effects model do not perform well. As proposed by [12], the logistic quantile regression is given by

$$\log\left(\frac{y_i - y_{min}}{y_{max} - y_i}\right) = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

where y_i's are the observed response variables and bounded from below and from above by two constants y_{min} and y_{max}, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ represents the p known covariates, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of the regression coefficients, and ε_i's are independent error terms whose distribution is unknown, but is restricted to have the γth quantile equal to zero.

We consider variable selection on the logistic quantile regression through Bayesian adaptive Lasso. The error distribution is chosen to be ALD$(0, \sigma, \gamma)$ due to the close relationship between the ALD and quantile regression. The adaptive Lasso estimates of the logistic quantile regression are given by (5.1). We can use the same prior selections as in Section 5.1.

In particular, we consider independent Laplace priors for the regression coefficients, thus the adaptive penalty term could be interpreted as a Bayesian posterior mode. Let $\eta_j = \lambda_j/\sigma$, due to the mixture representation of the Laplace distribution, we have

$$\pi(\beta_j \mid \eta_j) = \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\} \frac{\eta_j^2}{2} \exp\left\{-\frac{\eta_j^2}{2} s_j\right\} ds_j,$$

where $\mathbf{s} = (s_1, s_2, \cdots, s_p)$ is a latent vector and s_j follows standard normal distribution, for $j = 1, 2, \cdots, p$.

We put an inverse Gamma prior on σ and a Gamma prior on η_j^2 due to the conjugacy. Due to the lack of prior information on γ, we put a uniform prior within $(0, 1)$ on it. However, in case which we have prior information about γ, a Beta prior with known shape parameters can also be considered.

Due to the mixture representation of the ALD, the Bayesian logistic quantile regression with

adaptive Lasso penalty is a hierarchical model given by

$$\begin{aligned}
\log\left(\frac{y_i - y_{min}}{y_{max} - y_i}\right) &= \mathbf{x}_i^T\boldsymbol{\beta} + \phi_1 v_i + \sqrt{\phi_2 \sigma v_i} z_i, \\
\mathbf{v} &\sim \prod_{i=1}^{n} \frac{1}{\sigma} \exp\left\{-\frac{v_i}{\sigma}\right\}, \\
\mathbf{z} &\sim \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z_i^2}{2}\right\}, \\
\boldsymbol{\beta} \mid \mathbf{s} &\sim \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi s_j}} \exp\left\{-\frac{\beta_j^2}{2s_j}\right\}, \\
\mathbf{s} \mid \boldsymbol{\eta} &\sim \prod_{j=1}^{p} \frac{\eta_j^2}{2} \exp\left\{-\frac{\eta_j^2}{2} s_j\right\}, \\
\sigma &\sim \frac{b^a}{\Gamma(a)} \sigma^{-a-1} \exp\left\{-\frac{b}{\sigma}\right\}, \\
\boldsymbol{\eta} &\sim \prod_{j=1}^{p} \frac{d^c}{\Gamma(c)} (\eta_j^2)^{c-1} \exp\{-d\eta_j^2\}, \\
\gamma &\sim 1,
\end{aligned}$$

where $a, b, c, d \geq 0$, $\phi_1 = (1-2\gamma)/(\gamma-\gamma^2)$, $\phi_2 = 2/(\gamma-\gamma^2)$, $\mathbf{v} = (v_1, v_2, \cdots, v_n)$, $\mathbf{z} = (z_1, z_2, \cdots, z_n)$, $\boldsymbol{\eta} = (\eta_1^2, \eta_2^2, \cdots, \eta_p^2)$. A Gibbs sampler algorithm can be derived from the hierarchical model.

Bibliography

[1] R. Alhamzawi and K. Yu. Variable selection in quantile regression via Gibbs sampling. *Journal of Applied Statistics*, 39(4):799–813, 2012.

[2] R. Alhamzawi, K. Yu, and D. F. Benoit. Bayesian adaptive Lasso quantile regression. *Statistical Modelling*, 12(3):279–297, 2012.

[3] L. An and P. Tao. Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of Global Optimization*, 11(3):253–285, 1997.

[4] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.

[5] R. B. Arellano-Valle, H. W. Gómez, and F. A. Quintana. Statistical inference for a general class of asymmetric distributions. *Journal of Statistical Planning and Inference*, 128(2):427–443, 2005.

[6] A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.

[7] S. Bang and M. Jhun. Simultaneous estimation and factor selection in quantile regression via adaptive sup-norm regularization. *Computational Statistics & Data Analysis*, 56(4):813–826, 2012.

[8] D. F. Benoit, R. Alhamzawi, and K. Yu. Bayesian Lasso binary quantile regression. *Computational Statistics*, 28(6):2861–2873, 2013.

[9] A. K. Bera, A. F. Galvao Jr, G. Montes-Rojas, and S. Y. Park. Which quantile is the most informative? Maximum likelihood, maximum entropy and quantile regression. *Working paper, Department of Economics, City University London*, 2010.

[10] J. O. Berger and J. M. Bernardo. On the development of reference priors. *Bayesian Statistics*, 4(4):35–60, 1992.

[11] J. O. Berger, V. De Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.

[12] M. Bottai, B. Cai, and R. E. McKeown. Logistic quantile regression for bounded outcomes. *Statistics in medicine*, 29(2):309–317, 2010.

[13] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[14] L. Demortier, S. Jain, and H. B. Prosper. Reference priors for high energy physics. *Physical Review D Particles & Fields*, 82(3):346–361, 2010.

[15] Y. Duan and K. Ye. *Normalized power prior Bayesian analysis*. UTSA, College of Business, 2008.

[16] D. B. Dunson. Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, 153(12):1222–1226, 2001.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[18] C. Fernández and M. F. Steel. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371, 1998.

[19] M. A. Ferreira and M. A. Suchard. Bayesian analysis of elapsed times in continuous-time Markov chains. *Canadian Journal of Statistics*, 36(3):355–368, 2008.

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

[21] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

[22] D. V. Gokhale. Maximum entropy characterizations of some distributions. *In A Modern Course on Statistical Distributions in Scientific Work*, 17:299–304, 1975.

[23] W. Hendricks and R. Koenker. Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, 87(417):58–68, 1992.

[24] J. G. Ibrahim and M. Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

[25] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620–630, 1957.

[26] M. Jones. On families of distributions with shape parameters. *International Statistical Review*, 83(2):175–192, 2015.

[27] R. Koenker and G. Bassett Jr. Regression quantiles. *Journal of the Econometric Society*, 46(1):33–50, 1978.

[28] R. Koenker and O. Geling. Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468, 2001.

[29] R. Koenker and J. A. F. Machado. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448):1296–1310, 1999.

[30] R. Koenker and B. J. Park. An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283, 1996.

[31] G. Kordas. Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407, 2006.

[32] S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Number 183. Springer, 2001.

[33] T. J. Kozubowski and K. Podgórski. Asymmetric Laplace laws and modeling financial data. *Mathematical and Computer Modelling*, 34(9):1003–1021, 2001.

[34] H. Kozumi and G. Kobayashi. Gibbs sampling methods for Bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578, 2011.

[35] C. Leng, M. Tran, and D. Nott. Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, 66(2):221–244, 2014.

[36] Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556, 2010.

[37] Y. Li and J. Zhu. L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163–185, 2008.

[38] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[39] G. S. Mudholkar and A. D. Hutson. The epsilon–skew–normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference*, 83(2):291–309, 2000.

[40] S. Nadarajah and S. Kotz. Skewed distributions generated by the normal kernel. *Statistics & Probability Letters*, 65(3):269–277, 2003.

[41] T. Park and G. Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[42] E. Purdom, S. P. Holmes, et al. Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1070–1105, 2005.

[43] A. E. Raftery and S. M. Lewis. One long run with diagnostics: implementation strategies for Markov chain Monte Carlo. *Statistical Science*, 7(1):493–497, 1992.

[44] F. M. Reza. *An Introduction to Information Theory*. Courier Corporation, 1961.

[45] G. O. Roberts. Markov chain concepts related to sampling algorithms. *In Markov chain Monte Carlo in practice*, pages 45–57, 1996.

[46] F. J. Rubio and M. F. Steel. Inference for grouped data with a truncated skew-Laplace distribution. *Computational Statistics & Data Analysis*, 55(12):3218–3231, 2011.

[47] F. J. Rubio and M. F. Steel. Inference in two-piece location-scale models with Jeffreys priors. *Bayesian Analysis*, 9(1):1–22, 2014.

[48] E. Salazar, M. A. Ferreira, and H. S. Migon. Objective Bayesian analysis for exponential power regression models. *Sankhya B*, 74(1):107–125, 2012.

[49] F. M. Scherer and D. Harhoff. Technology policy for a world of skew-distributed outcomes. *Research Policy*, 29(4):559–566, 2000.

[50] R. L. Smith and J. Naylor. A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Applied Statistics*, 36(3):358–369, 1987.

[51] T. A. Stamey, J. N. Kabalin, J. E. McNeal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. *The Journal of Urology*, 141(5):1076–1083, 1989.

[52] D. Sun and J. O. Berger. Reference priors with partial information. *Biometrika*, 85(1):55–71, 1998.

[53] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[54] E. G. Tsionas. Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, 73(9):659–674, 2003.

[55] H. Wang, G. Li, and G. Jiang. Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *Journal of Business & Economic Statistics*, 25(3):347–355, 2007.

[56] N. Wichitaksorn, S. Choy, and R. Gerlach. A generalized class of skew distributions and associated robust quantile regression models. *Canadian Journal of Statistics*, 42(4):579–596, 2014.

[57] N. Wichitaksorn, J. J. Wang, S. B. Choy, and R. Gerlach. Analyzing return asymmetry and quantiles through stochastic volatility models using asymmetric Laplace error via uniform scale mixtures. *Applied Stochastic Models in Business & Industry*, 31(5):584–608, 2015.

[58] Y. Wu and Y. Liu. Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817, 2009.

[59] N. Yi and S. Xu. Bayesian Lasso for quantitative trait loci mapping. *Genetics*, 179(2):1045–1055, 2008.

[60] K. Yu, C. W. Chen, C. Reed, and D. Dunson. Bayesian variable selection in quantile regression. *Statistics and Its Interface*, 6:261–274, 2013.

[61] K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447, 2001.

[62] K. Yu and J. Zhang. A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics-Theory and Methods*, 34(9-10):1867–1879, 2005.

[63] M. Yuan and Y. Lin. Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association*, 100(472):1215–1225, 2005.

[64] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[65] X. Zhao, H. Zhang, K. Lai, and S. Wang. A method for evaluating mutual funds performance based on asymmetric Laplace distribution and DEA approach. *Systems Engineering: Theory and Practice*, 27(10):1–10, 2007.

[66] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[67] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.