ABSTRACT

JANG, WOO SUNG. Semiparametric Bayesian Quantile Regression. (Under the direction of Huixia Judy Wang.)

We propose semiparmetric Bayesian quantile regression methods for analyzing independent and clustered data and Bayesian regularized quantile regression methods for quantile and interquantile shrinkage.

Bayesian quantile regression is challenging without making any parametric likelihood assumptions. Instead of posing any parametric distributional assumptions on the random errors, we approximate the central density by linearly interpolating the conditional quantile functions of the response at multiple quantiles and estimate the tail densities by adopting extreme value theory.

In Chapter 2, we propose a semiparametric Bayesian quantile regression method for analyzing independent data based on the proposed approximate likelihood. In Chapter 3, we develop a semiparametric Bayesian quantile regression method for analyzing clustered data, where random effects are included to accommodate the intra-cluster dependence. Through joint-quantile modeling, the proposed methods yield the joint posterior distribution of quantile coefficients at multiple quantiles and meanwhile avoid the quantile crossing issue. Through simulation studies, we demonstrate that the proposed methods lead to more efficient multiple-quantile estimation than existing methods for quantile regression in finite samples.

In Chapter 4, we propose penalization methods for quantile and interquantile shrinkage for linear quantile regression. Our proposed methods shrink quantile slopes and their interquantile differences towards zero in some or the entire quantile region by using fused adaptive LASSO or group LASSO penalties based on Laplace priors. Our numerical investigations show that the proposed methods yield higher estimation efficiency than the conventional quantile regression method when there exists sparsity in the quantile coefficients and/or interquantile differences.

Semiparametric Bayesian Quantile Regression

by
Woo Sung Jang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2014

APPROVED BY:

Sujit Ghosh	Daowen Zhang
Wenbin Lu	Huixia Judy Wang Chair of Advisory Committee

BIOGRAPHY

Woo Sung Jang was born in Seoul, South Korea. He received his B.A. in Statistics from Korea University. After his Master's program in Industrial Engineering at Seoul National University, he worked as a researcher in the center for Defense Modeling&Simulation at Korea Institute for Defense Analyses before joining the Department of Statistics at North Carolina State University in August 2009.

ACKNOWLEDGEMENTS

First I would like to express my first and deepest gratitude to Dr. Huixia Wang for her excellent guidance, patience, financial and emotional support. I could not have imagined having a better mentor for my Ph.D. study.

I am especially thankful to my committee members, Dr. Sujit Ghosh, Dr. Daowen Zhang and Dr. Wenbin Lu for joining my committee. I greatly appreciate constructive feedback they have provided. I also thank Dr. Ratna Sharma-Shivappa for serving on my committee as the graduate school representative.

Most of all, I thank my parents, Minjung, Lia and friends for their love and support.

This research was supported in part by the National Science Foundation Grant DMS-1149355.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Introduction to Quantile Regressioin for Independent Data

1.1.1 Least Squares versus Quantile Regression

The linear regression model is one of the most commonly used tools in modeling the relationship between a response variable and independent variables. The classic linear regression model is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is a $N \times 1$ vector of the response, $\mathbf{X}$ is a $N \times p$ design matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the unknown regression coefficients and $\boldsymbol{\epsilon}$ is a $N \times 1$ vector of errors with zero mean.

The least squares (LS) method is most commonly used to estimate $\boldsymbol{\beta}$ since the LS estimator for $\boldsymbol{\beta}$ has the closed-form solution $\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. In addition, it is the best linear unbiased estimator under the additional assumption: $Var(\boldsymbol{\epsilon}) = \sigma^2 I_N$, where I_N is the $N \times N$ identity matrix, that is, the errors are independent and have a constant variance.

In some applications, the homoscedastic assumption may not hold, or the main focus may be on the tails instead of the center of the response distribution. For example, the years of work experience may have different impacts on the salary at the tails than at the center of the salary distribution. In such cases, quantile regression (QR), introduced by Koenker and Bassett (1978), provides a comprehensive analysis of the relationship

between the response and covariates, as it can automatically capture the popular heterogeneity. In addition, since quantile regression does not make any parametric distributional assumptions, the inference can be regarded as nonparametric and regression at not very extreme quantiles is more robust to outlier than the least squares regression.

1.1.2 Univariate Quantile Function

For any univariate random variable Z, its distribution is completely determined by the cumulative distribution function (CDF) $F(z)$. Conversely, we may also describe the distribution function $F(z)$ by its inverse function, also known as the quantile function. For a given quantile lever $\tau \in (0,1)$, the τ-th quantile function of Z is defined by

$$Q_\tau(z) = \inf\{z : F(z) \geq \tau\}.$$

Suppose we observe a random samples $\{z_1, \cdots, z_n\}$ of Z. The τ-th sample quantile of z, $\hat{Q}_\tau(z)$, can be obtained by using the order statistics. Equivalently, $\hat{Q}_\tau(z)$ can be obtained by

$$\hat{Q}_\tau(z) = \arg\min_{q \in R} \sum_{i=1}^{n} \rho_\tau(z_i - q),$$

where $\rho_\tau(u) = u\{\tau - I(u \leq 0)\}$ is the check loss function. The check loss function is a piecewise linear function that assigns weight τ to positive and $1-\tau$ to negative residuals.

1.1.3 Linear Quantile Regression for Independent Data

The idea of estimating the τth quantile of a single variable can be extended to regression setups to estimate the τ-th conditional quantile of a response given the covariates. Let Y be a univariate response variable and $\mathbf{X}$ be a p-dimensional covariate vector. Denote $F_Y(\cdot|\mathbf{X} = \mathbf{x})$ as the conditional cumulative distribution function (CDF) of Y given $\mathbf{X} = \mathbf{x}$. Then, at a given quantile level $\tau \in (0,1)$, the τ-th conditional quantile of Y given $\mathbf{X} = \mathbf{x}$ is defined as

$$Q_\tau(Y|\mathbf{X} = \mathbf{x}) = \inf\{t : F_Y(t|\mathbf{X} = \mathbf{x}) \geq \tau\}.$$

Throughout, we assume that the first element of $\mathbf{X}$ is one corresponding to the intercept.

The linear QR model assumes that

$$Q_\tau(Y|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau, \ i = 1, ..., n,$$

where $\boldsymbol{\beta}_\tau$ is a p-dimensional vector of unknown quantile coefficients. Given a random sample $\{(y_i, \mathbf{x}_i), i = 1, ..., n\}$, $\boldsymbol{\beta}_\tau$ can be estimated by solving the following optimization problem:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{1.1}$$

The quantile regression problem (1.1) can be reformulated as a linear programing problem:

$$\min_{(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v}) \in R^p \times R_+^n \times R_+^n} \{\tau 1_n^T \mathbf{u} + (1 - \tau) 1_n^T \mathbf{v} | \mathbf{x}_i^T \boldsymbol{\beta} + u_i - v_i = y_i\},$$

where 1_n is the n-dimensional vector of ones, $\mathbf{u}$ and $\mathbf{v}$ denote n-dimensional vectors of the positive and the negative parts of the residuals, and u_i and v_i are the i-th elements of $\mathbf{u}$ and $\mathbf{v}$, respectively. Details on the linear programming formulation and computation for quantile regression can be found in Chapter 6 of Koenker (2005).

1.2 Quantile Regression for Models with Random Effects

1.2.1 Clustered Data

Clustered data are encountered in many applications, for instance in medical studies with repeated measurements from the same individual and in educational studies with test scores of students in the same class. The key feature of clustered data is that measurements from the same cluster often have some common characteristics and thus tend to be correlated.

Similarly, longitudinal data are also characterized by multiple observations of subjects that are measured over time. For example, Yu et al. (2000) studied different child asthma management regimes and analyzed the relationship between daily measures of pollution and asthma symptoms. In this example, children form the cluster and asthma symptoms

from the same child are likely to be correlated.

Throughout this dissertation, we do not distinguish these two types of data, and we use clustered data to refer to both data with repeated measurements of the same subject taking at different times or measurements from different subjects in the same cluster.

The clustered data analysis is to study the impact of covariates on the conditional mean or quantiles of the response for a given cluster. Clustered data analysis can be challenging because observations from the same cluster are often dependent. Thus, we need to take into account the correlation among repeated measurements within the same cluster. In the literature, two different models are often used for analyzing clustered data: the marginal and the conditional models. Next, we will introduce and compare two types of models for mean regression and quantile regression separately.

1.2.2 Mean Regression for Clustered Data

Suppose we observe data $(y_{ij}, \mathbf{x}_{ij})$, $i = 1, ..., n$ and $j = 1, ..., n_i$, where y_{ij} and $\mathbf{x}_{ij}$ are the univariate response and p-dimensional covariate vector associated with cluster (or subject) i for the j-th subject (or measuring time). Write $\mathbf{x}_i = (\mathbf{x}_{i1}^T, ..., \mathbf{x}_{in_i}^T)^T$ and $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})^T$ as the design matrix and the response vector for cluster i, respectively.

The marginal linear regression model captures the population-average trend among all clusters. The marginal regression models specify marginal mean $E(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ as a function of the covariate $\mathbf{x}_{ij}$ (Gardiner et al., 2009). An example of the marginal model is

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_{ij}, \tag{1.2}$$

where u_{ij} are the random errors that are independent across clusters but may be dependent within the same cluster i and $\boldsymbol{\beta}$ is a p-dimensional marginal regression coefficient. The marginal effect $\boldsymbol{\beta}$ can be obtained by solving generalized estimating equation (GEE), introduced by Liang and Zeger (1986).

The conditional linear regression model captures the individual trend for each cluster. The conditional model specifies the mean of Y conditional on both the covariate $\mathbf{x}_{ij}$ and unobserved random effects $\mathbf{b}_i$:

$$E(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i,$$

where $\mathbf{z}_{ij}$ is often a sub-vector of $\mathbf{x}_{ij}$. Correlation among repeated measurements within the same cluster arises from sharing the unobserved random effect $\mathbf{b}_i$. A typical example of the conditional model is linear mixed models (LMM) from Laird and Ware (1982):

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i + \epsilon_{ij}, \tag{1.3}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{z}_{ij}$ is a $q \times 1$ covariate vector associated with cluster-specific random effects $\mathbf{b}_i$, and ϵ_{ij} are independent random measurement errors. We usually assume that $E(\mathbf{b}_i|\mathbf{x}_{ij}) = \mathbf{0}$ and $E(\epsilon_{ij}) = 0$ and $\mathbf{b}_i$ are mutually independent of each other and independent of ϵ_{ij}. The conditional regression coefficients can be estimated by maximum likelihood estimation or restricted maximum likelihood estimation.

In the conditional model (1.3), the random effect $\mathbf{b}_i$ measures the covariate effects that vary across clusters in the population. The assumptions on e_{ij} and $\mathbf{b}_i$ fully determine the correlation structure of the response. Thus, we can specify the model for the population, and obtain the marginal effects. If we assume that $E(\mathbf{b}_i|\mathbf{x}_{ij}) = \mathbf{0}$ and $E(\epsilon_{ij}) = 0$ in (1.3), then model (1.3) implies that $E(Y_{ij}|x_{ij}) = x_{ij}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the same as in model (1.2) and can be interpreted as the population-averaged effects of covariates.

The main difference between the marginal models and the conditional models is whether the focus of the study is on comparing the responses within the same cluster or across clusters. Consider the case $\mathbf{z}_{ij} = \mathbf{x}_{ij}$. In conditional models, the sum of fixed and random effects $\boldsymbol{\beta} + \mathbf{b}_i$ has the cluster-specific interpretation. Then, we may interpret $\boldsymbol{\beta} + \mathbf{b}_i$ as the change in the mean of the response when covariates $\mathbf{x}$ changes by one unit for cluster i. In marginal models, the fixed effect $\boldsymbol{\beta}$ has the population-averaged interpretation and describes how the population marginal mean of the response changes as the covariate $\mathbf{x}$ changes by one unit, regardless of the cluster.

1.2.3 Quantile Regression Models for Clustered Data

Most linear mixed model analyses for clustered data also assume that both the random effects and random errors are normally distributed with constant variances. Such assumptions imply that the covariates affect only the location of the response distribution and thus cannot accommodate population heterogeneity. However, in some applications, the covariates may have different impacts at different locations of the response distribution. For instance, in a birth weight study, Abrevaya and Dahl (2008) found that some covari-

ates such as the gender of the baby and the mother's prenatal-care visits have different effects at the lower and upper quantiles of the infant birth weight distribution. By focusing on the conditional quantiles, quantile regression (Koenker and Bassett, 1978) offers an alternative tool that can automatically capture the heterogeneity in covariate effects at different quantiles of the response distribution without modeling the heteroscedasticity.

As in the mean regression models, we may consider the marginal and conditional quantile regression models for analyzing clustered data.

The marginal quantile regression model captures the average trend among all clusters. Similar to the mean regression models, the marginal linear quantile regression model assumes that the τ-th quantile function $Q_\tau(Y_{ij}|\mathbf{x}_{ij})$ is linear in the covariates, such as in the following quantile regression model:

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau,m}, \tag{1.4}$$

where τ is a given quantile level of interest and $\boldsymbol{\beta}_{\tau,m}$ is a p-dimensional vector of the marginal effect of $\mathbf{x}_{ij}$ on the τ-th quantile of Y_{ij}.

The conditional quantile regression model captures an individual trend for each cluster and specifies the τ-th conditional quantile function $Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i)$ conditional on both the covariates and unobserved cluster-specific random effects $\mathbf{b}_i$, that is,

$$Q_\tau(Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau,c} + \mathbf{z}_{ij}^T\mathbf{b}_i, \tag{1.5}$$

where $\boldsymbol{\beta}_{\tau,c}$ is a p-dimensional vector of conditional quantile coefficient at the τ-th quantile.

We need to point out one difference between the mean and quantile regression models. For mean regression models, the parameter vector $\boldsymbol{\beta}$ in the marginal and conditional models are the same since $E_{\mathbf{b}}\{E(y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i)\} = E(y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$. However, for quantile regression models, we usually do not expect that $\boldsymbol{\beta}_{\tau,m} = \boldsymbol{\beta}_{\tau,c}$. For example, consider the model (1.3), where e_{ij} are random errors. The fixed effect $\boldsymbol{\beta}_{\tau,c}$ is not the population-averaged effect unless the τ-th conditional quantile of $\mathbf{z}_{ij}^T\mathbf{b}_i + e_{ij}$ given $\mathbf{x}_{ij}$ is zero, that is, $Q_\tau(\mathbf{z}_{ij}^T\mathbf{b}_i + e_{ij}|\mathbf{x}_{ij}) = 0$ for all i and j.

Model (1.5) assumes that the random effects are constant across all quantile levels. We give another example below to show that the random effects could depend upon the

quantile level. Consider the following heteroscedastic model

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{b}_i \gamma \epsilon_{ij}, \tag{1.6}$$

where $\mathbf{b}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \sigma_b^2 I_p)$, $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\boldsymbol{\gamma}$ is a scalar scale parameter. The corresponding conditional quantile regression models is

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{x}_{ij}^T \mathbf{b}_i \gamma \sigma \Phi^{-1}(\tau). \tag{1.7}$$

The cluster-specific effect of the l-th predictor on the τ-th quantile of Y in cluster i is $\beta_l + \gamma b_{il} \sigma \Phi^{-1}(\tau)$, where β_l and b_{il} are the l-th element of $\boldsymbol{\beta}$ and $\mathbf{b}_i$, respectively. Hence, the cluster-specific effect depends on τ through the τ-th quantile of a normal distribution. Since both $\mathbf{x}_{ij}^T \mathbf{b}_i \gamma$ and ϵ_{ij} are independent normal random variables, the density function of the product of two normals $\mathbf{x}_{ij}^T \mathbf{b}_i \gamma \epsilon_{ij}$ is proportional to a modified Bessel function of the second kind (Glen et al., 2004). Therefore, the marginal distribution of Y_{ij} given $\mathbf{x}_{ij}$ has the same shape as in Figure 1.1 with mean $\mathbf{x}_{ij}^T \boldsymbol{\beta}$.

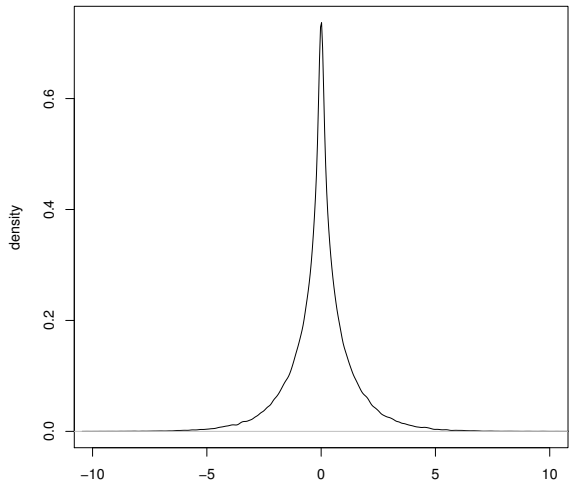


Figure 1.1: The density function of $W = UV$, where U and V are independent standard normal random variables.

In mean regression models, the marginal effects can be obtained by integrating out cluster-specific random effects in the conditional model. However, such relationship does not hold for the marginal and conditional quantile regression models. The marginal quantile function derived from the conditional linear quantile regression model may not even be linear in $\mathbf{x}_{ij}$. For example, consider the following heteroscedastic error model

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ij}^T\mathbf{b}_i + \mathbf{x}_{ij}^T\boldsymbol{\gamma}\epsilon_{ij},$$

where $\mathbf{b}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \Sigma)$, $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\boldsymbol{\gamma}$ is a p-dimensional scale vector. Both the marginal mean $E(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$ and the conditional mean $E(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{x}_{ij}^T\mathbf{b}_i$ are linear in $\mathbf{x}_{ij}$. In contrast, the conditional quantile function

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T\{\boldsymbol{\beta} + \boldsymbol{\gamma}\sigma\Phi^{-1}(\tau)\} + \mathbf{x}_{ij}^T\mathbf{b}_i$$

is linear in $\mathbf{x}_{ij}$ but the marginal quantile function

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \Phi^{-1}(\tau)\sqrt{\mathbf{x}_{ij}^T(\Sigma + \sigma^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T)\mathbf{x}_{ij}}$$

is not linear in $\mathbf{x}_{ij}$ unless $\sigma^2\boldsymbol{\gamma}\boldsymbol{\gamma}^T = \mathbf{0}$, where Φ^{-1} is the inverse CDF of the standard normal distribution.

Similar to mean regression models, marginal quantile regression models have the population quantile interpretation and the conditional quantile regression models have the cluster-specific quantile interpretation. Appropriate model can be selected based on whether the focus of the analysis is on the population quantile or on cluster-specific quantiles.

In Section 1.2.4, we will review existing methods for the marginal and conditional quantile regression models, respectively.

1.2.4 Methods for Quantile Regression Models with Random effects

Existing work for quantile regression with random effects is limited. The main challenges are that quantile regression usually does not make any parametric distributional assumptions, and unlike mean, quantiles are not additive, that is, quantiles of a sum of two

random variables are often not the sum of their quantiles. To bypass these challenges, some researchers analyzed clustered data by considering marginal quantile regression models that treat the sum of random effects and random errors as a unit and focus on the covariate effects averaged over clusters; see for instance, Jung (1996), Wei and He (2006), Wang (2009), Mu and Wei (2009), Tang and Leng (2011), and Liya and Wang (2012).

For a conditional quantile regression model with a random intercept, Koenker (2004) proposed a regularization method, where a L_1 penalty is introduced to shrink the random effects towards a common value. Koenker (2004) proposed to estimate b_i and $\boldsymbol{\beta}_{\tau,c}$ at multiple quantile levels $\tau_k, k = 1, ..., q$ by minimizing the following penalized objective function:

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\sum_{j=1}^{n_i} w_k \rho_\tau\{y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_k} - b_i\} + \lambda\sum_{i=1}^{n}|b_i|,$$

where w_k is the weight on the quantile level τ_k and λ is the regularization parameter that controls the variation of b_i and helps shrink b_i towards a common value. Koenker (2004) assumed that b_i are the same across quantiles and proposed to estimate b_i by combining information across quantiles. As in most regularization problems, the choice of the penalty parameter λ is crucial.

Several authors proposed parametric or semiparametric Bayesian approaches for quantile regression with random effects. Geraci and Bottai (2007), Yuan and Yin (2010) and Geraci and Bottai (2013) extended the asymmetric Laplace distribution idea in Yu and Moyeed (2001) for quantile regression with independent data to clustered data, where the conditional distribution of the response is assumed to follow an asymmetric Laplace distribution with the mean depending on the covariates and the skewness parameter depending on the quantile level of interest. Instead of posing a parametric likelihood, Reich et al. (2010) proposed to model the likelihood nonparametrically by an infinite mixture of quantile-restricted two-component Gaussian mixtures and to accommodate error heteroscedasticity by specifying its form parametrically. Yang and He (2012) proposed the empirical likelihood as a working likelihood for Bayesian quantile regression for independent data and this method was extended to clustered data by Kim and Yang (2011). Unlike the other methods, the empirical likelihood approach avoids modeling the error distribution. One commonality of these existing Bayesian methods is that the analysis

(and modeling) is carried out at a single quantile level separately. Such separate analyses have two major limitations. First, the resulting estimated conditional quantiles are not guaranteed to be monotonically increasing in the quantile level and thus quantile crossing may be encountered. Second, for the methods in Geraci and Bottai (2007), Yuan and Yin (2010), Geraci and Bottai (2013) and Reich et al. (2010), since the error distribution is modeled at each quantile level separately, there is a lack of consistency of likelihood across quantiles. Since Bayesian quantile regression for clustered data is the main focus of Chapter 3, we will revisit this issue and discuss more in Section 3.1.

1.3 Introduction to Extreme Value Theory

Extreme value theory (EVT) studies the limiting behaviors of extreme values at the tails of the distribution of interest. There are two common approaches to model extreme values: block maxima (BM) and peak-over-threshold (POT). Roughly speaking, the BM approach models the extremely rare events in the long period of time and the POT approach models the extremely rare events that exceed a high threshold. For example, the BM approach deals with yearly maximal rainfalls and the POT approach deals with rainfalls exceeding 4 inches.

1.3.1 Block Maxima (BM) Method

Fisher and Tippett (1928) and Gnedenko (1943) established the limiting distribution of the normalized maximum of a random sample, which is considered as the cornerstone of the BM approach. Let $X_1, ..., X_n$ be a sequence of mutually independent random variables from a common distribution F. Fisher and Tippett (1928) and Gnedenko (1943) focused on the behavior of the largest member of a random sample $M_n = \max\{X_1, X_2, ..., X_n\}$. Since the distribution of M_n degenerates to a point mass on x that satisfies $F(x) = 1$ as $n \to \infty$, Fisher and Tippett (1928) linearly normalized M_n and identified the three types of limiting distributions of the normalized M_n. Fisher-Tippett theorem states that if there exist sequences of normalizing constants $a_n > 0$ and b_n such that

$$P\left(\frac{M_n - b_n}{a_n} < x\right) \to G(x) \text{ as } n \to \infty, \tag{1.8}$$

where G is a nondegenerate distribution, then G must belong to one of the three families: Gumbel, Fréchet and Weibull. These three types of families can be combined into the generalized extreme value (GEV) distribution of the form

$$G(x) = G_{GEV}(x; \sigma, \xi, \mu) = \begin{cases} \exp\{-(1+\xi\frac{x-\mu}{\sigma})^{-1/\xi}\} & \text{for } 1+\xi\frac{x-\mu}{\sigma} \geq 0, \xi \neq 0 \\ \exp\{-\exp(-\frac{x-\mu}{\sigma})\} & \text{for } x \in \mathbb{R}, \xi = 0, \end{cases}$$

where $\sigma > 0$ is the scale parameter, ξ is the shape parameter and μ is the location parameter. If $\xi = 0$, G is the Gumbel-type distribution and it is associated with distributions with moderate tails such as normal, exponential and gamma distributions. If $\xi > 0$, G is the Fréchet-type distribution and it is associated with heavy-tailed distributions such as t-distribution and Pareto distributions. If $\xi < 0$, G is the Weibull-type distribution, has finite support and is associated with light-tailed distributions such as Beta distribution. A more comprehensive review can be found in Coles (2001).

1.3.2 Peak-Over-Threshold (POT) Method

The POT method, introduced by Pickands (1975) and Balkema and de Haan (1974), models extreme events that exceed a high threshold. Suppose $X_1, ..., X_n$ is a sequence of mutually independent random variables from a common distribution F and Y_i is the excess amount over the threshold u, that is, $Y_i = X_i - u$. Then, for any $y > 0$, the probability of Y being greater than y is given by

$$P(Y > y|Y > 0) = P(X > u + y|X > u) = \frac{P(X > u + y)}{P(X > u)} = \frac{1 - F(u + y)}{1 - F(u)}. \tag{1.9}$$

By Fisher-Tippett theorem in Fisher and Tippett (1928), we can approximate the distribution of the normalized $M_n = \max\{X_1, X_2, ..., X_n\}$ for large n,

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \approx G_{GEV}(x; \sigma', \xi', \mu') \tag{1.10}$$

for some parameters $\sigma' > 0$, $\xi' \in \mathbb{R}$ and $\mu' \in \mathbb{R}$. Let $z = a_n x + b_n$ and then (1.10) is equivalent to

$$P(M_n \leq z) = \{F(z)\}^n \approx G_{GEV}\left(\frac{z - b_n}{a_n}; \sigma', \xi', \mu'\right) = G^*_{GEV}(z; \sigma, \xi, \mu),$$

where G^* is the GEV distribution with parameters $\sigma > 0$, $\xi \in \mathbb{R}$ and $\mu \in \mathbb{R}$. The first order Taylor expansion gives us

$$n \log F(z) \approx -n\{1 - F(z)\}. \tag{1.11}$$

After plugging $G^*_{GEV}(z; \sigma, \xi, \mu)$ into the left-hand side of (1.11), we obtain

$$1 - F(z) \approx \frac{1}{n} \log G^*_{GEV}(z; \sigma, \xi, \mu) \tag{1.12}$$

for large n. Thus, combining (1.9) and (1.12) gives us the following approximation:

$$\begin{aligned} P(Y > y|Y > 0) &= \frac{1 - F(u + y)}{1 - F(u)} \\ &\approx \begin{cases} \left(1 + \frac{\xi y}{\sigma + \xi(u-\mu)}\right)^{-1/\xi} & \text{for } 1 + \frac{1}{\sigma}\xi(y + u - \mu) \geq 0, \xi \neq 0 \\ \exp(-\frac{y}{\sigma}) & \text{for } y \in \mathbb{R}, \xi = 0. \end{cases} \end{aligned}$$

By letting $\sigma + \xi(u - \mu) = \varrho$, we obtain the generalized Pareto distribution $GPD(\varrho, \xi)$

$$H(y) = H_{GPD}(y; \varrho, \xi) = \begin{cases} 1 - (1 + \xi \frac{y}{\varrho})^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp(-\frac{y}{\varrho}), & \xi = 0, \end{cases}$$

where $y > 0$ for $\xi \geq 0$ and $0 \leq y < -\varrho/\xi$ for $\xi < 0$ with ξ and $\varrho > 0$ being the shape and shape parameter of the GPD, respectively. In summary, Pickands-Balkema-de Haan theorem states that, if F satisfies Fisher-Tippett theorem, then the excess conditional distribution function of $Y = X - u|X > u$ can be approximated by $H_{GPD}(y; \varrho, \xi)$ with some parameters $\varrho > 0$ and ξ for large enough threshold u.

Parameters ϱ and ξ can be estimated by maximum likelihood (ML) method. The log likelihood function of $y_1, ..., y_k$, excesses of a given threshold u, can be written as

$$\log l(\varrho, \xi; y) = \begin{cases} -k \log \varrho - (1 + 1/\xi) \sum_{i=1}^{k} \log(1 + \xi \frac{y_i}{\varrho}), & \xi \neq 0 \\ -k \log \varrho - \sum_{i=1}^{k} y_i/\varrho, & \xi = 0. \end{cases} \tag{1.13}$$

The likelihood cannot be maximized analytically and thus numerical maximization is required. We may use the function $fitgpd$ in the R package "POT" for numerical maximum likelihood estimation.

The threshold u creates a trade-off between bias and variance. A high threshold value will produce high variances because it leads to few excesses and the estimation will rely on a small sample size. In contrast, a low threshold value will result in bias because the approximation in the Pickands-Balkema-de Haan theorem is less accurate if threshold u is not large enough. For more details on ML estimation and threshold selection, please refer to Davison and Smith (1990) and Chapter 4 of Coles (2001). More details on applications of the POT method and suggested reading list can also be found in Chapter 4 of Coles (2001).

Chapter 2

Linearly Interpolated Density (LID) Method for Linear Quantile Regression

In this chapter, we propose a semiparametric Bayesian quantile regression (QR) method based on the linearly interpolated density (LID) method originally proposed by Feng (2011). The LID method is a Bayesian QR method for cross sectional data using linear interpolations of quantiles to approximate the likelihood.

In Section 2.1, we describe linearly interpolated density using a univariate distribution. Then, we discuss the motivation of using the LID method for quantile regression and introduce the LID method proposed by Feng (2011) in Section 2.2, and present our proposed LID method in Section 2.3, which adopts the extreme value theory to improve the tail density estimation. Simulation study is carried out in Section 2.4 to compare the performance of existing methods.

2.1 Linearly Interpolated Density (LID) Method

We introduce the idea of the LID method by considering a univariate distribution. Suppose that we have a univariate continuous random variable, Z, with continuous distribution function F and density function f. Let $F^{-1}(\tau) = Q(\tau)$ denote the τ-th quantile function of Z. Note that $\tau = F\{F^{-1}(\tau)\}$. By taking the derivative with respect to τ on

both sides, we get

$$1 = \frac{d}{d\tau} F\{F^{-1}(\tau)\} = f\{F^{-1}(\tau)\} \frac{d}{d\tau} F^{-1}(\tau).$$

Therefore,

$$\frac{d}{d\tau} F^{-1}(\tau) = \frac{1}{f\{F^{-1}(\tau)\}}.$$

Let $0 < \tau_1 < ... < \tau_K < 1$ be a sequence of quantile levels. By mean value theorem, there exists a $z^* = F^{-1}(\tau_{z^*}) \in (F^{-1}(\tau_j), F^{-1}(\tau_{j+1}))$ such that

$$\left. \frac{dF^{-1}(\tau)}{d\tau} \right|_{\tau=\tau_{z^*}} = \frac{F^{-1}(\tau_{j+1}) - F^{-1}(\tau_j)}{\tau_{j+1} - \tau_j},$$

where $\tau_j < \tau_z^* < \tau_{j+1}$ and consequently,

$$f(z^*) = \frac{\tau_{j+1} - \tau_j}{F^{-1}(\tau_{j+1}) - F^{-1}(\tau_j)}.$$

To approximate $f(z)$ with $\tau_j < \tau_z = F(z) < \tau_{j+1}$, note that if $\tau_j \to \tau_z^-$ and $\tau_{j+1} \to \tau_z^+$,

$$f(z^*) = \lim_{\tau_j \to \tau_z^-, \tau_{j+1} \to \tau_z^+} \frac{\tau_{j+1} - \tau_j}{F^{-1}(\tau_{j+1}) - F^{-1}(\tau_j)} = f(z).$$

Therefore, for any z such that $F^{-1}(\tau_j) < z < F^{-1}(\tau_{j+1})$ for some $j = 1, ..., K$, we can approximate $f(z)$ by

$$f(z) \approx \frac{\tau_{j+1} - \tau_j}{F^{-1}(\tau_{j+1}) - F^{-1}(\tau_j)}. \tag{2.1}$$

In practice, $F^{-1}(\tau_j)$ is unknown but can be estimated by the sample quantile $\hat{Q}(\tau_j), j = 1, \cdots, K$, based on the observed sample. Then, the substitution of $F^{-1}(\tau_j)$ by $\hat{Q}(\tau_j)$ in (2.1) enables us to estimate $f(z)$ and we refer to the approximate density as linearly interpolated density (LID).

To further illustrate how the LID method operates, we randomly generated two data sets of size $n = 1000$, one from the standard normal distribution and the other from χ^2 distribution with $df = 3$. Figure 2.1 shows that the LID well describes the overall shapes of both symmetric and skewed distributions. For both data sets, fifteen equally

spaced quantile levels from 0.0625 to 0.9375 were used. The dashed curves depict the true densities and the solid lines correspond to the approximate density estimates by the LID method. The approximate densities are piecewise constant functions and inversely proportional to the distance between two adjacent sample quantiles.

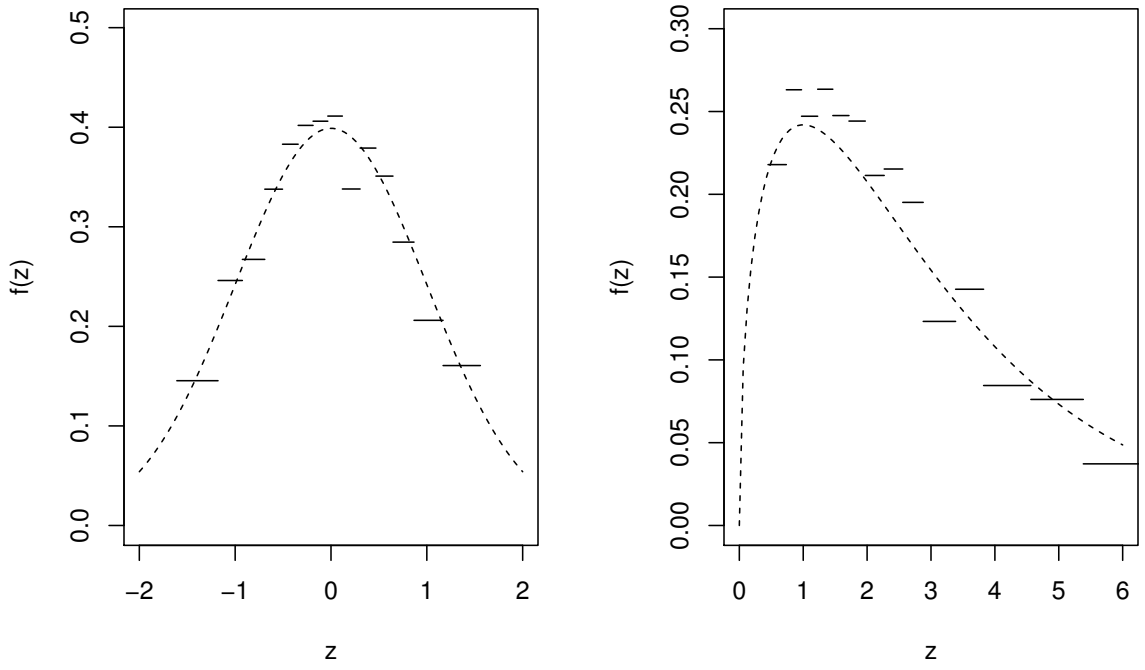


Figure 2.1: The LID estimates for data sets generated from $N(0,1)$ and χ^2_3. The dashed curves are the true density functions and the solid lines are the LID estimations to $N(0,1)$ (the left plot) and χ^2_3 (the right plot) .

The LID method can be applied to estimate the density of z at any point that lies between the τ_1 and τ_K-th quantiles. However, it is difficult to use the LID method for estimating the tail density since the tail quantiles are hard to be estimated accurately for small samples without any parametric distributional assumption. For tails, Feng (2011) proposed to use a truncated normal distribution to model the tail density. We will discuss more about the parametric tail density estimation approach suggested by Feng (2011)

and its limitations in Section 2.3 under the regression setup.

2.2 LID Method for Linear Quantile Regression

2.2.1 Benefits of the LID method

Most existing Bayesian methods focus on a single quantile level and those approaches may encounter quantile crossing problem and inconsistency of likelihood when the analysis is carried out at multiple quantile levels.

In contrast, the LID method avoids those limitations by joint-modeling of multiple quantiles under the assumption that the quantiles of the response are linear in covariates. With this joint-modeling, we can avoid quantile crossing and achieve higher efficiency of estimating quantile coefficients. The method will be discussed in Section 2.2.2 and the efficiency will be assessed via simulation studies in Section 2.4.

2.2.2 LID for Linear Quantile Regression

We introduce the LID method of Feng (2011) for Bayesian quantile regression. Let's consider the linear quantile regression model,

$$Q_\tau(Y|\mathbf{X}) = \mathbf{X}^T\boldsymbol{\beta}_\tau,\ 0 < \tau < 1, \tag{2.2}$$

where Y is the response, $\mathbf{X}$ is a p-dimensional covariate vector and $\boldsymbol{\beta}_\tau$ is the p-dimensional vector of quantile coefficients. Suppose that we observe a random sample, $(\mathbf{y}, \mathbf{x}) = \{(y_i, \mathbf{x}_i), i = 1, ...n\}$ of $(\mathbf{Y}, \mathbf{X})$, and we are interested in estimating $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_{\tau_1}^T, \cdots, \boldsymbol{\beta}_{\tau_k}^T)^T$, where $0 < \tau_1 < \cdots < \tau_K < 1$ is a sequence of quantile levels.

By the Bayes rule, the posterior distribution of $\boldsymbol{\beta}_K$ given $(\mathbf{y}, \mathbf{x})$ is

$$f(\boldsymbol{\beta}_K|\mathbf{y}, \mathbf{x}) \propto \prod_{i=1}^{n} f_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K)\pi(\boldsymbol{\beta}_K), \tag{2.3}$$

where $\pi(\boldsymbol{\beta}_K)$ denotes the prior distribution of $\boldsymbol{\beta}_K$.

In quantile regression, usually no parametric form on the conditional distribution $f_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K)$ is assumed. Instead, we assume the quantile regression model (2.2). Note that there may exist multiple conditional distributions that lead to the same conditional

quantile functions at $\tau_k, k = 1, \cdots, K$. Using the idea of the integrated likelihood methods by Berger et al. (1999), the conditional density $f_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K)$ is defined by

$$f_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K) = \int f_j(y_i|\mathbf{x}_i)\pi(f_j|\mathbf{x}_i, \boldsymbol{\beta}_K)df_j,$$

where $f_j(y|\mathbf{x}_i)$ is the conditional density function of Y given $\mathbf{X} = \mathbf{x}_i$ that leads to the same conditional quantile functions at τ_k, $k = 1, \cdots, K$ and $\pi(f_j|\mathbf{x}_i, \boldsymbol{\beta}_K)$ is the prior on $f_j(Y_i|\mathbf{x}_i)$. Then, we can interpret $f_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K)$ as the approximate likelihood averaged over all possible f_j's that share the same conditional quantiles at the K quantile levels.

Note that we can link the conditional density function $f_Y(\cdot|\mathbf{x}, \boldsymbol{\beta}_K)$ with model (2.2) by the following equation:

$$f_Y(y|\mathbf{x}, \boldsymbol{\beta}_K) = \lim_{\delta \to 0} \frac{\delta}{\mathbf{x}^T(\boldsymbol{\beta}_{u_y+\delta} - \boldsymbol{\beta}_{u_y})},$$

where $u_y = \{\tau \in (0, 1) : \mathbf{x}^T\boldsymbol{\beta}_\tau = y\}$. Let $0 < \tau_1 < \cdots < \tau_K < 1$ be a grid of quantile levels. Therefore, for any y with associated covariate $\mathbf{x}$ such that $F_Y^{-1}(\tau_j|\mathbf{x}, \boldsymbol{\beta}_K) < y < F_Y^{-1}(\tau_{j+1}|\mathbf{x}, \boldsymbol{\beta}_K)$ for some $j = 1, \ldots, K-1$, we can approximate $f_Y(y|\mathbf{x}, \boldsymbol{\beta}_K)$ by

$$\frac{\tau_{j+1} - \tau_j}{\mathbf{x}^T(\boldsymbol{\beta}_{\tau_{j+1}} - \boldsymbol{\beta}_{\tau_j})}.$$

Feng (2011) proposed a Bayesian quantile regression procedure based on the following linearly interpolated density (LID)

$$\begin{aligned}\hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K) = {} & I\{y_i \in (-\infty, \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1})\}\tau_1 f_1(y_i) \\ & + \sum_{k=1}^{K-1} I\{y_i \in [\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}, \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_{k+1}})\}\frac{\tau_{k+1} - \tau_k}{\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_{k+1}} - \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}} \\ & + I\{y_i \in [\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_K}, \infty)\}(1 - \tau_K)f_2(y_i),\ i = 1, \ldots, n, \end{aligned} \tag{2.4}$$

where f_1 and f_2 are some prespecified left and right tail density functions. Particularly, Feng (2011) assumed $f_1(\cdot)$ and $f_2(\cdot)$ to be the truncated normal distributions $N_{(-\infty, \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}]}(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}, \sigma_1^2)$ and $N_{(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}, \infty)}(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}, \sigma_2^2)$, respectively, where σ_1 and σ_2 are some pre-specified scale parameters. The LID method may be senesitive to the seletion of σ^2. Hence, in the simulation studies, we consider several values for σ_1^2 and σ_2^2. In addition,

we can avoid pre-specifying σ_1^2 and σ_2^2 by matching the linearly interpolated density and truncated normal tail density at $\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}$ and $\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}$, that is, we can estimate σ_1^2 and σ_2^2 by

$$\hat{\sigma}_1 = 2\tau_1 \sum_{i=1}^{n} \frac{\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_2} - \mathbf{x}_i^T\boldsymbol{\beta}(\tau_1)}{n\sqrt{2\pi}(\tau_2 - \tau_1)} \text{ and}$$
$$\hat{\sigma}_2 = 2\{1-\tau_K) \sum_{i=1}^{n} \frac{\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k} - \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_{K-1}}}{n\sqrt{2\pi}(\tau_K - \tau_{K-1})}.$$

Algorithm of LID for Linear Quantile Regression

We describe the algorithm for drawing samples from the approximate posterior distributions. Given the approximate likelihood $\hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K)$ and the prior $\pi(\boldsymbol{\beta}_K)$, we draw the posterior samples of $\boldsymbol{\beta}_K$ using the following algorithm.

Step 1. Choose initial values for the quantile coefficients $\boldsymbol{\beta}_K^{(0)}$.

(a) Apply the conventional quantile regression. First, regress y_i on $\mathbf{x}_i$ at each quantile level $\tau_k, k = 1, ..., K$, separately and obtain $\boldsymbol{\beta}_K^{(0)}$. We can use the rq function in the R package "quantreg."

Step 2. Propose a move.

(a) Randomly pick a quantile level, τ_k, and select a random number from 1 to p, say l. Then, let β^*_{l,τ_k} denote the candidate for β_{l,τ_k}, the lth element of $\boldsymbol{\beta}_{\tau_k}$.

(b) To avoid quantile crossing, we draw β^*_{l,τ_k} from a proposal distribution that satisfies the following monotonicity constraints for all $i = 1, \cdots, n$:

$$\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_{K-1}} \leq \mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k}, k = 2, ..., K.$$

The further description of the proposal distribution is given in the Appendix.

The l-th element of $\boldsymbol{\beta}_{\tau_k}$ will be replaced by β^*_{l,τ_k}, generating the candidate $\boldsymbol{\beta}^*_K$.

Step 3. Take

$$\boldsymbol{\beta}_K^{(t+1)} = \begin{cases} \boldsymbol{\beta}^*_K & \text{with probability} \quad \phi(\boldsymbol{\beta}_K^{(t)}, \boldsymbol{\beta}^*_K) \\ \boldsymbol{\beta}_K^{(t)} & \text{with probability} \quad 1 - \phi(\boldsymbol{\beta}_K^{(t)}, \boldsymbol{\beta}^*_K), \end{cases}$$

where

$$\phi(\boldsymbol{\beta}_K^{(t)}, \boldsymbol{\beta}_K^*) = \min\left\{ \frac{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K^*)\pi(\boldsymbol{\beta}_K^*)}{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K^{(t)})\pi(\boldsymbol{\beta}_K^{(t)})}, 1 \right\},$$

where $\pi(\cdot)$ is the prior of $\boldsymbol{\beta}_K$ and $\hat{f}_Y(y|\mathbf{x}, \boldsymbol{\beta}_K)$ is the approximate density as in (2.4).

Step 4. Repeat Steps 2 and 3 until the MCMC chain converges. After a burn-in period, we obtain posterior samples of $\boldsymbol{\beta}_K$. We estimate each parameter by using the mean of the samples.

2.3 LID Method with Improved Tail Density Estimation

The method of Feng (2011) relies on the specification of the tail densities f_1 and f_2 and the scale parameter σ. Our numerical studies suggest that the method could be sensitive to the choice of σ; see Section 2.4. Moreover, truncated normal distributions might not work well for estimating tails of asymmetric, skewed or heavy-tailed error distributions. To improve the tail estimation, we propose a modified approximate likelihood, which estimates the tail densities based on extreme value theory and does not require any prespecification of parameters.

Pickands (1975) and Balkema and de Haan (1974) showed that if the distribution function F of a random variable Y belongs to the domain of attraction of generalized extreme value distribution, then as $u \to y_F$, the end point of F, the conditional distribution function $P(Y \le u + y|Y > u)$ approaches the distribution function of a generalized Pareto distribution (GPD). That is, as $u \to y_F$,

$$P(Y \le u + y|Y > u) \to G(y|\varrho, \xi) = \begin{cases} 1 - (1 + \xi y/\varrho)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp(-y/\varrho), & \xi = 0, \end{cases}$$

where $y > 0$ for $\xi \ge 0$ and $0 \le y < -\varrho/\xi$ for $\xi < 0$ with ξ and $\varrho > 0$ being the shape and scale parameters of the GPD, respectively.

In view of the above discussion, we propose to approximate the conditional density of

Y by linear interpolation of conditional quantiles at the central quantiles while use the GPD to approximate the tail densities.

For right tails, we define the upper thresholds $u_i = \{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_K} + \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{K-1}}\}/2$ and approximate the tail distribution by the generalized Pareto distribution (GPD) with scale parameter $\varrho_2 > 0$ and shape parameter ξ_2, $G(y|\varrho_2, \xi_2)$, that is,

$$P(Y < y | y > u_i, \mathbf{x}_i) \approx G(y - u_i | \varrho_2, \xi_2).$$

Note that

$$P(Y < y | y > u_i, \mathbf{x}_i) = \frac{P(Y < y|\mathbf{x}_i) - P(Y < u_i|\mathbf{x}_i)}{P(Y > u_i|\mathbf{x}_i)} \approx \frac{F(y|\mathbf{x}_i) - (\tau_K + \tau_{K-1})/2}{1 - (\tau_K + \tau_{K-1})/2}.$$

Therefore, the right tail density for $y > \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_K}$ can be approximated by

$$\hat{f}_Y(y|\mathbf{x}_i, \boldsymbol{\beta}_K) = (1 - \tau_K) f_2(y|\mathbf{x}_i, \boldsymbol{\beta}_K, \varrho_2, \xi_2) \approx \{1 - (\tau_{K-1} + \tau_K)/2\} g(y - u_i|\varrho_2, \xi_2), \tag{2.5}$$

where $g(\cdot, \varrho, \xi)$ is the density function of GPD(ϱ, ξ) defined as

$$g(y|\varrho, \xi) = \begin{cases} \varrho^{-1}(1 + \xi y/\varrho)^{-(1+1/\xi)}, & \xi \neq 0 \\ \varrho^{-1} \exp(-y/\varrho), & \xi = 0, \end{cases}$$

with $y > 0$ for $\xi \geq 0$ and $0 \leq y < -\varrho/\xi$ for $\xi < 0$.

The scale parameter ϱ in the GPD can be estimated by matching the LID $\hat{f}_Y(y|\mathbf{x}_i, \boldsymbol{\beta}_K)$ and $g(y - u_i|\varrho_r, \xi_r)$ at $y = u_i$. Specifically, for $u_i = \mathbf{x}_i^T(\boldsymbol{\beta}_{\tau_K} + \boldsymbol{\beta}_{\tau_{K-1}})/2$, note that

$$\begin{aligned} &\hat{f}_Y(u_i|\mathbf{x}_i, \boldsymbol{\beta}_K) \approx \{1 - (\tau_{K-1} + \tau_K)/2)\} g(0|\varrho_2, \xi_2) \\ \Leftrightarrow\ &\frac{\tau_K - \tau_{K-1}}{\mathbf{x}_i^T(\boldsymbol{\beta}_{\tau_K} - \boldsymbol{\beta}_{\tau_{K-1}})} \approx \{1 - (\tau_{K-1} + \tau_K)/2)\} \frac{1}{\varrho_2}. \end{aligned}$$

Then, the scale parameter ϱ_2 can be estimated by

$$\hat{\varrho}_2 = \{1 - (\tau_{K-1} + \tau_K)/2)\} \sum_{i=1}^{n} \frac{\mathbf{x}_i^T(\tilde{\boldsymbol{\beta}}_{\tau_K} - \tilde{\boldsymbol{\beta}}_{\tau_{K-1}})}{n(\tau_K - \tau_{K-1})}, \tag{2.6}$$

where $\tilde{\boldsymbol{\beta}}_{\tau_k}$ is the initial estimator of $\boldsymbol{\beta}_{\tau_k}$ for $k = 1, \ldots, K$. In our implementation, we

choose $\tilde{\boldsymbol{\beta}}_{\tau_k}$ as the conventional (frequentist) quantile regression estimator defined as

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

with $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ being the quantile loss function; see Koenker and Bassett (1978). The estimator $\tilde{\boldsymbol{\beta}}_{\tau_k}$ is obtained by applying the "rq" function in the R package *quantreg* at the τ_kth quantile.

Note that if we define $W = -Y$, then $Q_\tau(W|\mathbf{x}) = -Q_{1-\tau}(Y|\mathbf{x})$, that is, the lower tail of Y corresponds to the upper tail of W. Define the lower thresholds $l_i = \{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_1} + \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_2}\}/2$. For $y < \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_1}$, the left tail density can be approximated by

$$\hat{f}_Y(y|\mathbf{x}_i, \boldsymbol{\beta}_K) = \tau_1 f_1(y|\mathbf{x}_i, \boldsymbol{\beta}_K, \varrho_1, \xi_1) \approx g(-y + l_i|\varrho_1, \xi_1)(\tau_1 + \tau_2)/2, \tag{2.7}$$

where $\varrho_1 > 0$ is the scale parameter and ξ_1 is the shape parameter.

Similarly, by matching the GPD density with LID at $y = l_i$, we can estimate the scale parameter ϱ_1 by

$$\hat{\varrho}_1 = (\tau_1 + \tau_2)/2 \sum_{i=1}^{n} \frac{\mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_{\tau_2} - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_{\tau_1}}{n(\tau_2 - \tau_1)}. \tag{2.8}$$

The shape parameters ξ_1 and ξ_2 can then be estimated by maximizing the GPD likelihood based on the exceedances $\{y_i : y_i < l_i, i = 1, \ldots, n\}$ and $\{y_i : y_i > u_i, i = 1, \ldots, n\}$, respectively.

In our implementation, we first estimate the shape and scale parameters ϱ_1, ϱ_2, ξ_1, ξ_2 based on the initial estimator $\tilde{\boldsymbol{\beta}}_K = (\tilde{\boldsymbol{\beta}}_{\tau_1}^T, \ldots, \tilde{\boldsymbol{\beta}}_{\tau_K}^T)^T$, and then keep them unchanged throughout the Markov chain Monte Carlo (MCMC) simulations. Alternatively, we could also update the shape and scale parameters in the MCMC. However, our numerical studies show that the latter approach complicates the algorithm without improving the performance.

The algorithm of the proposed LID method with GPD tail approximation requires an additional procedure in Step 1.

Step 1. Choose initial values for the quantile coefficients $\boldsymbol{\beta}_K^{(0)}$.

(a) Apply the conventional quantile regression. First, regress y_i on $\mathbf{x}_i$ at each

quantile level $\tau_k, k = 1, ..., K$, separately and obtain $\boldsymbol{\beta}_K^{(0)}$. We can use the rq function in the R package "quantreg."

(b) Obtain the scale parameter estimates, $\hat{\varrho}_1$ and $\hat{\varrho}_2$ from the equations (2.8) and (2.6), respectively, using the initial $\boldsymbol{\beta}_K^{(0)}$. Then, obtain the shape parameter estimates, $\hat{\xi}_1$ and $\hat{\xi}_2$, by using the $fitgpd$ function in the R package "POT."

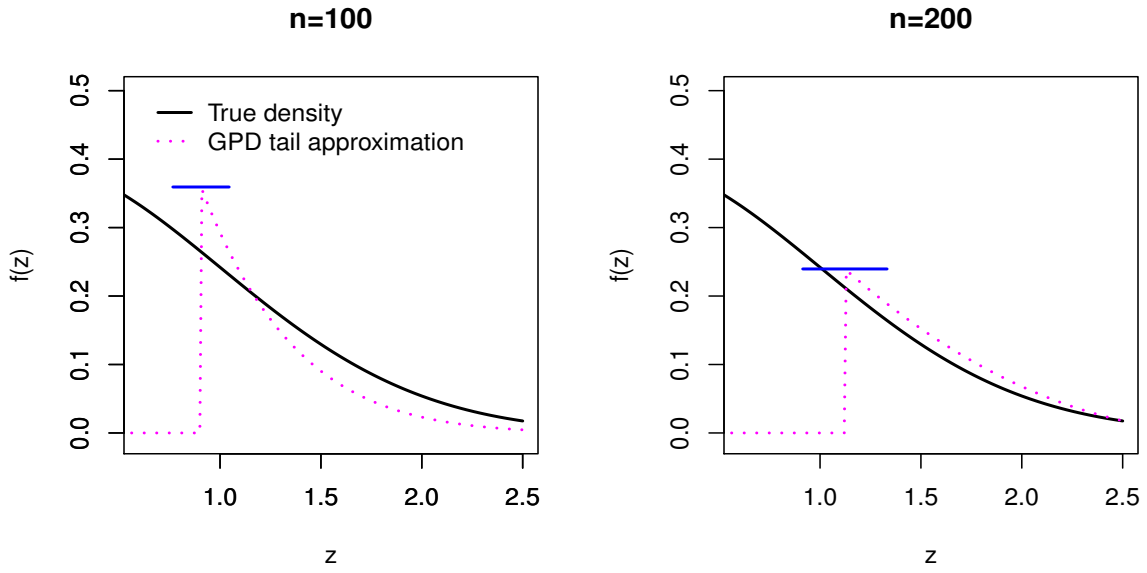


Figure 2.2: The estimated upper tail densities for random samples of size 100 and 200 from the standard normal distribution. The horizontal axis represents z and the vertical axis represents the density. The dotted vertical lines are $z = \hat{\Phi}^{-1}(0.85)$. The solid curves are the true density and the dotted curves correspond to the GPD tail approximation.

Figure 2.2 illustrates right-tail density approximations using the proposed method for random samples of size 100 and 200 from the standard normal distribution. The shape and scale parameters are estimated using exceedances over the 85-th percentile of each random sample. We can observe that the GPD tail approximations are getting finer with larger sample size.

In Appendix A.2, we propose different versions of the LID and LIGPD methods. For example, we can update the shape and scale parameters in the MCMC. One may simplify

the constraints in the proposal distributions. Appendix A.2 summarizes different versions of LIGPD and LID. These versions lead to more complicated algorithms without much improvement in the performance. Hence, our proposed LIGPD method first estimate the shape and scale parameters based on the initial estimator and then keep them unchanged throughout the MCMC simulations.

2.4 Simulation Studies

We carry out two simulation studies to compare the performance of the LID method (Feng, 2011), the proposed LIGPD method, and the conventional frequentist quantile regression method (Koenker and Bassett, 1978).

2.4.1 Simulation Study I

We consider the following model for generating the simulation data:

$$y_i = \beta_0 + \beta_1 x_i + (1 + x_i)\epsilon_i,\ i = 1, \ldots, n, \tag{2.9}$$

where $x_i \sim LN(0, 1)$, $(\beta_0, \beta_1) = (2, 4)$ and $LN(\mu, \varsigma)$ is the log normal distribution with location parameter μ and scale parameter ς. We consider three distributions for generating the random errors: the standard normal, t_3, and χ^2_1. The corresponding conditional quantile function is

$$Q_\tau(Y|x_i) = \beta_{0,\tau} + \beta_{1,\tau} x_i,$$

where $\beta_{0,\tau} = 2 + F_\epsilon^{-1}(\tau)$, $\beta_{1,\tau} = 4 + F_\epsilon^{-1}(\tau)$ and F_ϵ is the distribution function of ϵ_i.

For each scenario, the simulation is repeated 200 times. In the following, we refer to the conventional quantile regression method as QR, the LID method proposed by Feng (2011) based on truncated normal tail distributions with pre-specified variance parameters as LID, the LID method with the proposed variance parameter selection method suggested in Section 2.2.2 as LID II and the proposed method that approximates the tail density using GPD as LIGPD.

We generate $n = 100, 200$ and 500 observations from model (2.9). For both LID and LIGID methods, we use a grid of $K = 7, 11$ and 15 equally spaced quantile levels $\tau_k = k/(1 + K), k = 1, \cdots, K$, and assign the normal priors $N(\mathbf{0}, \Sigma)$ to $(\boldsymbol{\beta}_{\tau_1}, \cdots, \boldsymbol{\beta}_{\tau_K})^T$, where $\boldsymbol{\beta}_{\tau_k} = (\beta_{0,\tau_k}, \beta_{1,\tau_k})^T$ and $\Sigma = \text{diag}(100, \ldots, 100)$. With 100,000 MCMC samples,

we discard the first 50,000 as burn-in, and estimate $\beta_{j,\tau}$ for $j = 0, 1$ by the posterior means of the thinned sample using every 100th draw of the MCMC chain. Following the suggestions in Feng et al. (2014), for the LID method, we take the left tail density $f_1(\cdot)$ as the left half of $N(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}, \sigma^2)$ and the right tail density $f_2(\cdot)$ as the right half of $N(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_K}, \sigma^2)$ with $\mathbf{x}_i = (1, x_i)^T$ and σ^2 a pre-specified constant. To examine the dependence of LID on the choice of σ, we consider $\sigma = 0.5, 1, 1.5, 2$ and 3 for both the normal and t_3 error distributions and $\sigma_1 = 0.5$ and $\sigma_2 = 1, 2, 3, 4, 5$ for χ_1^2 errors.

The posterior means were used as the estimates for $\beta_{j.\tau}, j = 0, 1, \tau = 0.25, 0.5, 0.75$. The mean squared errors (MSE) were obtained by

$$\frac{1}{200}\sum_{m=1}^{200}\{\hat{\beta}_{j,\tau}^{(m)} - \beta_{j,\tau}\}^2, \tag{2.10}$$

where $\hat{\beta}_{j,\tau}^{(m)}, j = 0, 1$ are the candidate estimators for the intercept and slope in the m-th simulation run.

Mean Squared Error Comparison

Tables 2.1–2.3 summarize the mean squared errors of coefficient estimates from different methods at three quartiles (τ=0.25, 0.5, 0.75) for the standard normal, t_3 and χ_1^2 distributions, respectively. The results for the LID estimator in Tables 2.1–2.3 are based on "optimal" σ that give us the smallest sums of the mean squared errors of $\beta_{0,\tau}$ and $\beta_{1,\tau}$ at $\tau = 0.25, 0.5, 0.75$, $\beta_{0,.75} - \beta_{0,.5}$ and $\beta_{1,.75} - \beta_{1,.5}$. Figures 2.3–2.5 show the mean squared errors of $\beta_{0,\tau}$ and $\beta_{1,\tau}$ at $\tau = 0.25, 0.5, 0.75$, $\beta_{0,.75} - \beta_{0,.5}$ and $\beta_{1,.75} - \beta_{1,.5}$ for QR and LID, LID II and LIGPD with $K = 7$ and the ranges of the mean squared errors for LID.

For normal errors, the LID, LID II and LIGPD estimators show comparable or higher efficiency than the QR estimator for all three sample sizes. One possible explanation is that the QR method fits quantile regression at each quantile level separately without making any distributional assumptions or attempting to model the distribution, while the LID, LID II and LIGPD aim to estimate the conditional distribution of Y by using the additional assumption of global linearity of quantile functions.

For t_3 errors, the LID, LID II and LIGPD estimators with $K = 7$ show higher efficiency than the QR estimator for all three sample sizes. The LIGPD estimator has slight efficient gain over the LID II estimator. Because the GPD tail density relies on large sample

approximation, the LIGPD method does not have clear efficiency gain over the LID with the "optimal" σ for small samples sizes. The LIGPD estimator has some efficiency gain over the LID II estimator.

For χ^2_1 random errors, the LID, LID II and LIGPD estimators show higher efficiency than the QR estimator for $n = 100$ and 200. The LID method with the "optimal" σ was the most efficient for $n = 100$ and 200 but the LID II and LIGPD show comparable results for all sample sizes. For $n = 500$, all three estimators have similar efficiency when $K = 7$. For $K = 15$ and $n = 100$, the LIGPD estimator shows higher MSEs than the QR and LID estimators because small sample size makes poor estimation of tail probability.

Tables 2.1–2.3 show that the LIGPD and LID estimators have comparable or higher efficiency than the QR estimator, especially for moderate sample size $n = 100$ and 200. In general, results suggest that, by using GPD approximation, the LIGPD leads to similar or more efficient estimation than the LID based on the "optimal" σ for $N(0, 1)$ and t_3 errors and than the LID II estimator for all error distributions. The efficiency gain of LIGPD over LID is not obvious in small sample sizes, since the GPD approximation to tail density does not work well for small sample sizes.

To examine the sensitivity of the LID method to the choice of σ, we plot the MSEs of the QR and LIGPD estimators and the ranges of MSEs of the LID estimators based on different choices of σ in Figures 2.3–2.5. These figures suggest that not-well selected σ in the LID estimator may result in high MSE and that the LID method is very sensitive to the choice of σ in all scenarios considered. If σ is chosen inappropriately, the performance of LID could be much worse than LIGPD and even worse than the regular QR method in some cases. Note that in our implementation of the LID, we choose σ involved in f_1 and f_2 among few choices that gives LID estimates with the smallest mean squared errors. This option is in favor of the LID method. However, in practice this choice is not feasible.

Quantile Crossing Issue

When we estimate multiple conditional quantiles, two estimated quantile functions may cross each other. This problem is referred to as the quantile crossing problem in the literature. The conventional quantile regression method can have this issue because it estimates each conditional quantile function separately at each quantile level τ_k, $k = 1, ..., K$. However, the LID and LIGPD estimates avoid the quantile crossing problem by simultaneously estimating quantile coefficients across quantiles and by the constraints

Table 2.1: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim N(0,1)$ from the conventional quantile regression method (QR), the LID method of Feng (LID) based on the "optimal" σ, the LID method with the proposed variance parameter selection method (LID II) and the proposed LIGPD method using GPD tail density approximation (LIGPD), where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	Method	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
100	QR	-	0.228	0.176	0.220	0.155	0.227	0.189	0.206	0.166
	LID	7	0.130	0.119	0.178	0.064	0.177	0.135	0.181	0.055
		11	0.208	0.136	0.175	0.055	0.197	0.150	0.160	0.040
		15	0.281	0.178	0.234	0.059	0.259	0.176	0.235	0.062
	LID II	7	0.180	0.138	0.134	0.039	0.207	0.155	0.183	0.052
		11	0.221	0.147	0.186	0.046	0.230	0.160	0.179	0.048
		15	0.281	0.173	0.220	0.060	0.276	0.187	0.220	0.061
	LIGPD	7	0.174	0.128	0.136	0.043	0.168	0.124	0.131	0.049
		11	0.205	0.141	0.166	0.046	0.185	0.143	0.155	0.041
		15	0.282	0.169	0.233	0.062	0.243	0.163	0.199	0.056
200	QR	-	0.096	0.080	0.105	0.092	0.096	0.076	0.098	0.093
	LID	7	0.065	0.037	0.059	0.027	0.082	0.059	0.061	0.045
		11	0.092	0.055	0.069	0.038	0.073	0.049	0.070	0.028
		15	0.084	0.060	0.081	0.029	0.090	0.061	0.087	0.034
	LID II	7	0.090	0.060	0.083	0.038	0.082	0.055	0.076	0.039
		11	0.094	0.061	0.075	0.029	0.102	0.059	0.083	0.036
		15	0.090	0.062	0.077	0.030	0.083	0.058	0.081	0.032
	LIGPD	7	0.092	0.057	0.082	0.043	0.076	0.051	0.063	0.037
		11	0.090	0.064	0.076	0.035	0.088	0.055	0.073	0.035
		15	0.093	0.063	0.078	0.028	0.085	0.058	0.077	0.030
500	QR	-	0.051	0.034	0.044	0.034	0.049	0.033	0.048	0.034
	LID	7	0.048	0.036	0.045	0.021	0.051	0.033	0.039	0.020
		11	0.048	0.036	0.047	0.020	0.042	0.033	0.044	0.014
		15	0.044	0.037	0.045	0.014	0.049	0.040	0.054	0.018
	LID II	7	0.046	0.035	0.050	0.021	0.050	0.035	0.048	0.020
		11	0.041	0.034	0.045	0.016	0.047	0.034	0.049	0.018
		15	0.046	0.039	0.047	0.015	0.047	0.041	0.057	0.017
	LIGPD	7	0.043	0.035	0.053	0.023	0.044	0.033	0.046	0.020
		11	0.041	0.035	0.044	0.018	0.041	0.033	0.045	0.018
		15	0.041	0.034	0.040	0.014	0.042	0.035	0.044	0.014

Table 2.2: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim t_3$ from the conventional quantile regression method (QR), the LID method of Feng (LID) based on the "optimal" σ, the LID method with the proposed variance parameter selection method (LID II) and the proposed LIGPD method using GPD tail density approximation (LIGPD), where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	Method	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
100	QR	-	0.371	0.220	0.330	0.246	0.362	0.208	0.331	0.247
	LID	7	0.276	0.160	0.286	0.160	0.265	0.152	0.218	0.082
		11	0.388	0.211	0.352	0.128	0.423	0.210	0.290	0.086
		15	0.469	0.253	0.517	0.202	0.588	0.270	0.460	0.168
	LID II	7	0.315	0.195	0.284	0.103	0.329	0.200	0.269	0.072
		11	0.429	0.257	0.431	0.143	0.478	0.259	0.375	0.098
		15	0.563	0.254	0.503	0.204	0.578	0.255	0.460	0.181
	LIGPD	7	0.273	0.170	0.254	0.100	0.248	0.163	0.219	0.061
		11	0.361	0.221	0.372	0.152	0.408	0.224	0.293	0.091
		15	0.622	0.270	0.507	0.206	0.590	0.275	0.461	0.154
200	QR	-	0.171	0.098	0.170	0.113	0.179	0.095	0.163	0.103
	LID	7	0.067	0.056	0.104	0.042	0.126	0.084	0.154	0.042
		11	0.234	0.117	0.202	0.078	0.210	0.131	0.209	0.075
		15	0.204	0.115	0.190	0.066	0.208	0.126	0.175	0.060
	LID II	7	0.132	0.078	0.127	0.056	0.153	0.102	0.158	0.054
		11	0.159	0.095	0.154	0.053	0.176	0.109	0.147	0.041
		15	0.174	0.108	0.175	0.051	0.184	0.118	0.165	0.051
	LIGPD	7	0.136	0.074	0.136	0.061	0.118	0.083	0.122	0.041
		11	0.141	0.085	0.143	0.059	0.134	0.086	0.112	0.046
		15	0.170	0.100	0.158	0.055	0.172	0.099	0.147	0.056
500	QR	-	0.059	0.038	0.061	0.044	0.060	0.044	0.067	0.046
	LID	7	0.047	0.032	0.045	0.024	0.070	0.045	0.056	0.026
		11	0.053	0.036	0.058	0.024	0.055	0.041	0.061	0.025
		15	0.079	0.034	0.061	0.029	0.066	0.042	0.070	0.027
	LID II	7	0.058	0.032	0.050	0.024	0.064	0.040	0.061	0.029
		11	0.061	0.034	0.057	0.024	0.059	0.042	0.062	0.026
		15	0.053	0.032	0.052	0.019	0.051	0.040	0.059	0.020
	LIGPD	7	0.052	0.027	0.047	0.028	0.050	0.032	0.049	0.028
		11	0.049	0.029	0.056	0.027	0.048	0.034	0.051	0.024
		15	0.049	0.030	0.049	0.022	0.042	0.033	0.050	0.022

Table 2.3: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim \chi^2_1$ from the conventional quantile regression method (QR), the LID method of Feng (LID) based on the "optimal" σ, the LID method with the proposed variance parameter selection method (LID II) and the proposed LIGPD method using GPD tail density approximation (LIGPD), where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	Method	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
100	QR	-	0.017	0.114	0.615	0.383	0.018	0.129	0.594	0.380
	LID	7	0.008	0.034	0.287	0.204	0.006	0.035	0.242	0.165
		11	0.003	0.067	0.656	0.380	0.004	0.076	0.541	0.277
		15	0.007	0.177	1.610	0.819	0.007	0.150	1.065	0.507
	LID II	7	0.013	0.075	0.531	0.292	0.007	0.073	0.402	0.205
		11	0.020	0.181	1.152	0.511	0.012	0.143	0.815	0.353
		15	0.032	0.323	2.287	1.038	0.021	0.263	1.583	0.673
	LIGPD	7	0.018	0.100	0.547	0.259	0.007	0.077	0.449	0.231
		11	0.020	0.168	1.147	0.531	0.012	0.140	0.780	0.334
		15	0.034	0.327	2.314	1.023	0.022	0.287	1.736	0.726
200	QR	-	0.008	0.064	0.354	0.221	0.008	0.058	0.356	0.227
	LID	7	0.005	0.018	0.110	0.098	0.005	0.022	0.116	0.073
		11	0.002	0.021	0.177	0.121	0.002	0.015	0.142	0.100
		15	0.002	0.034	0.307	0.184	0.001	0.027	0.233	0.139
	LID II	7	0.007	0.029	0.167	0.117	0.004	0.023	0.139	0.091
		11	0.007	0.047	0.277	0.149	0.003	0.031	0.220	0.120
		15	0.009	0.073	0.484	0.237	0.006	0.055	0.358	0.179
	LIGPD	7	0.007	0.029	0.167	0.116	0.004	0.026	0.172	0.109
		11	0.007	0.050	0.296	0.161	0.004	0.037	0.280	0.159
		15	0.011	0.087	0.517	0.237	0.007	0.063	0.380	0.174
500	QR	-	0.002	0.025	0.121	0.080	0.003	0.023	0.121	0.083
	LID	7	0.006	0.028	0.117	0.063	0.006	0.029	0.133	0.069
		11	0.003	0.012	0.069	0.048	0.003	0.013	0.084	0.053
		15	0.001	0.012	0.083	0.053	0.001	0.010	0.082	0.056
	LID II	7	0.007	0.029	0.118	0.064	0.006	0.026	0.126	0.064
		11	0.003	0.015	0.086	0.053	0.002	0.013	0.087	0.053
		15	0.003	0.020	0.119	0.066	0.001	0.011	0.082	0.055
	LIGPD	7	0.007	0.029	0.110	0.058	0.005	0.024	0.101	0.052
		11	0.004	0.020	0.116	0.067	0.002	0.012	0.101	0.068
		15	0.003	0.021	0.123	0.066	0.002	0.014	0.100	0.062

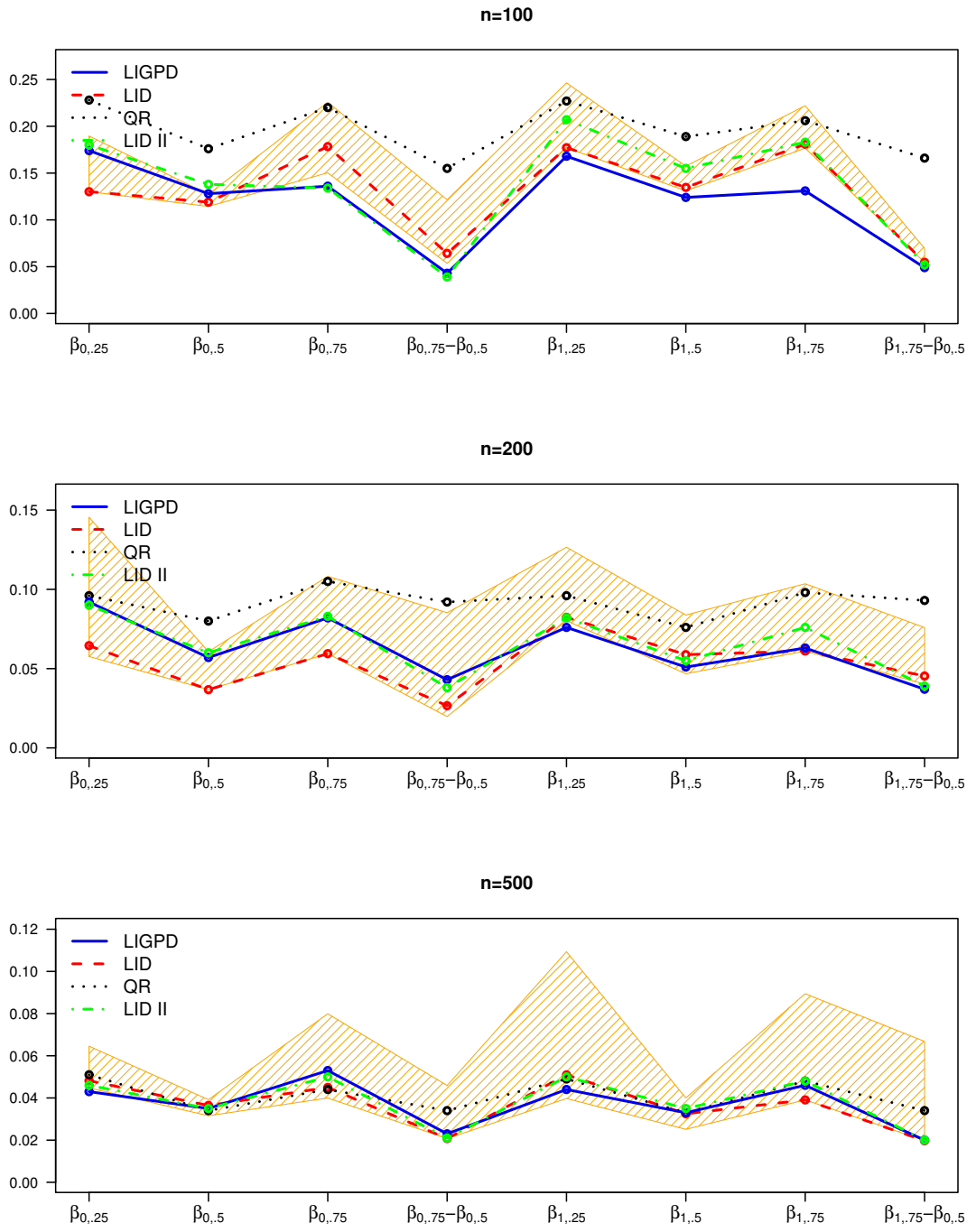


Figure 2.3: Mean squared errors of the QR, LID, LID II and LIGPD estimators for data generated from model (2.9) with $\epsilon_i \sim N(0,1)$. The shaded bands represent the range of MSEs of the LID estimator with $\sigma = 0.5, 1, 1.5, 2, 3$.

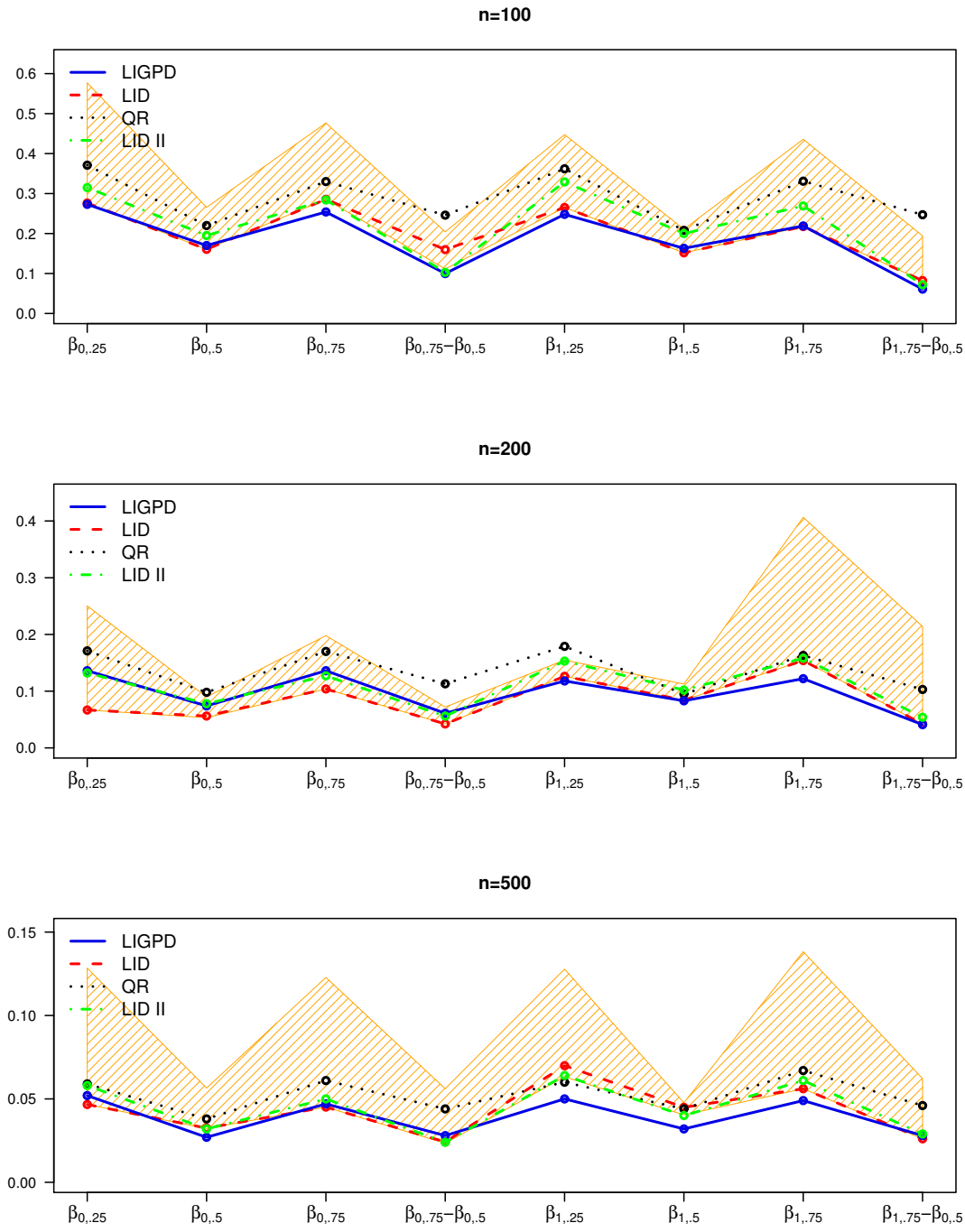


Figure 2.4: Mean squared errors of the QR, LID, LID II and LIGPD estimators for data generated from model (2.9) with $\epsilon_i \sim t_3$. The shaded bands represent the range of MSEs of the LID estimator with $\sigma = 0.5, 1, 1.5, 2, 3$.

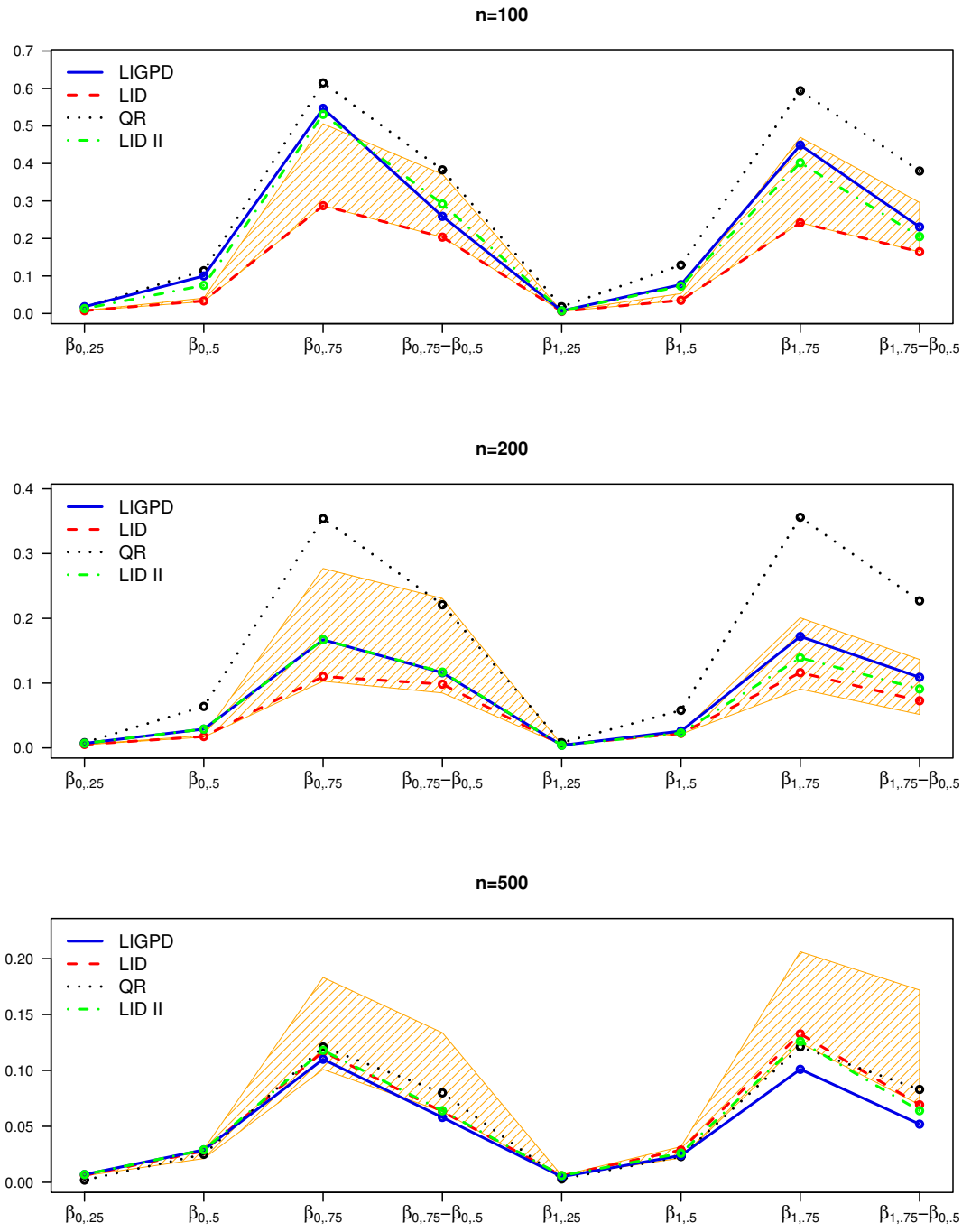


Figure 2.5: Mean squared errors of the QR, LID, LID II and LIGPD estimators for data generated from model (2.9) with $\epsilon_i \sim \chi^2_1$. The shaded bands represent the range of MSEs of the LID estimator with $\sigma_1 = 0.5$ and $\sigma_2 = 1, 2, 3, 4, 5$.

Table 2.4: Quantile crossing rates in percentage (%) for data generated from model (2.9) with $\epsilon_i \sim N(0,1)$, t_3, and χ^2_1 from the conventional quantile regression method.

		Sample size		
Error	K	100	200	500
$N(0,1)$	7	2.7	3.4	3.1
	11	7.5	8.8	9.1
	15	15.0	14.7	15.6
t_3	7	0.5	0.9	1.3
	11	2.2	3.1	4.3
	15	5.9	7.5	7.4
χ^2_1	7	0.0	0.0	0.0
	11	0.3	0.2	0.4
	15	0.8	0.7	1.3

posed in Step 2 of the procedures.

Table 2.4 summarizes the quantile crossing rates for the conventional quantile regression method. The quantile crossing rates were obtained by

$$\frac{1}{200}\sum_{j=1}^{200}\left(1-\frac{\sum_{i=1}^{n} I\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_1} < \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_2} < \cdots < \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_k}\}}{n}\right), \tag{2.11}$$

where j indicates the j-th simulated data set and $\sum_{i=1}^{n} I\{\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_1} < \cdots < \mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_k}\}$ is the number of observations that do not violate quantile monotonicity. As expected, we do not observe any quantile crossing for the LID and LIGPD estimates. However, for the QR estimates, we observe positive quantile crossing rates, ranging from 0.0002% to 15%. The quantile crossing rate of the QR method increases when more quantile levels are involved.

Credible Intervals and Coverage Probability

To construct 95% credible intervals, we considered two different methods based on MCMC sample quantiles and MCMC sample variance. Table 2.5 shows the coverage probabilities of 95% confidence intervals of the QR method and 95% credible intervals constructed by MCMC sample quantiles and variances of the LIGPD method. The credible intervals constructed by MCMC sample variances are based on the normality assumption. For

Table 2.5: Coverage probabilities (%) of 95% credible intervals (LIGPD) and 95% confidence intervals (QR), where $x_i \sim LN(0,1)$, δ_0=$\beta_{0,.75}$ - $\beta_{0,.5}$ and δ_1=$\beta_{1,.75}$ - $\beta_{1,.5}$.

Method	ϵ_i	n	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
QR	$N(0,1)$	100	0.89	0.89	0.87	0.94	0.94	0.94	0.94	0.95
		200	0.89	0.88	0.90	0.95	0.93	0.94	0.96	0.95
		500	0.89	0.91	0.88	0.96	0.95	0.95	0.95	0.95
	t_3	100	0.88	0.87	0.87	0.96	0.92	0.93	0.94	0.95
		200	0.89	0.90	0.88	0.95	0.94	0.92	0.94	0.94
		500	0.88	0.89	0.88	0.96	0.94	0.94	0.94	0.96
LIGPD	$N(0,1)$	100	0.91	0.91	0.91	0.99	0.91	0.92	0.93	1.00
Posterior		200	0.87	0.90	0.90	0.98	0.91	0.91	0.92	0.97
variance		500	0.83	0.85	0.82	0.94	0.87	0.88	0.87	0.95
	t_3	100	0.93	0.90	0.92	0.99	0.92	0.92	0.93	1.00
		200	0.92	0.90	0.89	0.98	0.91	0.90	0.91	0.98
		500	0.85	0.90	0.85	0.95	0.86	0.87	0.86	0.93
LIGPD	$N(0,1)$	100	0.92	0.92	0.93	1.00	0.91	0.92	0.94	1.00
Quantile		200	0.89	0.90	0.91	0.98	0.91	0.92	0.92	0.98
based		500	0.86	0.88	0.85	0.94	0.90	0.90	0.88	0.96
	t_3	100	0.94	0.91	0.93	1.00	0.93	0.91	0.94	1.00
		200	0.92	0.91	0.90	0.98	0.92	0.92	0.90	0.99
		500	0.86	0.91	0.86	0.97	0.89	0.87	0.88	0.94

example, 95% credible interval for $\beta_{1,\tau}$ is $(\bar{\beta}_{1,\tau} - 1.96 \times \text{sd}(\hat{\beta}_{1,\tau}), \hat{\beta}_{1,\tau} + 1.96 \times \text{sd}(\hat{\beta}_{1,\tau}))$, $\bar{\beta}_{1,\tau}$ and $\text{sd}(\hat{\beta}_{1,\tau})$ are the posterior sample mean and standard deviation, respectively.

For $n = 100$ and 200, coverage probabilities of 95% credible intervals are comparable with coverage probabilities of 95% confidence intervals constructed by the conventional quantile regression method. For interquantile differences, coverage probabilities are higher with shorter credible intervals than the QR. We notice that the coverage probabilities are not getting better as sample size increases since more sample size brings more restriction on the proposal distributions. It leads to the underestimation of posterior variances and narrower quantile-based credible intervals.

2.4.2 Simulation Study II

We consider a model with two predictors. The data were generated from the following linear model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + (1 + x_{1i} + x_{2i})\epsilon_i,\ i = 1, ..., n, \tag{2.12}$$

where $x_{1i} \sim LN(0, 1)$, $x_{2i} \sim Bernoulli(0, 5)$, $\beta_0 = 5$, $\beta_1 = 1$, $\beta_2 = 1$ and $\epsilon_i \sim N(0, 1)$ or t_4. The corresponding quantile regression model is

$$Q_\tau(Y|\mathbf{x}_i) = \beta_{0,\tau} + \beta_{1,\tau} x_{1i} + \beta_{2,\tau} x_{2i}, \tag{2.13}$$

where $\beta_{0,\tau} = 5 + F_e^{-1}(\tau)$, $\beta_{1,\tau} = 1 + F_e^{-1}(\tau)$ and $\beta_{2,\tau} = 1 + F_e^{-1}(\tau)$ with F_e denoting the CDF of ϵ_i.

We generate $n = 100, 200$ and 500 observations from model (2.12). For the standard normal random errors, we use a grid of $K = 7, 11$ and 15 equally spaced quantile levels $\tau_k = k/(1 + K), k = 1, \cdots, K$, and assign the normal priors $N(\mathbf{0}, \Sigma)$ to $(\boldsymbol{\beta}_{\tau_1}, \cdots, \boldsymbol{\beta}_{\tau_K})^T$, where $\boldsymbol{\beta}_{\tau_k} = (\beta_{0,\tau_k}, \beta_{1,\tau_k}, \beta_{2,\tau_k})^T$ and $\Sigma = \text{diag}(100, \ldots, 100)$. With 100,000 MCMC samples, we discard the first 50,000 as burn-in, and estimate $\beta_{j,\tau}$ for $j = 0, 1$ by the posterior means of the thinned sample using every 100th draw of the MCMC chain. The simulation is repeated 200 times for each σ considered. For t_4 errors, K is set as seven.

Similar to Section 2.4.1, we report mean squared errors of the QR, LID and LIGPD estimators in Tables 2.6–2.7 and Figures 2.6 and 2.7. Note that the LID is based on the "optimal" σ that gives the smallest sum of the mean squared errors among $\sigma = \{0.5, 1, 1.5, 2, 3\}$ for the standard normal and t_4 errors.

There are some similar patterns in MSE for both the standard normal and t_4 errors. For $n = 100$ and 200, the LID and LIGPD estimators with $K = 7$ show clear efficiency gain over the QR estimator for estimating β_0 and β_1 and similar performance for estimating β_2. The results for $n = 100$ are sensitive to the number of quantile levels, K, because the LID and LIGPD methods have 21, 33, or 45 quantile coefficients to estimate. For moderate sample size $n = 200$ and large sample size $n = 500$, the results are less sensitive to the choice of K.

In sum, through simulation studies I and II, we observe that the LID and LIGPD methods perform generally better than the conventional quantile regression in terms of efficiency. Especially, for heavy-tailed distributions such as t distributions, the LIGPD

Table 2.6: Mean squared errors of parameter estimators for data generated from model (2.12) with $\epsilon_i \sim N(0,1)$ from the conventional quantile regression method (QR), the LID method of Feng (LID) based on the "optimal" σ, the proposed LID method (LIGPD), where δ_0=$\beta_{0,.75} - \beta_{0,.5}$, δ_1=$\beta_{1,.75} - \beta_{1,.5}$ and δ_2=$\beta_{2,.75} - \beta_{2,.5}$.

n	Method	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1	$\beta_{2,.25}$	$\beta_{2,.5}$	$\beta_{2,.75}$	δ_2
100	QR	-	0.32	0.28	0.34	0.28	0.23	0.19	0.28	0.20	0.55	0.50	0.59	0.35
	LID	7	0.38	0.24	0.26	0.11	0.17	0.15	0.18	0.04	0.63	0.47	0.54	0.22
		11	0.81	0.29	0.37	0.27	0.24	0.21	0.27	0.04	0.86	0.56	0.66	0.28
		15	1.56	0.40	0.50	0.40	0.39	0.28	0.42	0.08	1.13	0.58	0.89	0.36
	LIGPD	7	0.39	0.24	0.25	0.12	0.17	0.16	0.18	0.03	0.64	0.44	0.51	0.21
		11	0.77	0.33	0.37	0.23	0.20	0.19	0.22	0.03	0.82	0.51	0.58	0.22
		15	1.49	0.40	0.54	0.39	0.46	0.29	0.40	0.07	1.13	0.61	0.79	0.32
200	QR	-	0.16	0.15	0.17	0.12	0.13	0.11	0.13	0.09	0.22	0.22	0.27	0.21
	LID	7	0.12	0.10	0.14	0.05	0.12	0.10	0.10	0.04	0.28	0.24	0.31	0.14
		11	0.19	0.11	0.18	0.10	0.09	0.08	0.10	0.02	0.29	0.25	0.32	0.14
		15	0.25	0.15	0.22	0.09	0.11	0.10	0.12	0.02	0.32	0.28	0.35	0.13
	LIGPD	7	0.13	0.10	0.16	0.07	0.09	0.08	0.09	0.03	0.29	0.24	0.31	0.16
		11	0.13	0.09	0.15	0.06	0.09	0.08	0.09	0.03	0.25	0.23	0.30	0.12
		15	0.20	0.13	0.19	0.09	0.11	0.10	0.11	0.02	0.36	0.28	0.32	0.12
500	QR	-	0.06	0.06	0.07	0.05	0.05	0.05	0.05	0.04	0.10	0.07	0.10	0.07
	LID	7	0.06	0.05	0.07	0.02	0.05	0.04	0.05	0.03	0.13	0.09	0.12	0.07
		11	0.08	0.05	0.09	0.04	0.05	0.04	0.05	0.02	0.13	0.10	0.13	0.07
		15	0.07	0.05	0.08	0.03	0.05	0.04	0.05	0.01	0.11	0.10	0.12	0.06
	LIGPD	7	0.07	0.06	0.08	0.03	0.04	0.04	0.04	0.02	0.12	0.09	0.13	0.08
		11	0.06	0.05	0.07	0.03	0.04	0.04	0.04	0.02	0.11	0.09	0.12	0.07
		15	0.07	0.06	0.08	0.02	0.05	0.05	0.05	0.01	0.12	0.10	0.14	0.06

shows clear efficiency gain. However, for small sample size $n = 100$, the LIGPD estimator sometimes perform worse than the LID method based on the "optimal" σ and the QR method because of poor tail estimation. Both LID and LIGPD methods can avoid the quantile crossing problem. The LIGPD method does not require pre-specifiy the parameters in the tail densities and thus is more stable than the LID method, which may give poor estimates when σ is chosen inappropriately.

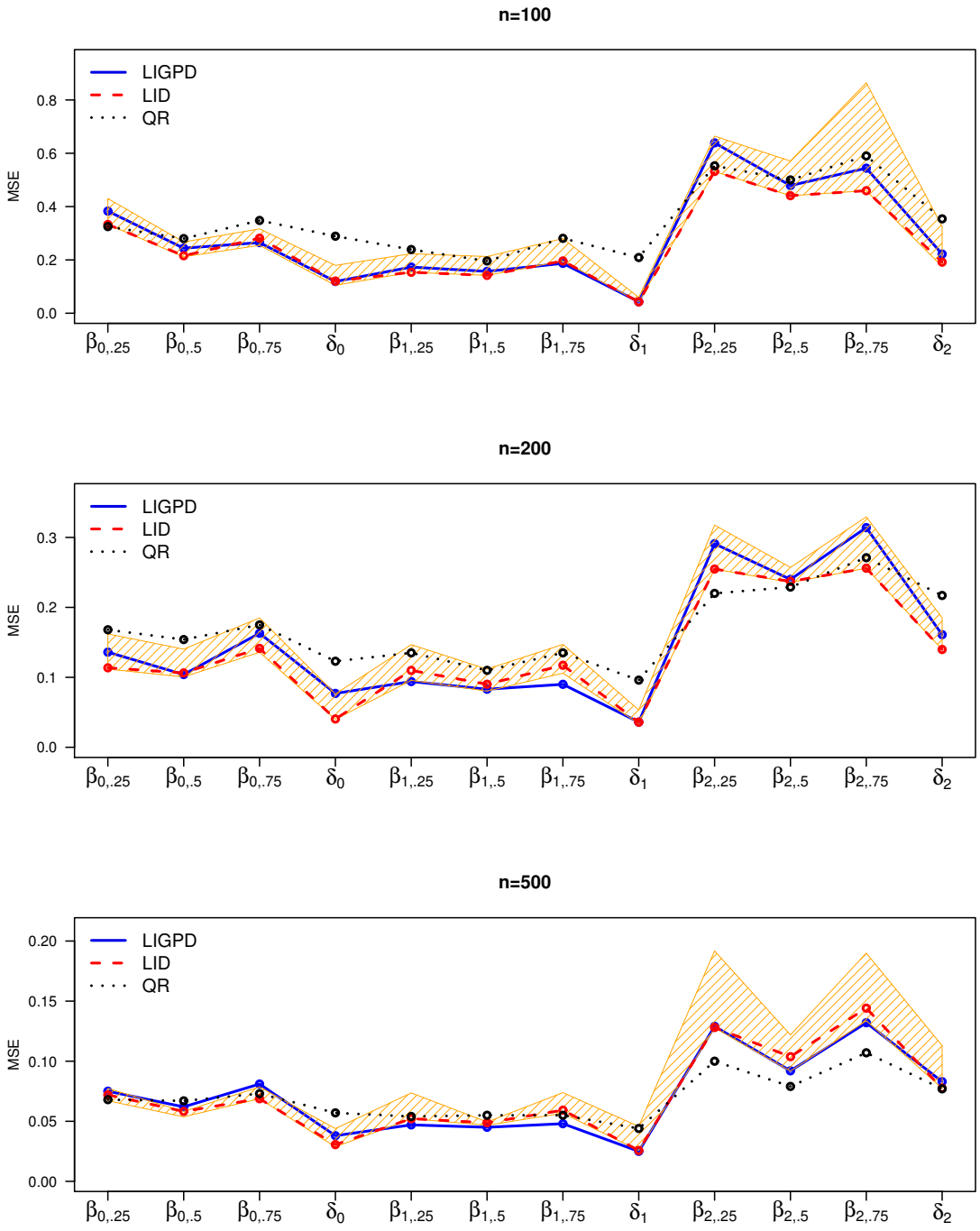


Figure 2.6: MSE of the QR, LID and LIGPD estimators for data generated from model (2.12) with $\epsilon_i \sim N(0,1)$, where δ_0=$\beta_{0,.75} - \beta_{0,.5}$, δ_1=$\beta_{1,.75} - \beta_{1,.5}$ and δ_2=$\beta_{2,.75} - \beta_{2,.5}$. The shaded bands represent the range of MSEs of the LID estimator with $\sigma = 0.5, 1, 1.5, 2, 3$.

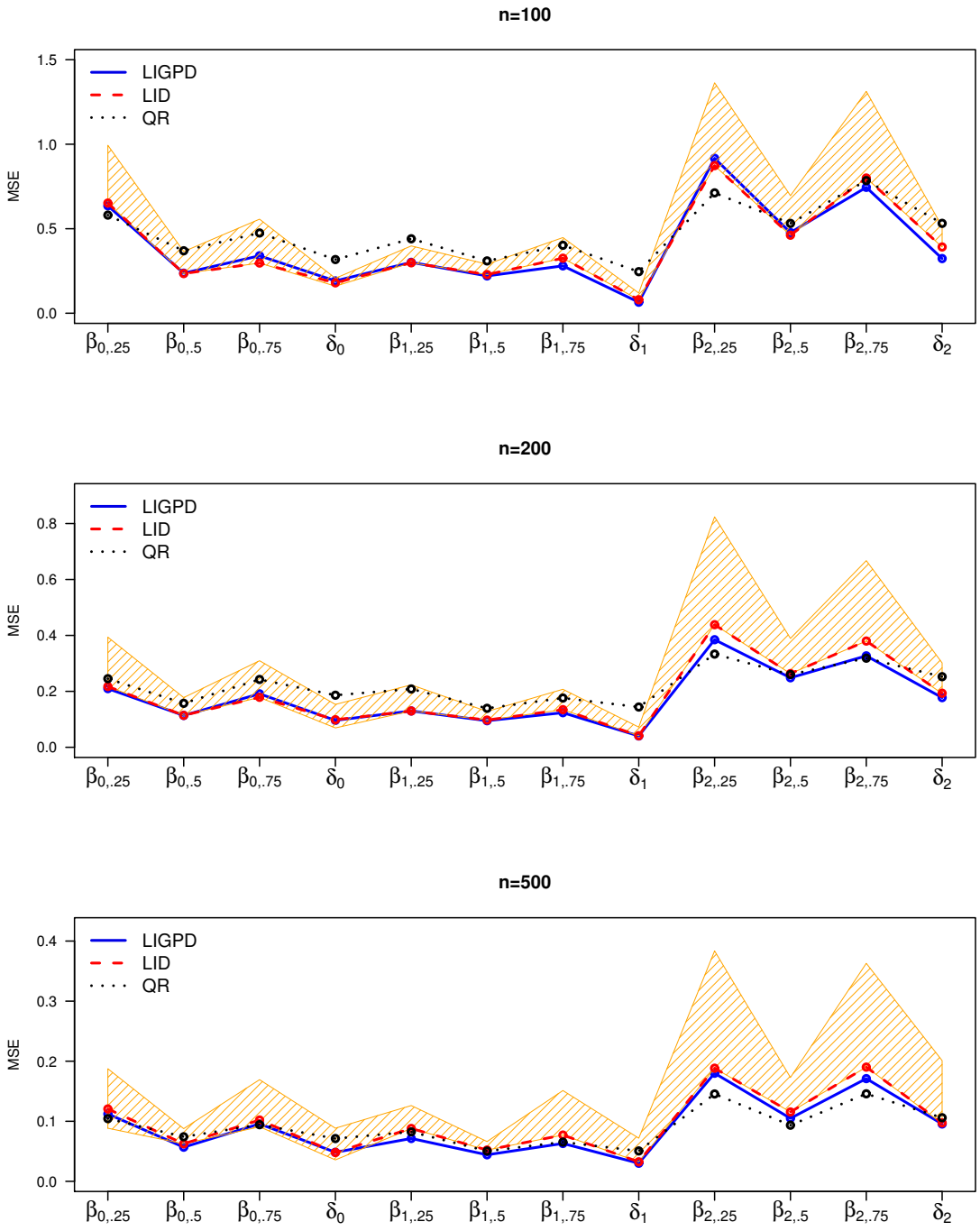


Figure 2.7: MSE of the QR, LID and LIGPD estimators for data generated from model (2.12) with $\epsilon_i \sim t_4$, where $\delta_0 = \beta_{0,.75} - \beta_{0,.5}$, $\delta_1 = \beta_{1,.75} - \beta_{1,.5}$ and $\delta_2 = \beta_{2,.75} - \beta_{2,.5}$. The shaded bands represent the range of MSEs of the LID estimator with $\sigma = 0.5, 1, 1.5, 2, 3$.

Table 2.7: Mean squared errors of parameter estimators for data generated from model (2.12) with $\epsilon_i \sim t_4$ from the conventional quantile regression method (QR), the LID method of Feng (LID) based on the "optimal" σ, the proposed LID method (LIGPD), where δ_0=$\beta_{0,.75} - \beta_{0,.5}$, δ_1=$\beta_{1,.75} - \beta_{1,.5}$ and δ_2=$\beta_{2,.75} - \beta_{2,.5}$.

n	Method	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1	$\beta_{2,.25}$	$\beta_{2,.5}$	$\beta_{2,.75}$	δ_2
100	QR	-	0.58	0.37	0.47	0.31	0.44	0.31	0.40	0.24	0.71	0.53	0.78	0.53
	LID	7	0.65	0.23	0.29	0.18	0.29	0.22	0.32	0.08	0.87	0.46	0.80	0.39
	LIGPD	7	0.63	0.23	0.34	0.19	0.30	0.22	0.28	0.06	0.91	0.47	0.74	0.32
200	QR	-	0.24	0.15	0.24	0.18	0.20	0.13	0.17	0.14	0.33	0.26	0.31	0.25
	LID	7	0.21	0.11	0.17	0.09	0.13	0.09	0.13	0.04	0.43	0.26	0.38	0.19
	LIGPD	7	0.21	0.11	0.19	0.09	0.13	0.09	0.12	0.04	0.38	0.24	0.32	0.17
500	QR	-	0.10	0.07	0.09	0.07	0.08	0.05	0.06	0.05	0.14	0.09	0.14	0.10
	LID	7	0.12	0.06	0.10	0.04	0.08	0.05	0.07	0.03	0.18	0.11	0.19	0.09
	LIGPD	7	0.11	0.05	0.09	0.04	0.07	0.04	0.06	0.03	0.18	0.10	0.17	0.09

Chapter 3

LID for Quantile Regression Models with Random Effects

In this chapter, we propose a semiparametric Bayesian method for quantile regression with random effects. We approximate the likelihood using the LIGPD method and propose a Metropolis-within-Gibbs algorithm to update fixed and random effects. The proposed algorithm avoids the quantile crossing problem, and yields the joint posterior distribution of quantile coefficients at multiple quantiles. Through simulation studies, we demonstrate that by approximating the likelihood through information-sharing across quantiles, the proposed method leads to more efficient multiple-quantile estimation than existing methods for quantile regression with random effects in finite samples.

3.1 Review on Bayesian Quantile Regression Models with Random Effects

As mentioned in Chapter 2, Bayesian quantile regression with random effects is challenging without any parametric likelihood assumptions. In the literature, there exist several likelihood estimation methods in parametric and semiparametric frameworks.

Yu and Moyeed (2001) proposed the asymmetric Laplace distribution as a working likelihood for Bayesian quantile regression for independent data. Geraci and Bottai (2007) and Yuan and Yin (2010) extended the asymmetric Laplace distribution idea to clustered data.

Geraci and Bottai (2007) considered a random intercept model at a given quantile level $\tau \in (0,1)$:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + b_i + e_{ij},\ i = 1, ..., n, j = 1, ..., n_i,$$

where y_{ij} are responses, $\mathbf{x}_{ij}$ are p-dimensional covariate vectors, $\boldsymbol{\beta}_\tau$ is a p-dimensional fixed effect, b_i are random subject effects, e_{ij} are independent random errors, n is the number of clusters and n_i is the number of subjects in the i-th cluster. Geraci and Bottai (2007) assumed that, given b_i, y_{ij} are independent of each other and follow the asymmetric Laplace distribution $ALD(\tau, \sigma, \mu_{ij})$:

$$f(y_{ij}|\tau, \sigma, \mu_{ij}) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\rho_\tau\ \left(\frac{y_{ij} - \mu_{ij}}{\sigma}\right)\right\},$$

where $\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + b_i$ and σ is the scale parameter. Thus, the corresponding quantile regression model is

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, b_i) = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + b_i.$$

In addition, Geraci and Bottai (2007) also assumed that b_i are independent of e_{ij} and they are identically distributed with some known density f_b, which is parameterized by ψ. The random effect distribution is assumed to be symmetric and centered at zero, for instance, normal or Laplace distributions. Geraci and Bottai (2007) treated the random effect b_i as missing data and used a Monte Carlo Expectation Maximization (EM) algorithm to approximate the marginal likelihood. In the E-step, the expected value of the log likelihood for the i-th cluster at the $(t+1)$-th iteration is defined by

$$\begin{aligned}
&Q_i\{\boldsymbol{\beta}_\tau, \sigma, \psi | \boldsymbol{\beta}_\tau^{(t)}, \sigma^{(t)}, \psi^{(t)}\} \\
&= \int \left\{ \sum_{j=1}^{n_i} \log f(y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}_\tau, \sigma, b_i) + \log f_b(b_i|\psi) \right\} f\left\{ b_i | y_{ij}, \mathbf{x}_{ij}, \boldsymbol{\beta}_\tau^{(t)}, \sigma^{(t)}, \psi^{(t)} \right\} db_i,
\end{aligned}$$

where $f\{y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}_\tau, \sigma, b_i\}$ is the pdf of $ALD(\tau, \sigma, \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + b_i)$ and the integration was taken with respect to b_i. To approximate the Q function, Geraci and Bottai (2007) adopted a Monte Carlo integration technique. Since the random effects b_i are not observable, k

samples of b_i, say $b_{i1}, ..., b_{ik}$, were generated from the following distribution

$$f\left\{b_i|y_{ij}, \mathbf{x}_{ij}, \boldsymbol{\beta}_\tau^{(t)}, \sigma^{(t)}, \psi^{(t)}\right\} \propto f\left\{y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}_\tau^{(t)}, \sigma^{(t)}, b_i\right\} f\left\{b_i|\psi^{(t)}\right\}$$

and the approximate expected log likelihood function was obtained by

$$Q\left\{\boldsymbol{\beta}_\tau, \sigma, \psi|\boldsymbol{\beta}_\tau^{(t)}, \sigma^{(t)}, \psi^{(t)}\right\} = \sum_{i=1}^{n} \frac{1}{k} \sum_{l=1}^{k} \sum_{j=1}^{n_i} \left[\log f\left\{y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}_\tau, \sigma, b_{il}\right\} + \log f(b_{il}|\psi)\right].$$

In the M-step, Geraci and Bottai (2007) used an iterative algorithm to find estimates of $\boldsymbol{\beta}_\tau$, σ and ψ.

Geraci and Bottai (2013) extended the work in Geraci and Bottai (2007) to a more general quantile regression model, which is able to accommodate multiple random effects. The model in Geraci and Bottai (2013) included the random intercept model in Geraci and Bottai (2007) as a special case. Two papers differ slightly in how to approximate the marginal likelihood and estimate the parameters.

Geraci and Bottai (2013) considered the following linear regression model

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}, \ i = 1, ..., n, j = 1, ..., n_i,$$

where y_{ij} are responses, $\mathbf{x}_{ij}$ are p-dimensional covariate vectors for the fixed effects, $\mathbf{z}_{ij}$ are q-dimensional covariate vectors for the cluster-specific random effects, $\boldsymbol{\beta}_\tau$ is a p-dimensional fixed effect, $\mathbf{b}_i$ are q-dimensional random effects and ϵ_{ij} are $i.i.d$ random errors that are independent of $\mathbf{b}_i$. Geraci and Bottai (2013) assumed that y_{ij} are independent of each other and follow the asymmetric Laplace distribution $ALD(\tau, \sigma, \mu_{ij})$ with $\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T \mathbf{b}_i$. The corresponding conditional linear quantile regression model is

$$Q_\tau(Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T \mathbf{b}_i.$$

Similar to Geraci and Bottai (2007), the random effect distribution, $f_{\mathbf{b}}$, is assumed to be symmetric and centered at zero. For the random effect distribution, Geraci and Bottai (2013) considered a multivariate normal distribution with mean $\mathbf{0}$ and covariance Ψ or q independent Laplace distributions with the scale parameters $\psi_1, ..., \psi_q$. Then, the marginal density of Y can be obtained after integrating out the q-dimensional random

effects. That is,

$$f\left\{y|\boldsymbol{\beta}_\tau, \sigma, \mathbf{x}, \Psi\right\} = \int_{\mathbb{R}^q} f\{y|\mathbf{x}, \boldsymbol{\beta}_\tau, \sigma, \mathbf{b}\} f_{\mathbf{b}}(\mathbf{b}|\Psi) d\mathbf{b},$$

where $f\{y|\mathbf{x}, \boldsymbol{\beta}_\tau, \sigma, \mathbf{b}\}$ is the pdf of $ALD(\tau, \sigma, \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T\mathbf{b}_i)$. Geraci and Bottai (2013) used Gaussian quadrature approximation to evaluate the integral. Because the check loss function is not differentiable at zero, they used a non-smooth optimization technique to obtain maximum likelihood estimates for $\boldsymbol{\beta}_\tau$, σ and Ψ. The algorithm was implemented in the R package *lqmm*.

Yuan and Yin (2010) proposed to minimize the following L_2-penalized objective function to shrink the subject-specific intercepts and slopes toward the common population values,

$$\sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_\tau\{y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau - \mathbf{z}_{ij}^T\mathbf{b}_i\} + \frac{1}{2}\sum_{i=1}^{n} \mathbf{b}_i^T\Sigma^{-1}\mathbf{b}_i,$$

where Σ is a symmetric non-singular matrix and controls the amount of shrinkage in the second term. This is equivalent to the conditional linear quantile regression model with the multivariate normal random effect with mean $\mathbf{0}$ and variance Σ as discussed in Geraci and Bottai (2013). Thus, we can express the above L_2-penalization problem as the following hierarchical model

$$\begin{aligned} Y_{ij} &\overset{i.i.d}{\sim} ALD(\tau, 1, \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T\mathbf{b}_i), \\ \mathbf{b}_i &\sim N(0, \Sigma). \end{aligned}$$

The marginal likelihood of the observed data was obtained by evaluating the following integral,

$$\prod_{i=1}^{n}\int\prod_{j=1}^{n_i} f\left\{y_{ij}|\mathbf{x}_{ij}, \boldsymbol{\beta}_\tau, \sigma, \mathbf{b}_i\right\} f(\mathbf{b}_i|\Sigma) d\mathbf{b}_i.$$

Yuan and Yin (2010) assigned a flat prior for $\boldsymbol{\beta}_\tau$ and Wishart prior for Σ. They used the adaptive rejection Metropolis-within-Gibbs algorithm (Gilks et al., 1995) to update $\boldsymbol{\beta}_\tau$ and Gibbs sampler to update Σ. Note that Yuan and Yin (2010) incorporated non-ignorable missing data but we omitted the details since missing data is not the main

focus of our study.

Instead of assuming Laplace distributions as in Geraci and Bottai (2007, 2013) and Yuan and Yin (2010), Reich et al. (2010) accommodated heteroscedastic error distributions by using a more flexible distribution than the ALD, an infinite mixture of two-component Gaussian mixtures. Reich et al. (2010) considered the following location-scale shift model (He, 1997) with the random subject effects b_i,

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i + \mathbf{x}_{ij}^T\boldsymbol{\gamma}e_{ij},$$

where $\mathbf{x}_{ij}^T\boldsymbol{\gamma} > 0$ for all $\mathbf{x}_{ij}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown coefficients, and e_{ij} are $i.i.d.$ random errors whose τ-th quantiles equals zero. Reich et al. (2010) proposed a more flexible error distribution than the ALD. They defined the random error distribution as the infinite mixture

$$h(e|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{l=1}^{\infty} p_l f(e|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l^2, q_l),$$

where p_l is the mixing proportion with $\sum_{l=1}^{\infty} p_l = 1$, and the base density $f(e|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l^2, q_l)$ is the two-component normal mixture

$$f(e|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l^2, q_l) = q_l\phi(\mu_{1l}, \sigma_{1l}^2) + (1 - q_l)\phi(\mu_{2l}, \sigma_{2l}^2),$$

where $\phi(\mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. The parameters q_l are chosen to ensure $\int_{-\infty}^{0} f(e|\boldsymbol{\mu}_l, \boldsymbol{\sigma}_l^2, q_l)de = \tau$, that is, to make the τ-th quantile of e to be zero. Although Reich et al. (2010) could accommodate more general forms of error distributions, their method requires correct specification of the heteroscedasticity function in the model.

Yang and He (2012) proposed the empirical likelihood as a working likelihood for Bayesian quantile regression for independent data and this method was extended to clustered data by Kim and Yang (2011). Unlike the previously introduced methods, the empirical likelihood approach avoids modeling the error distribution.

Kim and Yang (2011) considered the following quantile regression model with random effects

$$Q_\tau(Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau + \mathbf{x}_{ij}^T\mathbf{b}_i = \mathbf{x}_{ij}^T\mathbf{b}_i^*(\tau),$$

where the random effects $\mathbf{b}_i^*(\tau)$ are cluster-specific with mean $\boldsymbol{\beta}_\tau$ and variance $\Sigma_i(\tau)$. The mean $\boldsymbol{\beta}_\tau$ is the population-averaged effect of X at the τ-th quantile of the response and $\Sigma_i(\tau)$ is the variability of the cluster-specific effects. For the i-th cluster, suppose $\widehat{\mathbf{b}}_i^*(\tau)$ is the solution to the following minimization problem,

$$\sum_{j=1}^{n_i} \rho_\tau(y_{ij} - \mathbf{x}_{ij}^T \mathbf{b}_i^*),$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$. Equivalently, $\widehat{\mathbf{b}}_i^*(\tau)$ is the solution to the following estimation equation,

$$\sum_{j=1}^{n_i} \varphi_\tau(y_{ij} - \mathbf{x}_{ij}^T \mathbf{b}_i^*)\mathbf{x}_{ij} \approx 0, \tag{3.1}$$

where $\varphi_\tau(u) = \tau - I(u < 0)$. The estimating equation (3.1) motivated the following empirical likelihood for cluster i,

$$L_i\{\mathbf{b}_i^*(\tau)\} = \max\left\{ \prod_{j=1}^{n_i} p_j \,\middle|\, \sum_{j=1}^{n_i} p_j \varphi_\tau\{y_{ij} - \mathbf{x}_{ij}^T \mathbf{b}_i^*(\tau)\}\mathbf{x}_{ij} = 0, s.t. \sum_{j=1}^{n_i} p_j = 1, 0 \le p_j \le 1 \right\}.$$

Let $g(\cdot|\boldsymbol{\theta})$ denote the density of $\mathbf{b}_i^*(\tau)$ with parameter $\boldsymbol{\theta}$. Kim and Yang (2011) proposed to estimate $\boldsymbol{\theta}$ and $\mathbf{b}_i^*(\tau)$ by

$$\widehat{\mathbf{b}}_i^*(\tau) = \arg\max_{\mathbf{b}_i^*(\tau)} L_i\{\mathbf{b}_i^*(\tau)\} g_i\{\mathbf{b}_i^*(\tau)|\boldsymbol{\theta}\},$$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \prod_{i=1}^{n} \int_{\mathbf{b}_i^*(\tau)} L_i(\mathbf{b}_i^*(\tau)) g_i(\mathbf{b}_i^*(\tau)|\boldsymbol{\theta}) d\mathbf{b}_i^*(\tau).$$

In the data analyses, Kim and Yang (2011) assumed that $\mathbf{b}_i^*(\tau)$ follow the multivariate normal distribution with mean $\boldsymbol{\beta}_\tau$ and variance $\Sigma_i(\tau) = \Sigma(\tau)$, so $\boldsymbol{\theta} = (\boldsymbol{\beta}_\tau, \Sigma(\tau))$. Kim and Yang (2011) proposed to use a Metropolis-Hastings algorithm to update $\boldsymbol{\beta}_\tau$ and a Gibbs sampler to update $\Sigma(\tau)$,which is similar to the MCMC algorithm in Yuan and Yin (2010).

3.2 LIGPD for Linear Quantile Regression with Random Effects

3.2.1 Model Setup

Suppose that we observe the clustered data $\{(y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}), i = 1, \ldots, n, j = 1, \ldots n_i\}$, where y_{ij} is the response, and $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are respectively the p- and q-dimensional covariate vectors associated with cluster i for the j-th subject (or measuring time). We assume the following conditional quantile regression model

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T\mathbf{b}_i,\ 0 < \tau < 1, \tag{3.2}$$

where $Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)$ is the τth conditional quantile of the response given covariates and the random cluster effects $\mathbf{b}_i$, $\boldsymbol{\beta}_\tau$ is a $p \times 1$ vector of fixed effects, $\mathbf{z}_{ij}$ is a $q \times 1$ covariate vector associated with the cluster-specific random effects $\mathbf{b}_i$. By including the random effects $\mathbf{b}_i$, the conditional quantile regression model (3.2) captures the cluster-specific effects of covariates on the conditional quantile of the response distribution. The fixed effect $\boldsymbol{\beta}_\tau$ is allowed to vary with the quantile level but we assume that the random effects $\mathbf{b}_i$ are the same across all quantile levels and that $\mathbf{b}_i \overset{i.i.d.}{\sim} N(0, \Sigma)$.

We can make more general distributional assumptions for $\mathbf{b}_i$ and the proposed method can be easily modified to accommodate quantile-dependent random effects. However, to do this, we need to make parametric assumptions specifying how the random effects depend on the quantile levels. This not only introduce more parameters but also the model assumptions can be difficult to verify in practice.

Instead of making any parametric assumptions on the conditional distribution of Y_{ij}, we assume in model (3.2) that the conditional quantile functions are linear for $\tau \in (0, 1)$. Under this global linearity assumption, we propose a Bayesian method based on an approximate likelihood for regression at multiple quantiles for clustered data. The idea of the approximate likelihood is to estimate the conditional density of Y_{ij} by linearly interpolating the conditional quantiles $Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)$ for a sequence of central quantiles τ, and to estimate the tail density by using extreme value theory.

Let $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_{\tau_1}^T, \ldots, \boldsymbol{\beta}_{\tau_K}^T)^T$ denote the vector of fixed quantile coefficients in model (3.2) at quantile levels $0 < \tau_1 < \cdots < \tau_K < 1$. Similarly as discussed in Section 2.2.2, given $\boldsymbol{\beta}_K$ and $\mathbf{b}_i$, we can approximate the conditional density of Y_{ij} by linearly interpolating the

conditional quantiles, $Q_{\tau_k}(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)$, at the central quantiles and by the generalized Pareto distribution at tails. Specifically, we define the approximate conditional density function as

$$\begin{aligned}\hat{f}_Y(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K) &= I\{y_{ij} \in (-\infty, \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_1} + \mathbf{z}_{ij}^T\mathbf{b}_i)\}\tau_1 f_1(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i, \boldsymbol{\beta}_K, \varrho_1, \xi_1) \\ &+ \sum_{k=1}^{K-1} I\{y_{ij} \in [\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_k} + \mathbf{z}_{ij}^T\mathbf{b}_i, \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_{k+1}} + \mathbf{z}_{ij}^T\mathbf{b}_i)\}\frac{\tau_{k+1} - \tau_k}{\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_{k+1}} - \mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_k}} \\ &+ I\{y_{ij} \in [\mathbf{x}_{ij}^T\boldsymbol{\beta}_{\tau_K} + \mathbf{z}_{ij}^T\mathbf{b}_i, \infty)\}(1 - \tau_K) f_2(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i, \boldsymbol{\beta}_K, \varrho_2, \xi_2), \end{aligned} \tag{3.3}$$

where $f_2(\cdot)$ and $f_1(\cdot)$ are the scaled generalized Pareto distributions defined in (2.5) and (2.7), respectively, and $\boldsymbol{\theta} = (\varrho_r, \varrho_l, \xi_r, \xi_l)^T$ contains all the GPD parameters. Thereafter, we refer to the approximate density (3.3) of our proposed model as LIGPD. Define $y_{ij}^* = y_{ij} - \mathbf{z}_{ij}^T\mathbf{b}_i{}^{(0)}$, where $\mathbf{b}_i^{(0)}, i = 1, \cdots, n$ are the initial estimates of the random effects. Then, we estimate the shape and scale parameters by the proposed method in Chapter 2.3 using the data $\{(y_{ij}^*, \mathbf{x}_{ij}), i = 1, \ldots, n, j = 1, \ldots n_i\}$ and the initial estimator $\tilde{\boldsymbol{\beta}}_K = (\tilde{\boldsymbol{\beta}}_{\tau_1}^T, \ldots, \tilde{\boldsymbol{\beta}}_{\tau_K}^T)^T$.

The parameters of interest are the fixed effects $\boldsymbol{\beta}_K$ and the covariance matrix Σ in the random effect distribution. Write $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \ldots, \mathbf{x}_{in_i}^T)^T$, $\mathbf{z}_i = (\mathbf{z}_{i1}^T, \ldots, \mathbf{z}_{in_i}^T)^T$, $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ and $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$. The approximate joint distribution of $(\boldsymbol{\beta}_K, \mathbf{b}_i, \Sigma)$ is

$$\hat{f}(\boldsymbol{\beta}_K, \mathbf{b}_i, \Sigma|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) \propto \hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K) f_b(\mathbf{b}_i|\Sigma)\pi(\boldsymbol{\beta}_K)\pi(\Sigma),$$

where $\hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \boldsymbol{\beta}_K, \mathbf{b}_i) = \prod_{j=1}^{n_i} \hat{f}_Y(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K)$ is the approximate conditional density of $\mathbf{Y}_i$ given the GDP parameters and the fixed and random effects, $f_b(\mathbf{b}_i|\Sigma)$ is the density of the Gaussian random effects, and $\pi(\boldsymbol{\beta}_K)$ and $\pi(\Sigma)$ are the priors for $\boldsymbol{\beta}_K$ and Σ, respectively. Throughout we assume $(\mathbf{b}_i, \Sigma)$ and $\boldsymbol{\beta}_K$ are independent of each other.

We put a truncated multivariate normal prior $N(\mathbf{0}, D)$ on $\boldsymbol{\beta}_K$ subject to $\mathbf{x}^T\boldsymbol{\beta}_{\tau_1} < \mathbf{x}^T\boldsymbol{\beta}_{\tau_2} < \cdots < \mathbf{x}^T\boldsymbol{\beta}_{\tau_K}$ and the inverse-Wishart $IW(\nu, S)$ on Σ, where $IW(\nu, S)$ is the inverse-Wishart distribution with the degree of freedom ν and the scale matrix S.

The corresponding joint posterior distribution can be expressed as

$$\hat{f}(\boldsymbol{\beta}_K, \mathbf{b}_i, \Sigma|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, D, \nu, S) \propto \hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K) f_b(\mathbf{b}_i|\Sigma)\pi(\boldsymbol{\beta}_K|D)\pi(\Sigma|\nu, S).$$

From this posterior distribution, we can obtain the full conditional distributions of each parameter by treating other parameters as fixed, and then update each parameter. The full conditional distributions for $\mathbf{b}_i$, Σ and $\boldsymbol{\beta}_K$ are

$$\begin{aligned}
\hat{f}(\mathbf{b}_i|\boldsymbol{\beta}_K, \Sigma, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, D, \nu, S) &\propto \hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K) f_b(\mathbf{b}_i|\Sigma), \\
f(\Sigma|\boldsymbol{\beta}_K, \mathbf{b}_i, \Sigma, \mathbf{y}, \mathbf{x}, \mathbf{z}, \boldsymbol{\theta}, D, \nu, S) &\propto \prod_{i=1}^{n} f_b(\mathbf{b}_i|\Sigma)\pi(\Sigma|\nu, S), \\
\hat{f}(\boldsymbol{\beta}_K|\mathbf{b}_i, \Sigma, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, D, \nu, S) &\propto \prod_{i=1}^{n} \hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i, \boldsymbol{\beta}_K)\pi(\boldsymbol{\beta}_K|D).
\end{aligned}$$

Note that Σ has a well-known full conditional distribution and thus we can easily update it with Gibb's sampler by using normal-inverse Wishart conjugacy. However, the full conditional distributions for $\mathbf{b}_i$ and $\boldsymbol{\beta}_K$ are based on the approximate densities and thus we update them by using the Metropolis-Hastings algorithm.

As mentioned in Section 2.3, we could pose a prior on the GPD parameter $\boldsymbol{\theta}$ and obtain its posterior distribution. However, our numerical studies show this does not lead to any further improvement but brings more computational complexity than fixing $\boldsymbol{\theta}$ at its initial estimate $\hat{\boldsymbol{\theta}}$ throughout. Therefore, we opt to estimate $\boldsymbol{\theta}$ following the similar method as introduced in Section 2.3, and fix $\hat{\boldsymbol{\theta}}$ throughout the MCMC.

3.2.2 Algorithm of the LIGPD Method for Clustered Data

In this subsection, we describe the algorithm for drawing samples from the approximate posterior distributions based on the proposed LIGPD likelihood.

Step 1. Obtain the initial values for $\mathbf{b}_i$, Σ and $\boldsymbol{\beta}_K$, denoted as $\mathbf{b}_i^{(0)}$, $\Sigma^{(0)}$, $\boldsymbol{\beta}_K^{(0)}$, respectively.

(a) Obtain $\mathbf{b}_i^{(0)}, i = 1, \ldots, n$, using the conventional quantile regression applied to data from each cluster separately. First, regress y_{ij} on $\mathbf{x}_{ij}, i = 1, \ldots, n, j = 1, \cdots, n_i$ at $\tau = 0.5$ and obtain $\boldsymbol{\beta}^0$. Then, we obtain $\mathbf{b}_i^{(0)}$ by regressing $y_{ij}^0 = y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}^0$ on $\mathbf{z}_{ij}$ using median regression with the data $\{y_{ij}^0, \mathbf{z}_{ij}, j = 1, \cdots, n_i\}$ for each cluster i separately.

(b) Define $\Sigma^{(0)} = (n-1)^{-1}\sum_{l=1}^{n} \mathbf{b}_i^{(0)}\mathbf{b}_i^{(0)^T}$.

(c) At each quantile level $\tau_k, k = 1, \ldots, K$, apply the conventional quantile regression by regressing $y^*_{ij} = y_{ij} - \mathbf{z}^T_{ij}\mathbf{b}^{(0)}_i$ on $\mathbf{x}_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i$ to obtain

$$\boldsymbol{\beta}^{(0)}_{\tau_k} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \rho_{\tau_k}(y^*_{ij} - \mathbf{x}^T_{ij}\boldsymbol{\beta}).$$

Define $\boldsymbol{\beta}^{(0)}_K = (\boldsymbol{\beta}^{(0)T}_{\tau_1}, \ldots, \boldsymbol{\beta}^{(0)T}_{\tau_K})^T$.

Step 2. Estimate the GPD parameters $\boldsymbol{\theta} = (\varrho_l, \varrho_r, \xi_l, \xi_r)$ based on $\{(y^*_{ij}, \mathbf{x}_{ij}), i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and $\boldsymbol{\beta}^{(0)}_K$.

(a) Estimate the scale parameters ϱ_l and ϱ_r by

$$\hat{\varrho}_l = \frac{(\tau_1 + \tau_2)}{2N(\tau_2 - \tau_1)} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathbf{x}^T_{ij}(\boldsymbol{\beta}^{(0)}_{\tau_2} - \boldsymbol{\beta}^{(0)}_{\tau_1}) \text{ and}$$
$$\hat{\varrho}_r = \frac{2 - (\tau_{K-1} + \tau_K)}{2N(\tau_K - \tau_{K-1})} \sum_{i=1}^{n}\sum_{j=1}^{n_i} \mathbf{x}^T_{ij}(\boldsymbol{\beta}^{(0)}_{\tau_K} - \boldsymbol{\beta}^{(0)}_{\tau_{K-1}}),$$

where $N = \sum_{i=1}^{n} n_i$.

(b) Obtain the shape parameter estimators, $\hat{\xi}_1$ and $\hat{\xi}_2$, by maximizing the GPD likelihood based on the exceedances $\{y^*_{ij} : y^*_{ij} < l_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and $\{y^*_{ij} : y^*_{ij} > u_{ij}, i = 1, \ldots, n, j = 1, \ldots, n_i\}$, respectively, where the lower thresholds $l_{ij} = \mathbf{x}^T_{ij}(\boldsymbol{\beta}^{(0)}_{\tau_1} + \boldsymbol{\beta}^{(0)}_{\tau_2})/2$ and the upper thresholds $u_{ij} = \mathbf{x}^T_{ij}(\boldsymbol{\beta}^{(0)}_{\tau_K} + \boldsymbol{\beta}^{(0)}_{\tau_{K-1}})/2$.

Step 3. For the $(t+1)$-th iteration, update $\mathbf{b}_i, i = 1, \cdots, n$, by using the following Metropolis-Hastings algorithm. Draw $\mathbf{b}^*_i$ from the candidate distribution, $N(\mathbf{b}^{(t)}_i, \Sigma^{(t)})$, and take

$$\mathbf{b}^{(t+1)}_i = \begin{cases} \mathbf{b}^*_i & \text{with probability } \phi(\mathbf{b}^{(t)}_i, \mathbf{b}^*_i), \\ \mathbf{b}^{(t)}_i & \text{with probability } 1 - \phi(\mathbf{b}^{(t)}_i, \mathbf{b}^*_i), \end{cases}$$

where

$$\phi(\mathbf{b}^{(t)}_i, \mathbf{b}^*_i) = \min\left\{ \frac{\hat{f}(\mathbf{b}^*_i | \boldsymbol{\beta}^{(t)}_K, \Sigma^{(t)}, D, \nu, S, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}})}{\hat{f}(\mathbf{b}^{(t)}_i | \boldsymbol{\beta}^{(t)}_K, \Sigma^{(t)}, D, \nu, S, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}})}, 1 \right\},$$

and the full conditional distribution of $\boldsymbol{b_i}$ is

$$\hat{f}(\mathbf{b}_i|\boldsymbol{\beta}_K, \Sigma, D, \nu, S, \mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}) \propto \hat{f}_Y(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \hat{\boldsymbol{\theta}}, \mathbf{b}_i, \boldsymbol{\beta}_K) f_b(\mathbf{b}_i|\Sigma).$$

Note that the transitional probabilities $P(\mathbf{b}_i^{(t)} \to \mathbf{b}_i^*)$ and $P(\mathbf{b}_i^* \to \mathbf{b}_i^{(t)})$ are the same since we use $N(\mathbf{b}_i^{(t)}, \Sigma^{(t)})$ as the proposal distribution. Thus, the transitional probabilities are ignored in $\phi(\mathbf{b}_i^{(t)}, \mathbf{b}_i^*)$.

Step 4. Update Σ using normal-inverse Wishart conjugacy. That is, we obtain $\Sigma^{(t+1)}$ from the full conditional distribution of Σ

$$\begin{aligned}
&f(\Sigma|\boldsymbol{\beta}_K, \mathbf{b}_i, D, \nu, S, \mathbf{y}, \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}}) \\
&\propto \prod_{i=1}^{n} f_b(\mathbf{b}_i|\Sigma) \times \pi(\Sigma|\nu, S) \\
&\propto |\Sigma|^{-n/2} \exp\left\{-\frac{tr(\Sigma^{-1}\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T)}{2}\right\} \frac{|S|^{-\nu/2}}{2^{\nu p/2}\Gamma_p(\nu/2)} |\Sigma|^{-(\nu+p+1)/2} \exp\left\{-\frac{tr(S\Sigma^{-1})}{2}\right\} \\
&\propto |\Sigma|^{-(\nu+p+1+n)/2} \exp\left[-\frac{1}{2} tr\left\{\Sigma^{-1}(\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T + S)\right\}\right] \\
&\propto IW(\nu+n, S+\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T),
\end{aligned}$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T$, $\mathbf{x} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$, $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_n^T)^T$ and $\Gamma_p(\cdot)$ is the multivariate gamma function.

Step 5. Update $\boldsymbol{\beta}_K$ based on the data $\{(y_{ij}^*, \mathbf{x}_{ij}), i = 1, \ldots, n, j = 1, \ldots, n_i\}$ with $y_{ij}^* = y_{ij} - \mathbf{z}_{ij}^T\mathbf{b}_i^{(t)}$. Randomly pick a quantile level, τ_k, and select a random number from 1 to p, say l. Then, we update β_{l,τ_k}, the lth element of $\boldsymbol{\beta}_{\tau_k}$. Let β_{l,τ_k}^* denote a candidate for β_{l,τ_k}. To avoid quantile crossing, the proposal distribution is chosen to have a finite support with lower and upper bounds. The specifications of bounds are provided in the Appendix. Let $\boldsymbol{\beta}_K^{(t+1)} = \boldsymbol{\beta}_K^{(t)}$ and $\boldsymbol{\beta}_K^* = \left(\boldsymbol{\beta}_{\tau_1}^{(t)T}, \ldots, \beta_{1,\tau_k}^{(t)}, \ldots, \beta_{l,\tau_k}^*, \beta_{l+1,\tau_k}^{(t)}, \ldots, \beta_{p,\tau_k}^{(t)}, \ldots, \boldsymbol{\beta}_{\tau_K}^{(t)T}\right)^T$.

We then update $\boldsymbol{\beta}_K^{(t+1)}$ by $\boldsymbol{\beta}_K^*$ with probability $\phi(\boldsymbol{\beta}_K^{(t)}, \boldsymbol{\beta}_K^*)$, where

$$\phi(\boldsymbol{\beta}_K^{(t)}, \boldsymbol{\beta}_K^*) = \min\left\{ \frac{\hat{f}(\boldsymbol{\beta}_K^*|\mathbf{b}^{(t)}, \Sigma^{(t)}, D, \nu, S, \mathbf{y}, \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}})}{\hat{f}(\boldsymbol{\beta}_K^{(t)}|\mathbf{b}^{(t)}, \Sigma^{(t)}, D, \nu, S, \mathbf{y}, \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}})}, 1 \right\},$$

$$\begin{aligned}\hat{f}(\boldsymbol{\beta}_K|\mathbf{b}, \Sigma, D, \nu, S, \mathbf{y}, \mathbf{x}, \mathbf{z}, \hat{\boldsymbol{\theta}}) &\propto \hat{f}_Y(\mathbf{y}|\mathbf{x}, \mathbf{z}, \boldsymbol{\beta}_K, \mathbf{b}, \hat{\boldsymbol{\theta}})\pi(\boldsymbol{\beta}_K|D) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \hat{f}_Y(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}_K, \mathbf{b}_i, \hat{\boldsymbol{\theta}})\pi(\boldsymbol{\beta}_K|D)\end{aligned}$$

with $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_n^T)^T$.

Step 6. Repeat Steps 3–5 until the MCMC chain converges. After a burn-in period, we obtain MCMC samples of $\boldsymbol{\beta}_K^{(t)}$, $\mathbf{b}^{(t)}$, and $\Sigma^{(t)}$, $t = 1, \ldots, N_t$. We estimate each parameter by using the mean of the corresponding MCMC samples.

3.3 Simulation Studies

In this subsection, we compare the performance of various methods for quantile regression with clustered data, including the proposed LIGPD method, the method in Geraci and Bottai (2013) based on the working asymmetric Laplace distribution (referred to as ALD), and the Bayesian empirical likelihood (BEL) approach proposed by Kim and Yang (2011). For regression at median, we also include the restricted maximum likelihood method (REML) for comparison.

3.3.1 Simulation Study I

We consider the following two models for generating the simulation data

$$\text{Model 1: } y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i}x_{ij} + \epsilon_{ij}, \tag{3.4}$$

$$\text{Model 2: } y_{ij} = \beta_0 + \beta_1 x_{ij} + b_{0i} + b_{1i}x_{ij} + x_{ij}\epsilon_{ij}, \tag{3.5}$$

where $x_{ij} \sim U(0, \sqrt{12})$, $\beta_0 = \beta_1 = 0$, $b_{ri} \overset{i.i.d.}{\sim} N(0, \sigma_r^2)$ with $\sigma_r^2 = 1$ for $r = 0, 1$, and ϵ_{ij} are independent random errors, $i = 1, \ldots, n = 20, j = 1, \ldots, n_i = 20$.

Model 1 has homoscedastic random errors, while model 2 has heteroscedastic random errors. For both models 1 and 2, we consider three distributions for generating ϵ_{ij}: $N(0,1)$, t_2 and $LN(0,1)$, where $LN(\mu,\varsigma)$ is the log normal distribution with location parameter μ and shape parameter ς. The corresponding conditional quantile function is

$$Q_\tau(Y_{ij}|x_{ij},b_{0i},b_{1i}) = \beta_{0,\tau} + \beta_{1,\tau}x_{ij} + b_{0i} + b_{1i}x_{ij}, \tag{3.6}$$

where $\beta_{0,\tau} = F_\epsilon^{-1}(\tau)$ and $\beta_{1,\tau} = 0$ for the homoscedastic model 1, $\beta_{0,\tau} = 0$ and $\beta_{1,\tau} = F_\epsilon^{-1}(\tau)$ for the heterosecdastic model 2, where F_ϵ is the CDF of the random error ϵ_{ij}.

For the LIGPD method, we consider a grid of $K = 9$ quantile levels $\tau_k = k/(1+K), k = 1,\ldots,K$, which are equally spaced between 0.1 and 0.9. We put a multivariate normal prior $N(0,D)$ on $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_{\tau_1}^T,\ldots,\boldsymbol{\beta}_{\tau_K}^T)^T$ with $D = \text{diag}(100,\ldots,100)$, and a flat prior on Σ. We define the LIGPD estimator as the posterior mean of every 100th draws of the MCMC chain (total 108,000 MCMC samples with the first 27,000 discarded as burn-in), and we use the "lqmm" function in the R package *lqmm* for the ALD method and the "lmer" function in the R package *lmer* for the REML method.

Note that we conduct the same simulation study as in Section 4.2 in Kim and Yang (2011) by using the LIGPD, ALD and REML methods. The results for the BEL method are taken from Table 3 in Kim and Yang (2011) since the program for the BEL method is not available. Kim and Yang (2011) assumed that the random intercept and slope are independent in their Monte Carlo simulations. For fair comparisons with the BEL method, the simulations for the REML, ALD and LIGPD methods were performed under the same independence assumption as in Kim and Yang (2011).

Comparison at Median

Table 3.3 and Figure 3.2 show the mean squared errors of the LIGPD, ALD and REML estimators at the median for the standard normal, t_2, and log normal random error distributions in both models 1 and 2. In Table 3.3 and Figure 3.2, the results for the BEL estimators are not included because Table 3 in Kim and Yang (2011) presents only the relative efficiency of the BEL estimator with respect to the REML estimator. To compare with the BEL estimator, Table 3.4 summarizes the relative efficiency of the LID, ALD and BEL estimators with respect to the REML estimator at median.

The REML method is for the mean regression model and can be used to estimate the

effects at median, which are the same as the mean effects when $\epsilon_{ij} \sim N(0,1)$ or t_2. The median and mean coincide at the center of the distribution for symmetric distributions but not for asymmetric distributions. The REML method estimates the conditional mean of the response while the other three methods estimate the conditional quantiles of the response. Since $LN(0,1)$ is an asymmetric distribution, the REML method and other three methods at median estimate different quantities. For example, the true intercept for the REML method is $\exp(1/2)$ and the true intercept at median for the LIGPD and ALD methods is $\exp(0) = 1$ for $\epsilon_{ij} \sim LN(0,1)$. Hence, we cannot directly compare the MSE of the REML estimator with those of the other three methods for the log normal error distribution. However, in Table 3.4, we can use the relative efficiencies for the log normal error to compare the LIGPD, ALD and BEL estimators since the MSE of the REML estimator will be canceled out when the relative efficiency for one estimator is divided by the relative efficiency for the other estimator.

The REML method assumed model (3.4) with the normal random error assumption. As expected, the REML method performs the best for estimating $\beta_{0,.5}$ and $\beta_{1,.5}$ for the homoscedastic and heteroscedastic normal errors and the results for the other three estimators are similar to each other.

For the homoscedastic t_2 errors, the LIGPD and ALD estimators show higher efficiency for estimating the intercepts than the REML method, while the LIGPD and ALD estimators are less efficient for estimating the slopes than the REML estimator. The BEL estimator is more efficient for estimating the fixed effects $\beta_{0,.5}$ and $\beta_{1,.5}$ than the other three estimators. The REML estimator is not as efficient as the other estimators for estimating σ_0^2 and σ_1^2. For the heteroscedastic t_2 errors, the REML estimator is less efficient than the other three estimators. Kim and Yang (2011) stated that the REML method reported unusually high estimates for σ_1^2 and σ_2^2 but we did not have such large variance estimates for the REML method. Thus, for the heteroscedastic t_2 distribution errors, the MSE ratio for the BEL method might be inflated due to a few very large variance estimates from the REML method obtained by Kim and Yang (2011).

For the homoscedastic and heteroscedastic log normal error distributions, the LIGPD method shows comparable or higher efficiency than the other methods.

In general, the LIGPD estimator produces the highest efficiency for estimating σ_1^2 and σ_2^2. The BEL estimator shows higher efficiency for estimating $\beta_{0,.5}$ and $\beta_{1,.5}$ for t_2 errors. The ALD and LIGPD estimators have similar efficiency for estimating $\beta_{0,.5}$ and $\beta_{1,.5}$ but

Table 3.1: Mean squared error of parameter estimators at median for the data generated from model (3.4) and (3.5) from the LIGPD, REML and ALD method. The REML estimators are grayed out when it cannot be directly compared to LIGPD and ALD.

Model	Distribution	Method	$\beta_{0,.5}$	$\beta_{1,.5}$	σ_0^2	σ_1^2
1	$N(0,1)$	LIGPD	0.099	0.074	0.222	0.120
		ALD	0.083	0.080	0.122	0.140
		REML	0.062	0.045	0.194	0.138
	t_2	LIGPD	0.095	0.067	0.260	0.100
		ALD	0.088	0.080	0.223	0.178
		REML	0.147	0.063	0.887	0.178
	$LN(0,1)$	LIGPD	0.092	0.063	0.159	0.087
		ALD	0.164	0.085	0.195	0.169
		REML	0.104	0.059	0.435	0.161
2	$N(0,1)$	LIGPD	0.080	0.085	0.142	0.136
		ALD	0.075	0.084	0.227	0.225
		REML	0.071	0.058	0.361	0.178
	t_2	LIGPD	0.078	0.110	0.185	0.162
		ALD	0.072	0.081	0.448	0.306
		REML	0.213	0.151	1.096	1.561
	$LN(0,1)$	LIGPD	0.108	0.078	0.146	0.128
		ALD	0.078	0.086	0.311	0.398
		REML	0.134	0.112	0.862	0.405

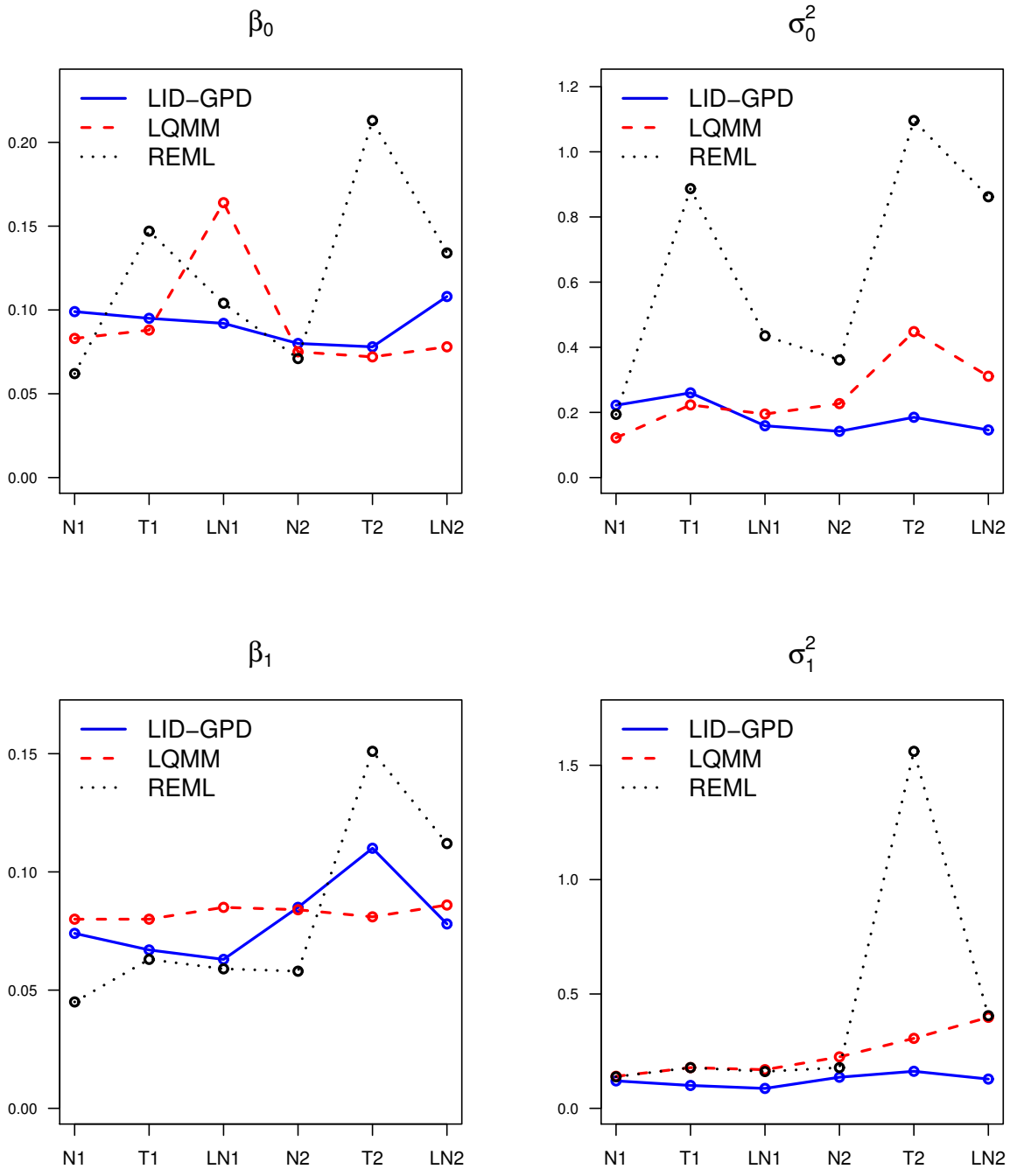


Figure 3.1: Mean squared error of parameter estimators at median for three models with homoscedastic errors (N1: $N(0,1)$, T1: t_2, LN1: $LN(0,1)$) and three models with heteroscedastic errors (N2: $N(0,1)$, T2: t_2, LN2: $LN(0,1)$) from the LIGPD (solid), REML (dotted) and ALD (dashed) methods.

Table 3.2: Relative efficiency of the LIGPD, ALD and BEL estimators at median with respect to the REML estimator at median. Relative efficiency larger than 1 means the competing method has smaller MSE and thus higher efficiency than the REML estimator. The results for BEL were from Table 3 in Kim and Yang (2011).

Model	Distribution	Method	$\beta_{0,.5}$	$\beta_{1,.5}$	σ_0^2	σ_1^2
1	$N(0,1)$	LIGPD	0.629	0.612	0.875	1.154
		ALD	0.751	0.563	1.591	0.986
		BEL	0.928	0.987	0.629	0.748
	t_2	LIGPD	1.545	0.945	3.412	1.789
		ALD	1.666	0.787	3.978	1.001
		BEL	3.195	1.639	2.718	1.871
	$LN(0,1)$	LIGPD	1.138	0.936	2.735	1.847
		ALD	0.637	0.694	2.227	0.951
		BEL	0.769	0.972	1.830	0.952
2	$N(0,1)$	LIGPD	0.887	0.688	2.547	1.308
		ALD	0.941	0.697	1.590	0.794
		BEL	0.909	0.885	1.432	0.734
	t_2	LIGPD	2.736	1.379	5.922	9.660
		ALD	2.961	1.860	2.444	5.105
		BEL	3.532	3.070	3.254	130.226
	$LN(0,1)$	LIGPD	1.246	1.433	5.913	3.175
		ALD	1.714	1.296	2.767	1.018
		BEL	1.509	0.711	2.919	1.421

Table 3.3: Mean squared error of parameter estimators at median for the data generated from model (3.4) and (3.5) from the LIGPD, REML and ALD method. The REML estimators are grayed out when it cannot be directly compared to LIGPD and ALD.

Model	Distribution	Method	$\beta_{0,.5}$	$\beta_{1,.5}$	σ_0^2	σ_1^2
1	$N(0,1)$	LIGPD	0.099	0.074	0.222	0.120
		ALD	0.083	0.080	0.122	0.140
		REML	0.062	0.045	0.194	0.138
	t_2	LIGPD	0.095	0.067	0.260	0.100
		ALD	0.088	0.080	0.223	0.178
		REML	0.147	0.063	0.887	0.178
	$LN(0,1)$	LIGPD	0.092	0.063	0.159	0.087
		ALD	0.164	0.085	0.195	0.169
		REML	0.104	0.059	0.435	0.161
2	$N(0,1)$	LIGPD	0.080	0.085	0.142	0.136
		ALD	0.075	0.084	0.227	0.225
		REML	0.071	0.058	0.361	0.178
	t_2	LIGPD	0.078	0.110	0.185	0.162
		ALD	0.072	0.081	0.448	0.306
		REML	0.213	0.151	1.096	1.561
	$LN(0,1)$	LIGPD	0.108	0.078	0.146	0.128
		ALD	0.078	0.086	0.311	0.398
		REML	0.134	0.112	0.862	0.405

the LIGPD estimators have higher efficiency for estimating σ_0^2 and σ_1^2.

Comparison between LIGPD and ALD across quantiles

Tables 3.5–3.10 summarize the mean squared errors of the parameters from the proposed LIGPD method and the ALD method proposed by Geraci and Bottai (2013) at nine quantile levels ($\tau = 0.1, 0.2, ..., 0.9$) for the standard normal, t_2, and the log normal random error distributions and their heteroscedastic versions.

For the LIGPD method, the parameters σ_0^2 and σ_1^2 are constant across quantile levels, so we only report the estimates at $\tau = 0.1$. However, even though σ_0^2 and σ_1^2 are quantile-invariant, ALD fits the data at each quantile level separately and thus the resulting estimations of σ_0^2 and σ_1^2 vary across quantiles.

For the homoscedastic and heteroscedastic normal errors, both LIGPD and ALD

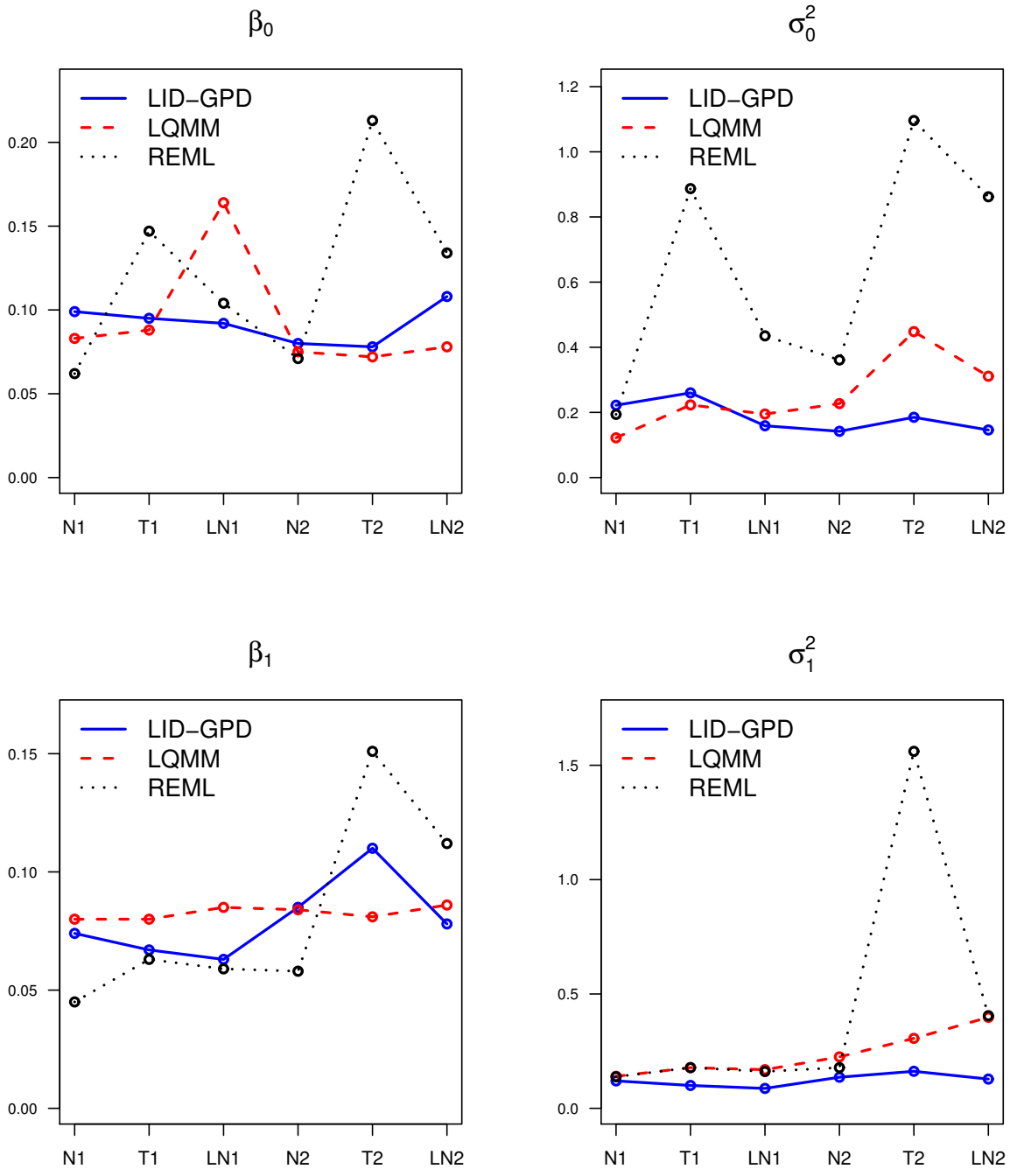


Figure 3.2: Mean squared error of parameter estimators at median for three models with homoscedastic errors (N1: $N(0,1)$, T1: t_2, LN1: $LN(0,1)$) and three models with heteroscedastic errors (N2: $N(0,1)$, T2: t_2, LN2: $LN(0,1)$) from the LIGPD (solid), REML (dotted) and ALD (dashed) methods.

Table 3.4: Relative efficiency of the LIGPD, ALD and BEL estimators at median with respect to the REML estimator at median. Relative efficiency larger than 1 means the competing method has smaller MSE and thus higher efficiency than the REML estimator. The results for BEL were from Table 3 in Kim and Yang (2011).

Model	Distribution	Method	$\beta_{0,.5}$	$\beta_{1,.5}$	σ_0^2	σ_1^2
1	$N(0,1)$	LIGPD	0.629	0.612	0.875	1.154
		ALD	0.751	0.563	1.591	0.986
		BEL	0.928	0.987	0.629	0.748
	t_2	LIGPD	1.545	0.945	3.412	1.789
		ALD	1.666	0.787	3.978	1.001
		BEL	3.195	1.639	2.718	1.871
	$LN(0,1)$	LIGPD	1.138	0.936	2.735	1.847
		ALD	0.637	0.694	2.227	0.951
		BEL	0.769	0.972	1.830	0.952
2	$N(0,1)$	LIGPD	0.887	0.688	2.547	1.308
		ALD	0.941	0.697	1.590	0.794
		BEL	0.909	0.885	1.432	0.734
	t_2	LIGPD	2.736	1.379	5.922	9.660
		ALD	2.961	1.860	2.444	5.105
		BEL	3.532	3.070	3.254	130.226
	$LN(0,1)$	LIGPD	1.246	1.433	5.913	3.175
		ALD	1.714	1.296	2.767	1.018
		BEL	1.509	0.711	2.919	1.421

Table 3.5: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim N(0,1)$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.114	0.126	0.075	0.132	0.222	1.364	0.120	0.455
0.2	0.112	0.137	0.075	0.104		0.437		0.321
0.3	0.101	0.129	0.073	0.089		0.162		0.187
0.4	0.099	0.100	0.073	0.079		0.138		0.136
0.5	0.099	0.083	0.074	0.080		0.122		0.140
0.6	0.097	0.097	0.076	0.084		0.123		0.135
0.7	0.101	0.141	0.076	0.090		0.180		0.180
0.8	0.101	0.154	0.079	0.100		0.302		0.321
0.9	0.104	0.134	0.081	0.132		0.882		0.323

estimators have comparable results for the middle quantiles and the efficiency gain of the LIGPD method is obvious for low and high quantiles.

For the homoscedastic t_2 errors, the LIGPD method shows higher efficiency for estimating the slope but the LIGPD and ALD methods show comparable results for estimating the intercept since MSEs of the ALD estimator lie in 95% confidence bands of MSE of the LIGPD estimator in most quantile region. For heteroscedastic t_2 errors, the LIGPD estimator shows higher efficiency for estimating the intercept and lower efficiency for estimating the slope. For both homoscedastic and heteroscedastic t_2 errors, the LIGPD shows clear efficiency gain for estimating σ_0^2 and σ_1^2, which is more obvious for low and high quantiles.

For the homoscedastic log normal errors, the LIGPD estimator is more efficient than the ALD estimator in most cases. For the heteroscedastic log normal errors, the LIGPD estimator shows higher efficiency for estimating β_0 and β_1 for low quantiles and the ALD estimator is more efficient for high quantiles. For both homoscedastic and heteroscedastic log normal errors, the LIGPD shows higher efficiency for estimating σ_0^2 and σ_1^2.

In general, the proposed LIGPD method shows higher efficiency than the ALD method for estimating σ_1^2 and σ_2^2. For the normal and log normal random errors, the LIGPD estimator shows higher efficiency for estimating the fixed effects β_0 and β_1 in most cases. For the t_2 errors, the LIGPD and ALD methods have comparable results for estimating the fixed effects $\beta_{0,\tau}$ and $\beta_{1,\tau}$.

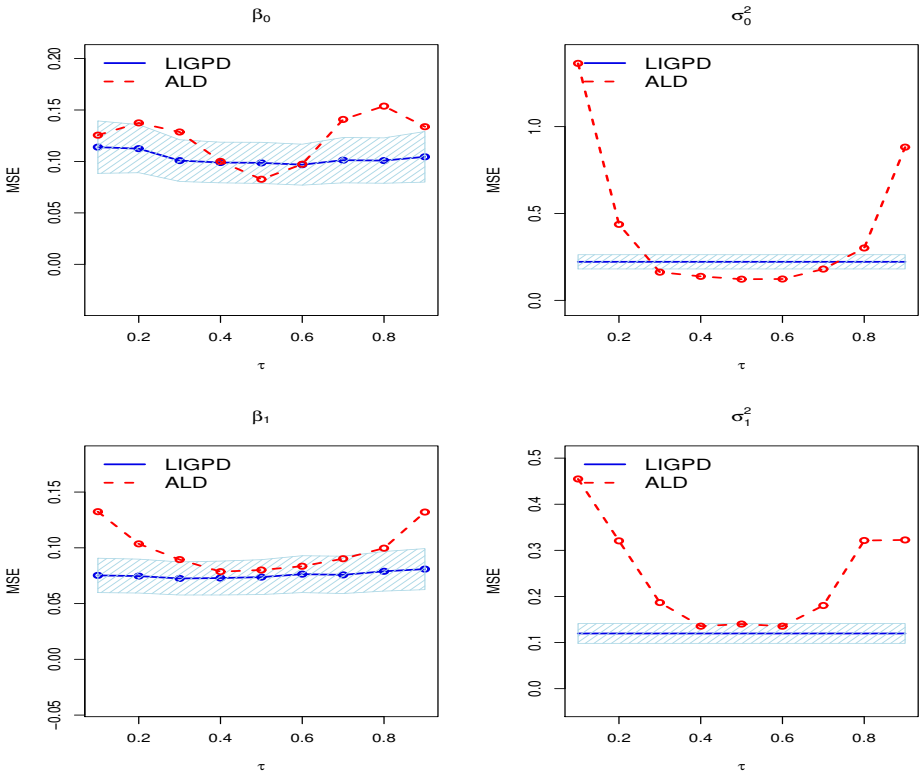


Figure 3.3: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim N(0, 1)$. MSE of $\beta_{0,.5}$ (top left), $\beta_{1,.5}$ (bottom left), σ_0^2 (top right) and σ_0^2 (bottom right) of LIGPD (solid line) and ALD (dashed line). The horizontal axis represents quantile levels and the vertical axis represents MSE. The shaded bands represent 95% confidence bands of MSE of the LIGPD estimator. For the LIGPD method, the estimates of random-effect variance parameters are constant across quantile levels.

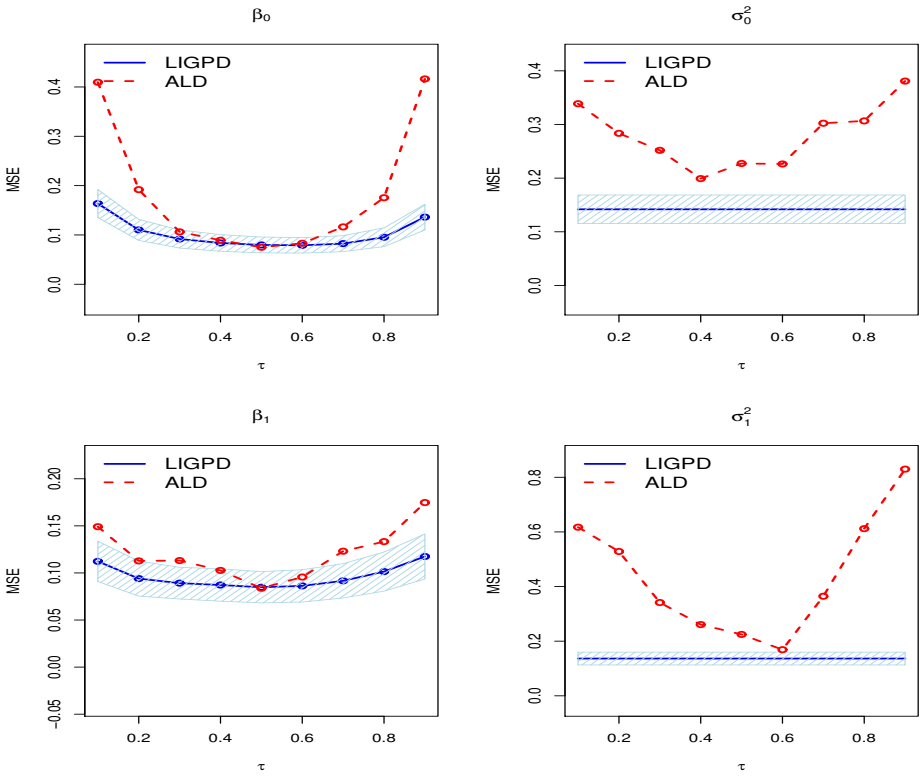


Figure 3.4: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim N(0, 1)$.

Table 3.6: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim N(0, 1)$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.164	0.410	0.112	0.149	0.142	0.339	0.136	0.617
0.2	0.110	0.192	0.094	0.113		0.283		0.528
0.3	0.092	0.106	0.089	0.113		0.252		0.342
0.4	0.084	0.089	0.087	0.103		0.199		0.261
0.5	0.080	0.075	0.085	0.084		0.227		0.225
0.6	0.079	0.083	0.086	0.096		0.226		0.169
0.7	0.082	0.116	0.092	0.123		0.302		0.365
0.8	0.095	0.175	0.101	0.133		0.307		0.612
0.9	0.136	0.416	0.117	0.175		0.381		0.830

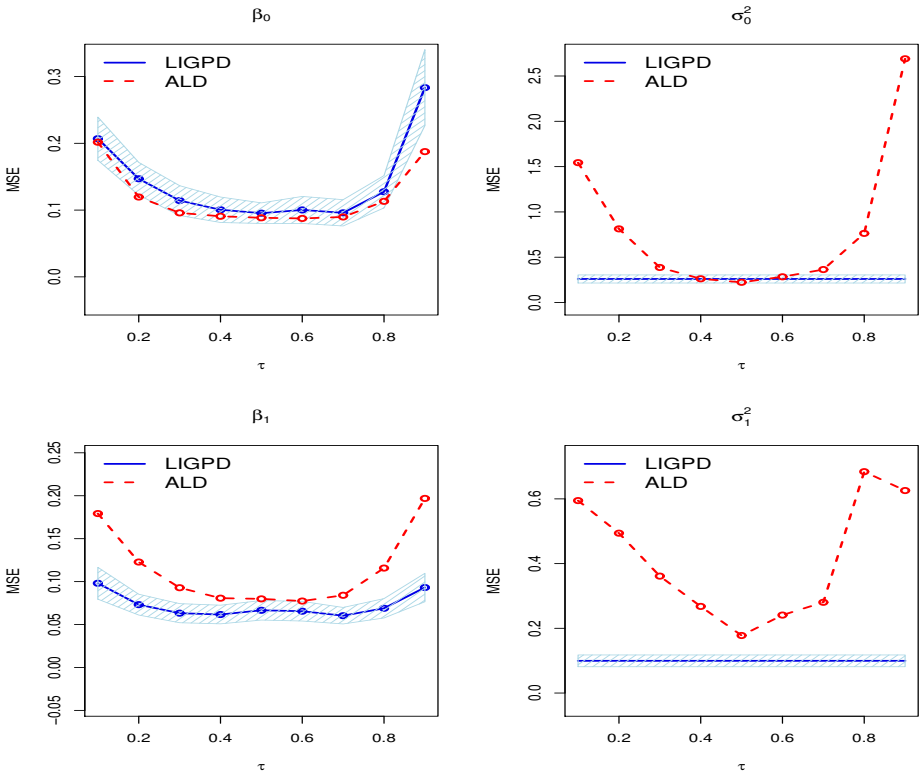


Figure 3.5: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim t_2$.

Table 3.7: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim t_2$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.207	0.202	0.098	0.179	0.260	1.544	0.100	0.595
0.2	0.147	0.120	0.073	0.123		0.813		0.494
0.3	0.114	0.096	0.063	0.093		0.387		0.361
0.4	0.100	0.091	0.062	0.081		0.262		0.268
0.5	0.095	0.088	0.067	0.080		0.223		0.178
0.6	0.100	0.088	0.066	0.077		0.284		0.241
0.7	0.096	0.090	0.060	0.084		0.364		0.280
0.8	0.127	0.113	0.069	0.116		0.763		0.684
0.9	0.283	0.188	0.093	0.197		2.690		0.626

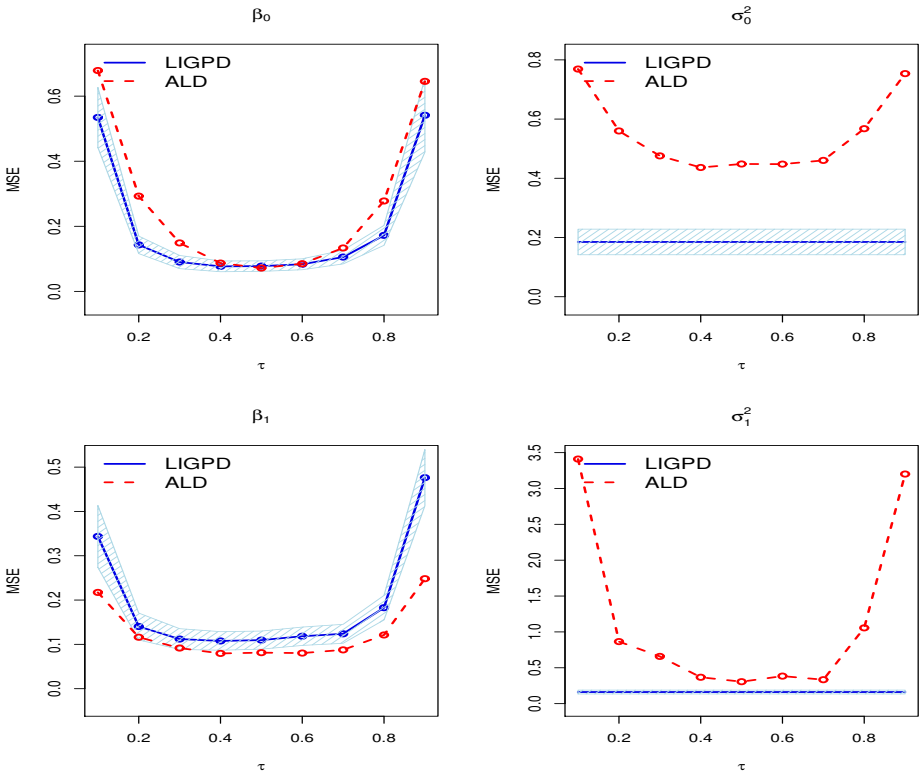


Figure 3.6: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim t_2$.

Table 3.8: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim t_2$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.535	0.679	0.344	0.217	0.185	0.769	0.162	3.410
0.2	0.143	0.293	0.140	0.116		0.560		0.865
0.3	0.090	0.149	0.112	0.092		0.476		0.658
0.4	0.077	0.087	0.108	0.080		0.436		0.368
0.5	0.078	0.072	0.110	0.081		0.448		0.306
0.6	0.083	0.085	0.118	0.080		0.448		0.384
0.7	0.106	0.133	0.124	0.088		0.460		0.334
0.8	0.172	0.278	0.183	0.121		0.568		1.056
0.9	0.541	0.645	0.476	0.248		0.754		3.201

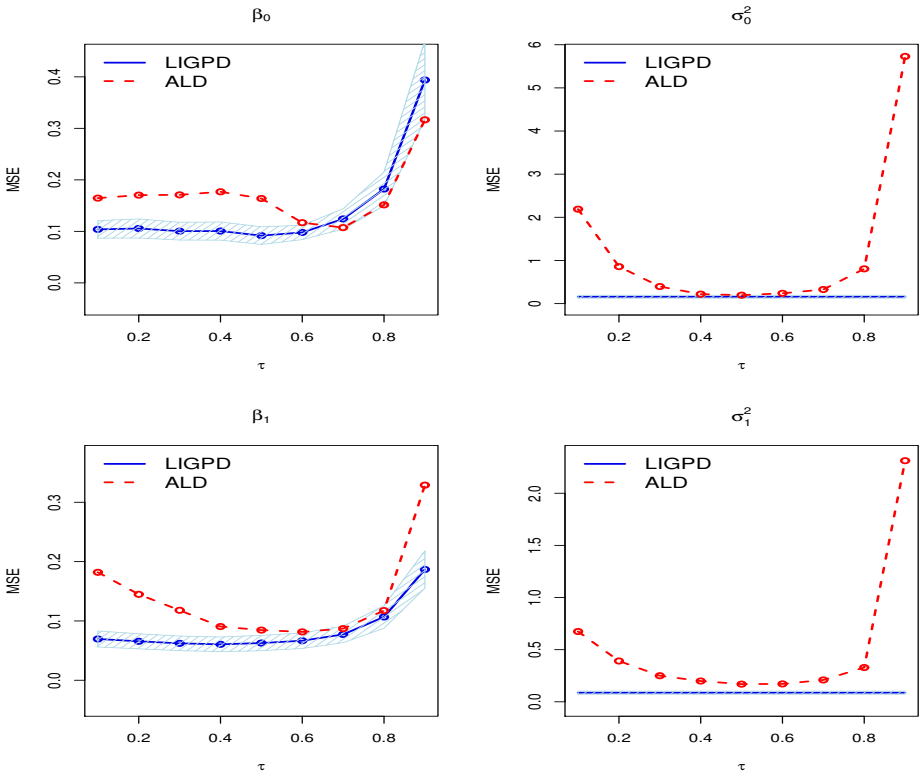


Figure 3.7: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim LN(0, 1)$.

Table 3.9: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) with $\epsilon_{ij} \sim LN(0, 1)$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.104	0.165	0.070	0.182	0.159	2.189	0.087	0.674
0.2	0.106	0.170	0.066	0.145		0.857		0.391
0.3	0.100	0.171	0.062	0.118		0.396		0.250
0.4	0.101	0.177	0.061	0.091		0.220		0.199
0.5	0.092	0.164	0.063	0.085		0.195		0.169
0.6	0.098	0.117	0.067	0.082		0.238		0.172
0.7	0.124	0.107	0.077	0.087		0.326		0.209
0.8	0.182	0.151	0.107	0.117		0.806		0.329
0.9	0.394	0.317	0.187	0.329		5.729		2.312

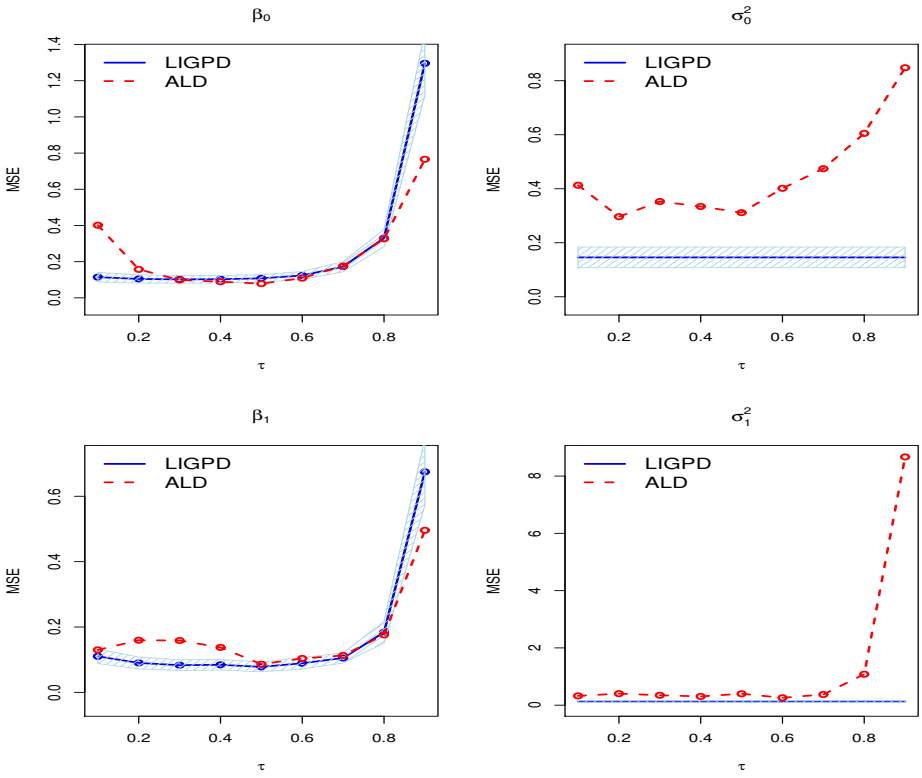


Figure 3.8: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim LN(0, 1)$.

Table 3.10: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.5) with $\epsilon_{ij} \sim LN(0, 1)$.

	β_{0,τ_k}		β_{1,τ_k}		σ_0^2		σ_1^2	
τ	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD	LIGPD	ALD
0.1	0.114	0.401	0.110	0.130	0.146	0.413	0.128	0.325
0.2	0.104	0.157	0.090	0.160		0.297		0.403
0.3	0.102	0.099	0.083	0.159		0.353		0.347
0.4	0.103	0.089	0.084	0.138		0.334		0.306
0.5	0.108	0.078	0.078	0.086		0.311		0.398
0.6	0.124	0.109	0.089	0.104		0.402		0.259
0.7	0.173	0.176	0.105	0.113		0.474		0.372
0.8	0.330	0.327	0.184	0.176		0.605		1.081
0.9	1.296	0.766	0.675	0.496		0.848		8.669

Table 3.11: Quantile crossing rates in percentage (%) of the ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.4) and (3.5) .

Model	Error	Quantile Crossing Rate
1	$N(0,1)$	85.7
	t_2	46.6
	$LN(0,1)$	78.2
2	$N(0,1)$	46.8
	t_2	10.1
	$LN(0,1)$	58.8

Quantile Crossing Issue

Table 3.11 summarizes the quantile crossing rates of the ALD method in different scenarios, where the quantile crossing rate is defined to be the proportion of simulated data in which the quantile monotonicity $\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\tau_1} \leq \cdots \leq \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\tau_K}$ is violated at least once. As expected, we do not observe any quantile crossing for the LIGPD method but the ALD has high quantile crossing rates. For instance, for homoscedastic $N(0,1)$ errors, 85.7% of the time the conditional quantile estimates from ALD may cross each other at least once in some region of the covariates and at two or more quantile levels.

3.3.2 Simulation Study II

We consider a more complex simulation study with two predictors. The data were generated from the following linear mixed model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{0i} + b_{1i} x_{ij1} + b_{2i} x_{ij2} + \epsilon_{ij}, i = 1, \cdots, n, j = 1, ..., n_i, \quad (3.7)$$

where $n = 20$, $n_i = 20$, $x_{ij1} \sim LN(0,1)$, $x_{ij2} \sim Bernoulli(0.5)$, $\beta_0 = \beta_1 = \beta_2 = 1$, $b_{ri} \sim N(0, \sigma_r^2)$ with $\sigma_r^2 = 1$ for $r = 0, 1, 2$, and ϵ_{ij} are independent random errors. The random effects b_{0i}, b_{1i} and b_{2i} are independent of each other and independent of ϵ_{ij}. We consider three distributions for generating the random errors ϵ_{ij}: $N(0,1)$, t_3 and χ_1^2.

The corresponding conditional quantile regression model is

$$Q_\tau(Y_{ij}|x_{ij1}, x_{ij2}, \mathbf{b}_i) = \beta_{0,\tau} + \beta_{1,\tau} x_{ij1} + \beta_{2,\tau} x_{ij2} + b_{0i} + b_{1i} x_{ij1} + b_{2i} x_{ij2}, \quad (3.8)$$

where $\beta_{0,\tau} = 1 + F_\epsilon^{-1}(\tau)$, $\beta_{1,\tau} = \beta_{2,\tau} = 1$ and F_ϵ is the CDF of the random error ϵ_{ij}.

For each scenario, the simulation is repeated 200 times. In this simulation, we focus on the comparison of the proposed LIGPD method and the ALD method proposed by Geraci and Bottai (2013). For both methods, we did not assume that b_{0i}, b_{1i} and b_{2i} are independent. Thus, we estimate three fixed effects $\beta_r, r = 0, 1, 2$ and six random effect variances $\sigma_0^2, \sigma_1^2, \sigma_2^2$, σ_{12}, σ_{13} and σ_{23}, where σ_{r_1,r_2} is the covariance between $b_{r_1 i}$ and $b_{r_2 i}$.

For the LIGPD method, we consider a grid of $K = 9$ quantile levels $\tau_k = k/(1 + K), k = 1, \ldots, K$, which are equally spaced between 0.1 and 0.9. We put a multivariate normal prior $N(0, D)$ on $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_{\tau_1}^T, \ldots, \boldsymbol{\beta}_{\tau_K}^T)^T$ with $D = \text{diag}(100, \ldots, 100)$, and a flat prior on Σ. We define the LIGPD estimator as the posterior mean of every 100th draws of the MCMC chain (total 81,000 MCMC samples with the first 40,500 discarded as burn-in).

Comparison between LIGPD and ALD across quantiles

Tables 3.12–3.14 and Figures 3.9–3.11 summarize the mean squared errors of the fixed and random-effect variance parameters from the LIGPD and ALD methods at nine quantile levels ($\tau_k = 0.1, 0.2, ..., 0.9$) for the standard normal, t_3, and χ_1^2 distributions, respectively.

In general, the proposed LIGPD method shows higher efficiency than the ALD method for estimating the random-effect variances in all scenarios considered. For fixed effects, the LIGPD estimators have either similar or higher efficiency than the ALD estimators especially at the tail quantiles.

Table 3.15 summarizes the quantile crossing rates for the ALD method. As expected, we do not observe any quantile crossing for the LIGPD estimator but the ALD has high quantile crossing rates.

Comparison of Computation Time between LIGPD and ALD

Both the proposed method and the ALD method require multi-dimensional integration using the Markov chain Monte Carlo algorithm and the Gaussian quadrature approximation, respectively. In Gaussian quadrature approximation, the number of functions to be evaluated grows exponentially in the dimension. In contrast, in the Markov chain Monte Carlo algorithm, the number of functions to be evaluated increases linearly in the dimension. Therefore, the computing of the ALD method is more challenging for higher dimensional data. For both ALD and LIGPD methods, all simulations were computed

using the same machine with Intel i7 CPU with 2.67 GHz clock speed and 6 GB RAM. In simulation study I, the average computing time for analyzing one data set was 1.12 seconds for the ALD method at one quantile level, while 38.62 seconds for the proposed LIGPD method for the analysis at nine quantile levels jointly. However, in simulation study II, the computation for the LIGPD method is approximately nine times faster than the ALD method; the average computing time for one data set was 42.15 seconds for the ALD method at one quantile level, while 46.72 seconds for the LID-GPD methods for the analysis at nine quantile levels.

In summary, the proposed LIGPD estimator has comparable or higher efficiency than the competing estimators in most cases, especially for estimating the random-effect variance parameters and fixed effects at tail quantiles. For models with multiple predictors, the LIGPD method is computationally faster than the ALD method. Lastly, the ALD and BEL methods are subject to quantile crossing, which is avoided by the proposed method through joint quantile modeling.

3.4 Apnea Duration Data Analysis

In this section, we analyze an apena duration data of elderly women for the swallow study conducted by Perlman et al. (2005). Apnea duration is the period of the absence of respiration during swallowing. This data has also been analyzed in Zhou and He (2008), Wang (2009), Reich et al. (2010) and Kim and Yang (2011). We use 856 Apnea durations measured on 31 healthy females, aged from 70 to 85.

We consider the following random intercept model with viscosity, feeding type and their interaction effects:

$$Q_\tau(Y_{ijkl}|V_j, F_k, b_i) = \beta_{0,\tau} + \beta_{1,\tau}V_j + \beta_{2,\tau}F_k + \beta_{3,\tau}VF_{jk} + b_i,$$

where Y_{ijkl} is the l-th repeated apnea duration (in seconds) of subject i for viscosity V_j (pudding=1 and water=0) and feeding type F_k (self-fed=1 and examiner-fed=0), VF_{jk} are the interaction effects and b_i are subject-specific random effects.

Figure 3.12 shows the box plots of the apnea duration for each of the four viscosity-feeding-type combinations. The box plots suggest that the covariate effects appear different at the central and upper quantiles. For example, the difference in the median apnea durations between the first two groups (water-examiner-fed and pudding-examiner-

Table 3.12: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim N(0,1)$. For the LIGPD method, the estimates of random-effect variance parameters are constant across quantile levels.

τ_k	β_{0,τ_k}	β_{1,τ_k}	β_{2,τ_k}	σ_0^2	σ_1^2	σ_2^2	σ_{12}	σ_{13}	σ_{23}
LIGPD									
0.1	0.134	0.085	0.135	0.151	0.120	0.237	0.084	0.088	0.070
0.2	0.118	0.075	0.108						
0.3	0.111	0.072	0.110						
0.4	0.105	0.069	0.114						
0.5	0.094	0.069	0.112						
0.6	0.104	0.069	0.104						
0.7	0.104	0.071	0.094						
0.8	0.106	0.074	0.097						
0.9	0.124	0.085	0.123						
ALD									
0.1	0.361	0.226	0.309	5.884	4.673	4.365	4.978	5.273	4.148
0.2	0.310	0.203	0.208	6.373	8.868	3.341	7.122	4.259	4.942
0.3	0.217	0.174	0.158	4.862	11.046	2.816	7.454	3.715	5.087
0.4	0.161	0.148	0.138	4.912	11.847	2.798	8.135	3.488	5.167
0.5	0.137	0.136	0.129	4.635	11.898	2.893	7.434	3.570	5.182
0.6	0.109	0.157	0.118	3.994	12.036	2.704	7.192	3.124	5.304
0.7	0.131	0.127	0.131	3.901	9.807	2.730	6.110	3.002	4.520
0.8	0.193	0.134	0.113	4.233	7.095	2.541	5.155	2.995	3.803
0.9	0.161	0.163	0.175	2.267	1.999	1.219	1.670	1.385	1.186

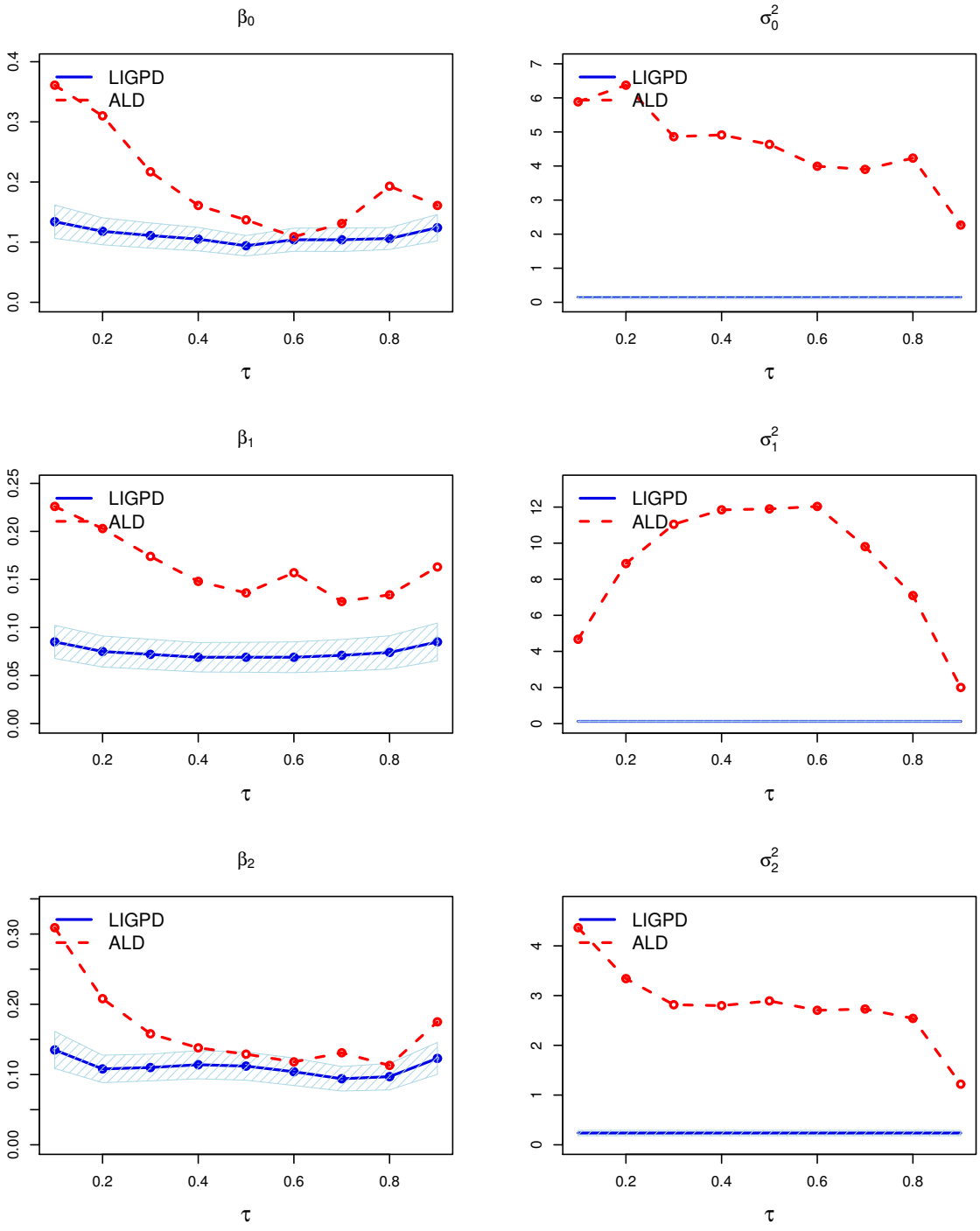


Figure 3.9: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim N(0, 1)$. The solid line is for LIGPD and the dashed line is for ALD. The shaded region indicates 95% pointwise confidence bands of the MSE of the LIGPD estimators.

Table 3.13: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim t_3$.

τ_k	β_{0,τ_k}	β_{1,τ_k}	β_{2,τ_k}	σ_0^2	σ_1^2	σ_2^2	σ_{12}	σ_{13}	σ_{23}
LIGPD									
0.1	0.209	0.134	0.340	0.250	0.132	0.321	0.106	0.160	0.099
0.2	0.135	0.093	0.145						
0.3	0.119	0.081	0.123						
0.4	0.114	0.075	0.121						
0.5	0.113	0.073	0.117						
0.6	0.113	0.072	0.117						
0.7	0.101	0.073	0.129						
0.8	0.117	0.080	0.144						
0.9	0.165	0.119	0.260						
ALD									
0.1	0.462	0.227	0.413	7.467	3.736	7.580	3.845	4.307	3.190
0.2	0.306	0.231	0.248	6.435	5.675	4.202	5.416	3.988	3.742
0.3	0.214	0.203	0.204	4.514	10.051	3.214	5.976	3.160	4.899
0.4	0.151	0.149	0.177	3.453	9.984	3.546	5.375	2.790	5.168
0.5	0.137	0.135	0.140	3.777	11.445	3.070	6.012	3.220	5.318
0.6	0.123	0.143	0.145	3.859	9.854	3.266	5.510	2.823	4.779
0.7	0.120	0.119	0.129	3.904	8.181	2.524	5.091	2.834	3.961
0.8	0.156	0.153	0.177	3.291	5.094	2.876	3.180	2.115	2.548
0.9	0.221	0.200	0.281	2.542	1.272	2.159	0.987	1.288	0.794

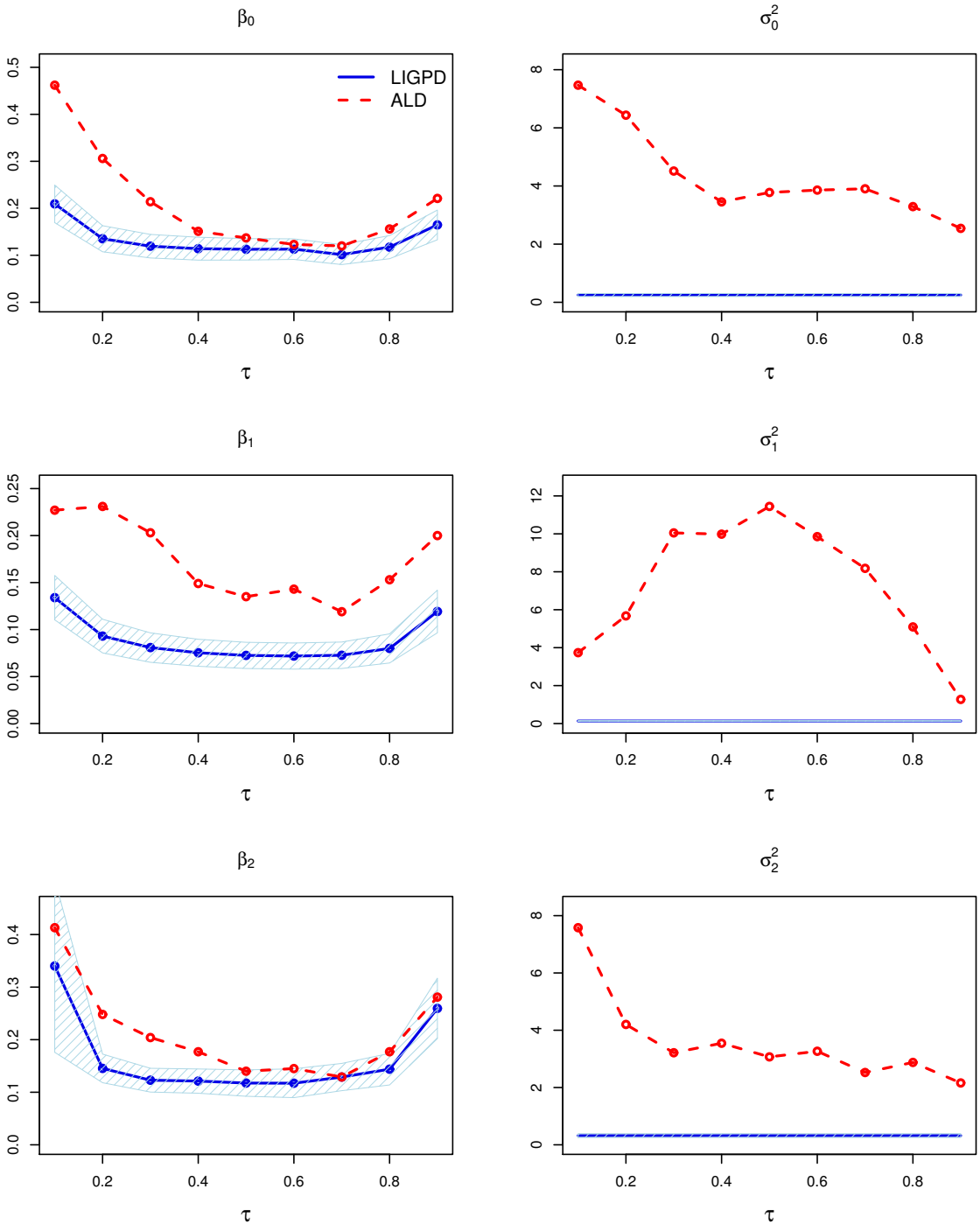


Figure 3.10: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim t_3$. The solid line is for LIGPD and the dashed line is for ALD. The shaded region indicates 95% pointwise confidence bands of the MSE of the LIGPD estimators.

Table 3.14: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim \chi^2_1$.

τ_k	β_{0,τ_k}	β_{1,τ_k}	β_{2,τ_k}	σ_0^2	σ_1^2	σ_2^2	σ_{12}	σ_{13}	σ_{23}
LIGPD									
0.1	0.083	0.057	0.061	0.102	0.111	0.143	0.063	0.062	0.051
0.2	0.087	0.056	0.061						
0.3	0.081	0.056	0.062						
0.4	0.079	0.056	0.069						
0.5	0.077	0.057	0.079						
0.6	0.086	0.060	0.097						
0.7	0.103	0.070	0.134						
0.8	0.145	0.094	0.204						
0.9	0.307	0.198	0.541						
ALD									
0.1	0.260	0.222	0.275	6.579	5.149	3.256	6.766	5.849	4.487
0.2	0.391	0.223	0.209	9.058	10.394	2.808	9.980	5.598	5.451
0.3	0.405	0.179	0.183	7.355	13.031	3.033	10.266	5.078	6.245
0.4	0.364	0.170	0.169	6.782	13.375	3.302	9.301	4.761	6.546
0.5	0.310	0.152	0.136	5.732	11.355	3.644	7.960	4.145	5.873
0.6	0.219	0.129	0.149	5.134	10.355	3.858	6.861	3.848	5.646
0.7	0.139	0.135	0.159	4.751	9.402	3.530	5.885	3.150	4.294
0.8	0.200	0.136	0.172	4.626	5.991	4.072	3.571	2.195	3.088
0.9	0.273	0.184	0.343	4.519	1.634	5.288	0.962	1.230	0.847

Table 3.15: Quantile crossing rates in percentage (%) of the ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7).

Error	Quantile Crossing Rate
$N(0,1)$	98.7
t_3	97.4
χ^2_1	97.9

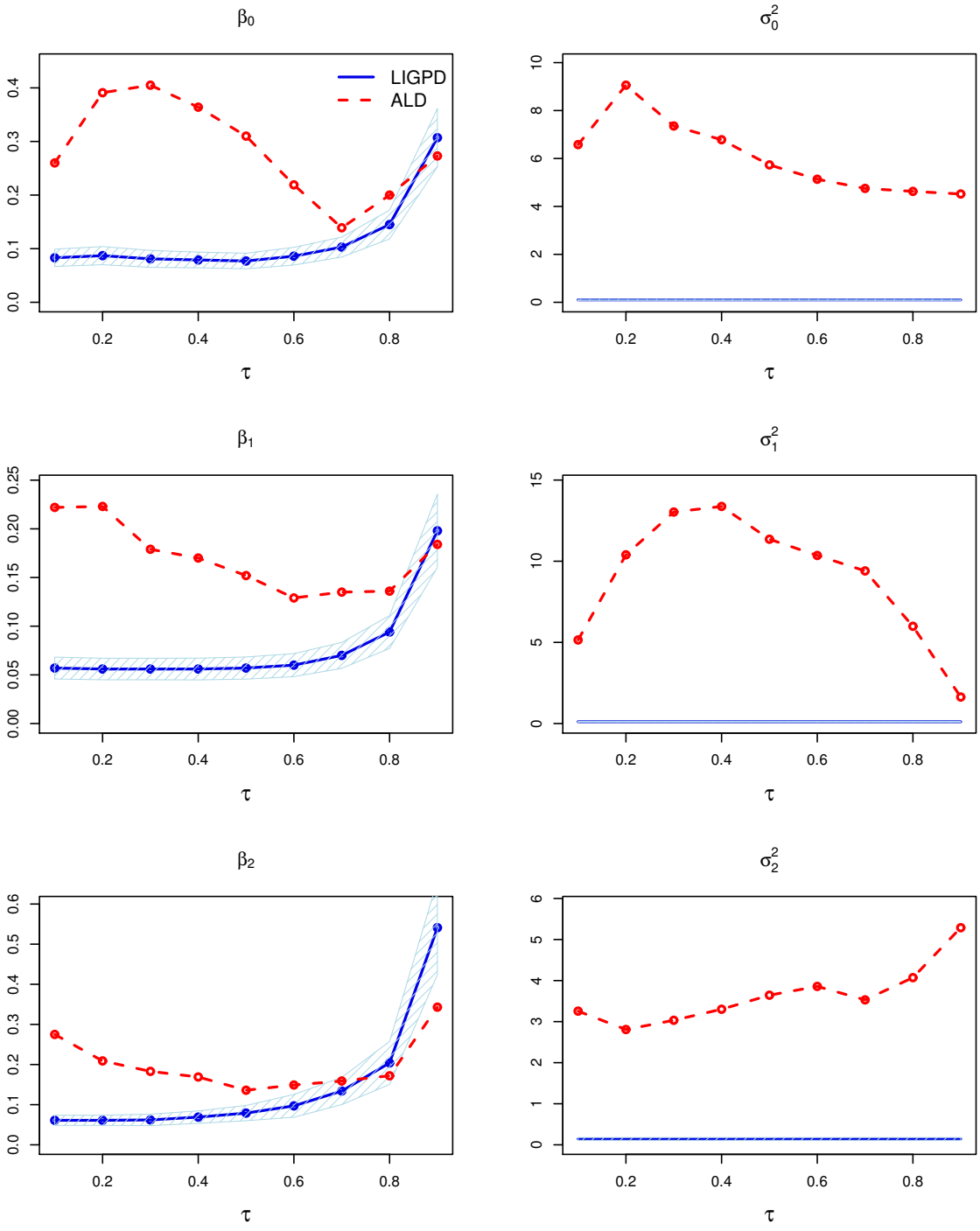


Figure 3.11: Mean squared errors of LIGPD and ALD estimators at $\tau = 0.1, \cdots, 0.9$ for data generated from model (3.7) with $\epsilon_{ij} \sim \chi^2_1$. The solid line is for LIGPD and the dashed line is for ALD. The shaded region indicates 95% pointwise confidence bands of the MSE of the LIGPD estimators.

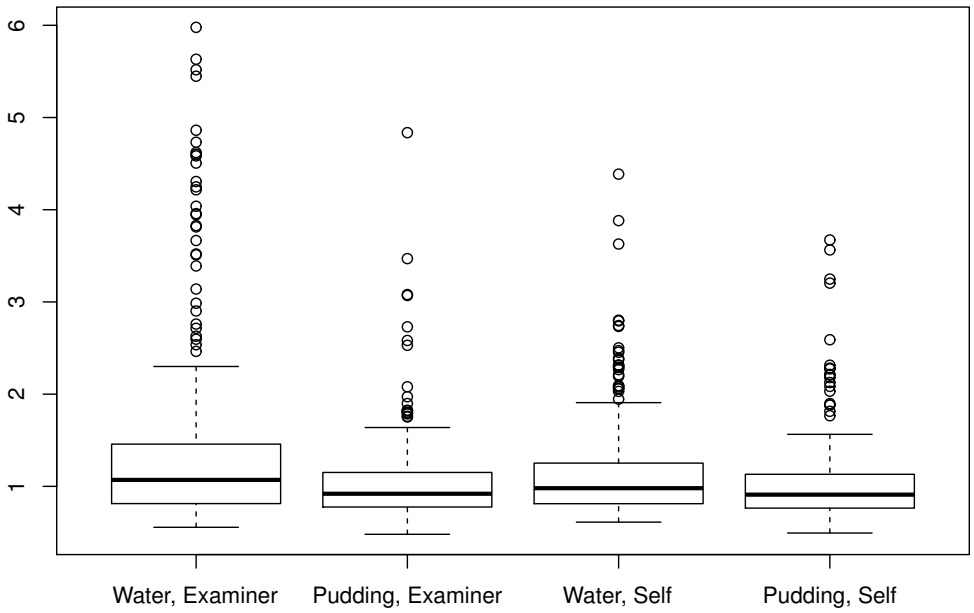


Figure 3.12: Boxplots of the apnea duration for each viscosity-feeding-type combination.

fed) appears to be smaller than the difference in the third quartiles. For this data set, the mean regression assuming homogeneous Gaussian random errors is inappropriate since apnea duration is right-skewed even after the log transform. Alternatively, we consider two quantile-regression-based methods: the proposed LIGPD method and the ALD method in Geraci and Bottai (2013) based on the working asymmetric Laplace distribution. For the LIGPD method, we use a grid of $K = 15$ equally spaced quantile levels $\tau_k = k/(1 + K), k = 1, \cdots, K$, and assign the normal priors $N(0, 100)$ to the random intercept effect b_i. With 75,000 MCMC samples, we discard the first 25,000 as burn-in, and estimate $\beta_{j,\tau}$ for $j = 0, 1, 2, 3$ with the posterior means of the thinned sample using every 100th draws of the MCMC chain.

Figure 3.13 presents the point estimations and 95% credible intervals for the fixed effects from the LIGPD and ALD methods across quantiles. The LIGPD method shows that the viscosity effects are significantly negative at all quantiles, suggesting that senior women tend to have longer apnea duration when drinking water than when swallowing pudding. The feeding condition and interaction effects get larger as the quantile level increases, and they tend to be significant at the upper quantiles. This suggests that the upper quantiles of the apnea duration are longer if subjects are fed by examiners, and that at the upper tails of the apnea duration the viscosity effect is stronger (in absolute value) for subjects being fed by examiners than by themselves. The estimations from the ALD method are similar to those from LIGPD at the lower and central quantiles. However, the main and interaction effects from ALD change slowly with the quantile level, leading to insignificance at the upper quantiles. In addition, examination of the estimated quantiles of apnea duration for the four viscosity-feeding-type combinations in Figure 3.14 shows that the ALD estimates suffer from quantile crossing but the estimates from LIGPD appear to be monotonically increasing with the quantile level, making the interpretation of estimation results easier.

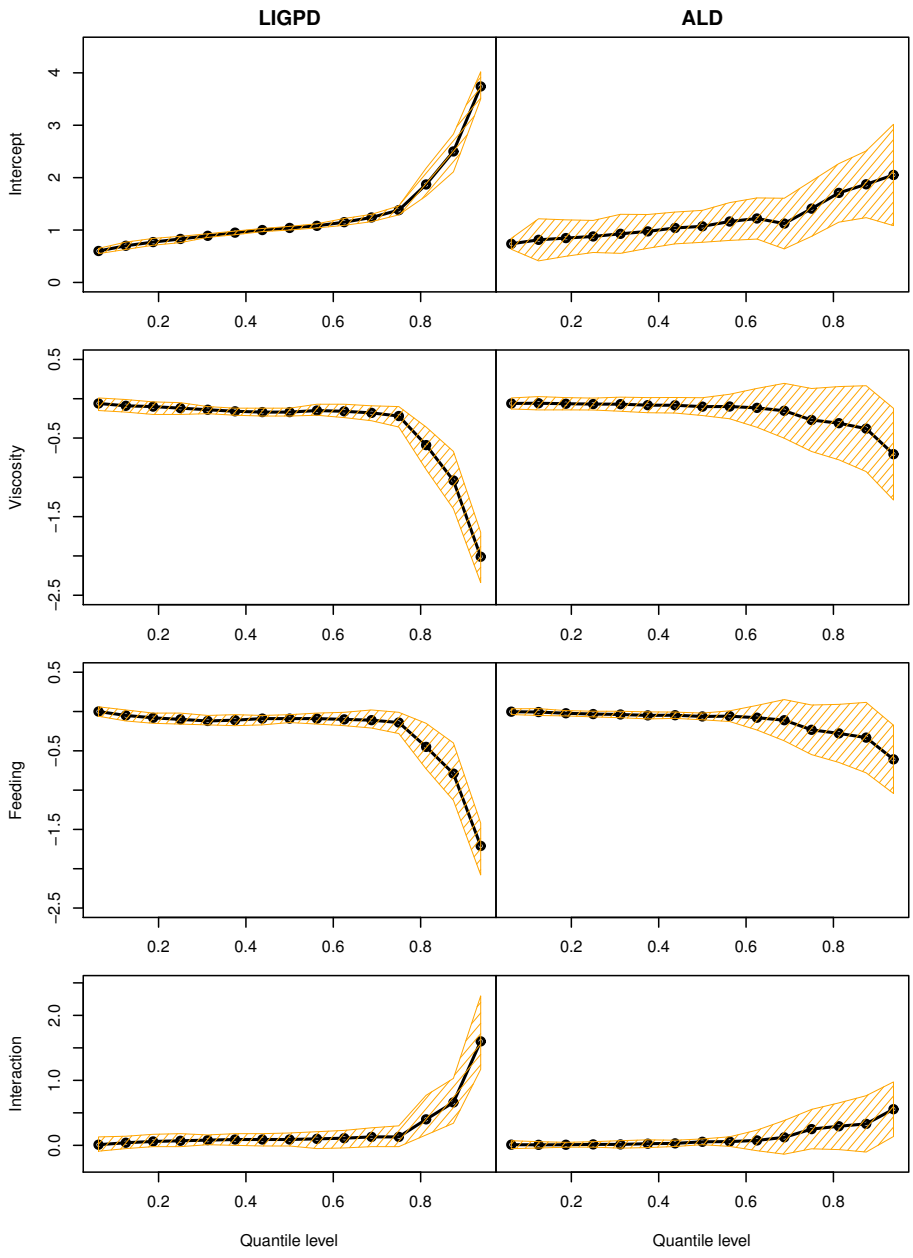


Figure 3.13: Point estimates (solid curves with dots) and 95% pointwise credible intervals (shaded regions) for the fixed effects at $\tau \in [0.0625, 0.9375]$ from LIGPD and ALD.

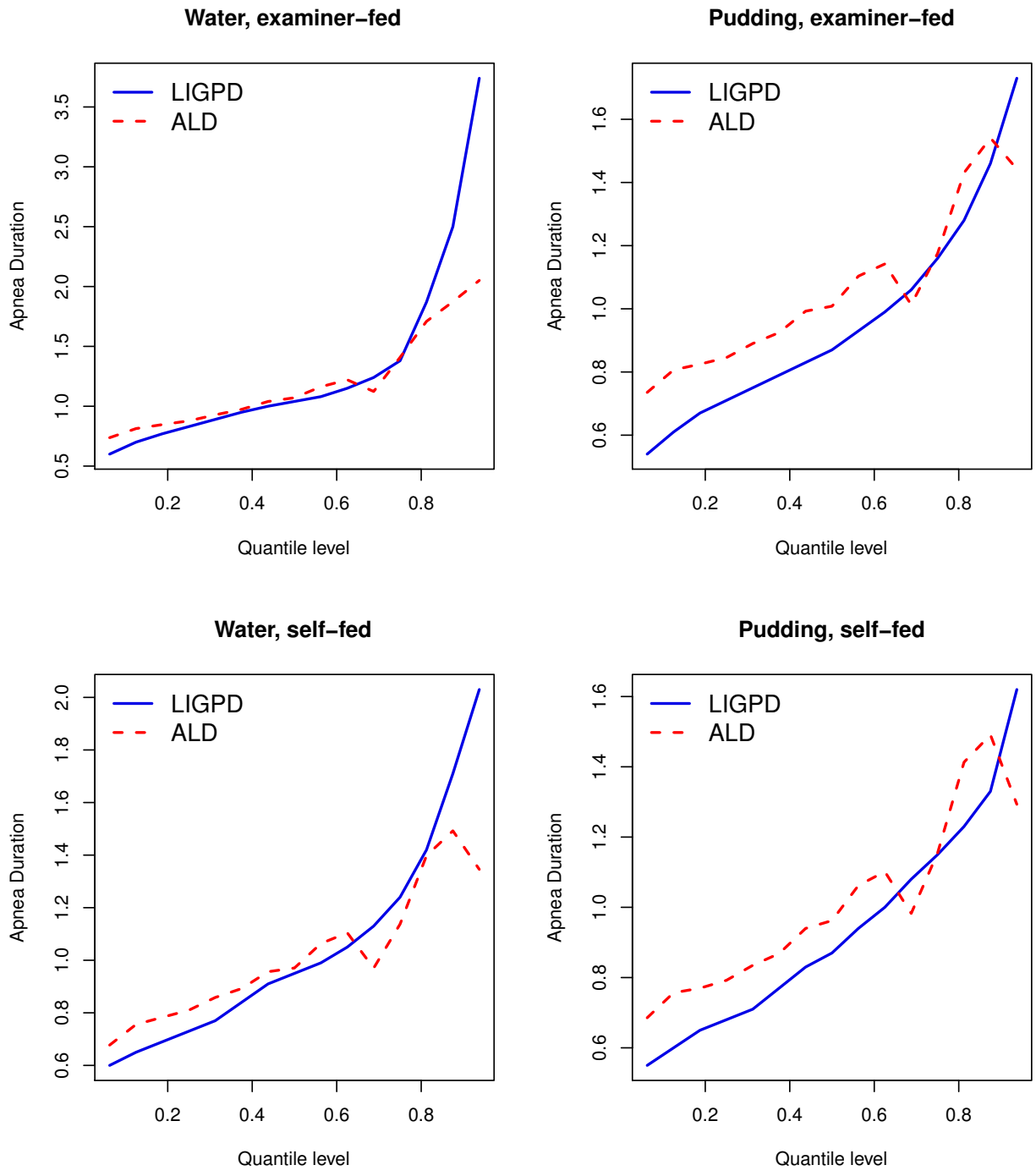


Figure 3.14: Estimated quantiles of apnea duration from LIGPD and ALD in four groups.

Chapter 4

LIGPD for Quantile and Interquantile Shrinkage in Quantile Regression Models

In this chapter, we propose and investigate the LIGPD methods with different types of penalizations. The proposed methods aim to shrink quantile slopes and their interquantile differences towards zero in some or the entire quantile region with fused LASSO or group LASSO penalties on quantile slopes and interquantile slope differences based on Laplace priors. Through numerical investigations, we show that the proposed methods yield higher estimation efficiency than the conventional quantile regression method when there exists sparsity in the quantile coefficients and/or interquantile differences.

4.1 Introduction

Penalized regression has been widely used to improve prediction accuracy and model interpretability. The exemplary penalization regression methods include the non-negative garrote in Breiman (1995), the LASSO in Tibshirani (1996), group LASSO in Bakin (1999) and Yuan and Lin (2006), the smoothly clipped absolute deviation (SCAD) in Fan and Li (2001) and the adaptive LASSO in Zou (2006).

In quantile regression analysis, regularization techniques have been considered in different contexts but most of them focused on quantile regression analysis at a single quantile level. For instance, the adaptive LASSO for median regression in Wang et al.

(2007), the LASSO and its solution path algorithm in Li and Zhu (2008), the SCAD in Wu and Liu (2009) and the adaptive LASSO penalization for censored quantile regression in Wang et al. (2013).

Focusing on a single quantile level may overlook the quantile coefficients that are significantly apart from zero at other quantile levels. Zou and Yuan (2008) proposed a simultaneous multiple quantile regression method (the composite quantile regression, CQR) that shares the same quantile slope coefficients across multiple quantile levels. By assuming constant quantile slope across multiple quantile levels, the CQR estimator has higher efficiency in cases where the assumptions are met. By assuming the same conditions for establishing asymptotic normality of the conventional quantile regression (QR) at a single quantile level in addition to the global linearity of quantile functions, the CQR enjoys asymptotic normality and may have higher efficiency than the QR. Jiang et al. (2012) extended the composite quantile regression to non-linear quantile regressions.

Both Zou and Yuan (2008) and Jiang et al. (2012) proposed the adaptive LASSO penalized CQR estimators for variable selection that remove covariates whose effects are zero at all quantile levels. The penalized CQR estimators are the best suited for the situations where non-sparse quantile slopes have zero interquantile slope differences at all quantile levels. The common quantile slope coefficients in the CQR are coincided with the homoscedastic random errors and zero interquantile slope differences across all quantile levels. The common quantile slope is a restrictive assumption that may be violated in practice. For example, the quantile slopes can be constant in some quantile region and vary across other quantile regions. Jiang et al. (2013, 2014) proposed methods that assign a weighted LASSO or sup-norm penalty on quantile slope coefficients and interquantile slope differences, and the methods can identify constant quantile slope coefficients in some or the entire quantile region.

In the Bayesian framework, Li et al. (2010) proposed hierarchical models for Bayesian quantile regression models with LASSO, group LASSO and elastic net penalties and developed Gibbs algorithms by using Bayesian hierarchical models and Gibbs sampling algorithms of Park and Casella (2008). Kyung et al. (2010) further extended the idea of Park and Casella (2008) for linear regressions with various types of penalizations including LASSO, group LASSO, fused LASSO and elastic net.

Motivated by the fused adaptive LASSO penalty of Jiang et al. (2014) and Bayesian hierarchical models in Kyung et al. (2010) and Li et al. (2010), we develop penaliza-

tion methods for quantile and interquantile shrinkage for linear quantile regression in the Bayesian framework using linear interpolated density. Li et al. (2010) focused on regression at a single quantile level and cannot shrink interquantile slope differences. Joint modeling of multiple quantiles in the LIGPD method does not require the strong constant slope assumption as in Zou and Yuan (2008) but only the linearity assumption on the quantile functions. By replacing the asymmetric Laplace error distribution in Li et al. (2010) with the linearly interpolated density, we can shrink quantile slopes and their interquantile differences simultaneously and improve the estimation efficiency through joint-modeling of multiple quantiles.

4.2 Proposed Method

4.2.1 Model Setup

We consider the following quantile regression model

$$Q_{\tau_k}(Y_i|\mathbf{x}_i) = \alpha_{\tau_k} + \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_k},\ i = 1, \ldots, n, k = 1, \ldots, K, 0 < \tau_k < 1, \tag{4.1}$$

where Y_i is the response, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is a p-dimensional covariate vector associated with the i-th observation, α_{τ_k} is a quantile intercept, $\boldsymbol{\beta}_{\tau_k} = (\beta_{1,\tau_k}, \ldots, \beta_{p,\tau_k})^T$ is a p-dimensional vector of quantile coefficients and β_{l,τ_k} is the quantile coefficient for the l-th predictor at the τ_k-th quantile level. Suppose we are interested in estimating $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_{\tau_1}^T, \ldots, \boldsymbol{\beta}_{\tau_k}^T)^T$, where $0 < \tau_1 < \cdots < \tau_K < 1$ is a sequence of quantile levels. To incorporate interquantile differences in the model, we define the slope differences as

$$d_{l,\tau_k} = \begin{cases} \beta_{l,\tau_k} - \beta_{l,\tau_{k-1}} & \text{if} \quad k = 2, \ldots, K, l = 1, \ldots, p, \\ \beta_{l,\tau_1} & \text{if} \quad k = 1,\ l = 1, \ldots, p. \end{cases}$$

Then we can rewrite the quantile function $Q_{\tau_k}(Y|\mathbf{x}_i)$ in model (4.1) as

$$Q_{\tau_k}(Y|\mathbf{x}_i) = \alpha_{\tau_k} + \sum_{j=1}^{k} \sum_{l=1}^{p} x_{il} d_{l,\tau_j}.$$

4.2.2 Bayesian Fused Adaptive Group LASSO

We consider a Bayesian fused adaptive group LASSO estimator using Laplace priors on quantile slopes and interquantile differences. To help the understanding of the propsoed method, we first illustrate the method in the frequentist framework.

Our first proposed method imposes a weighted fused group LASSO penalty on the quantile slope coefficients and interquantile slope differences. In the frequentist framework, the fused adaptive group LASSO estimator can be obtained by

$$\hat{\boldsymbol{\beta}}_K = \arg\min \sum_{k=1}^{K} \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - \alpha_{\tau_k} - \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_k}\} + \lambda_1 \sum_{l=1}^{p} \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2 + \lambda_2 \sum_{l=1}^{p} \tilde{v}_l ||\mathbf{d}_{l,\tau_K}||_2, \tag{4.2}$$

where λ_1 and λ_2 are regularization parameters, $\tilde{w}_l$ and $\tilde{v}_l$ are the adaptive weights for $\boldsymbol{\beta}_{l,\tau_K} = (\beta_{l,\tau_1}, \ldots, \beta_{l,\tau_K})^T$ and $\mathbf{d}_{l,\tau_K} = (\beta_{l,\tau_2} - \beta_{l,\tau_1}, \ldots, \beta_{l,\tau_K} - \beta_{l,\tau_{K-1}})^T$, respectively, and $||\cdot||_2$ is the l_2-norm. The adaptive weights are obtained by using the estimates $\tilde{\boldsymbol{\beta}}_k$ from the conventional quantile regression and we define $\tilde{w}_l = ||\tilde{\boldsymbol{\beta}}_{l,\tau_K}||_2^{-1}$ and $\tilde{v}_l = ||\tilde{\mathbf{d}}_{l,\tau_K}||_2^{-1}$. With the adaptive weights, zero quantile coefficients and interquantile differences will be shrunk towards zero at a higher speed than non-zero ones.

The regularization parameters λ_1 and λ_2 control the overall speed of shrinkage for quantile slopes and interquantile slope differences, respectively. In the frequentist framework, we can select λ_1 and λ_2 by information criteria such as Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwarz, 1978). In the Bayesian framework, we can choose λ_1 and λ_2 by putting gamma hyperpriors on them. Park and Casella (2008) suggested to use "relatively flat" hyperpriors that put posterior estimates near the MLE instead of improper priors because improper priors lead to improper posteriors.

The fused adaptive group LASSO method treats quantile slope coefficients or interquantile slope differences corresponding to a single predictor as a group and impose a group-wise penalty on them. Hence, in the frequentist framework, the fused adaptive group LASSO estimator will eliminate the predictors that have no impact across the entire quantile region and shrink quantile slopes at different quantile levels towards a constant value if the quantile slopes are constant across all quantile levels. Park and Casella (2008) discussed that Bayesian LASSO estimates with Laplace priors are very

close to frequentist LASSO estimates but cannot be exactly zero. Hence, we expect our Bayesian fused adaptive group LASSO estimator to shrink β_{l,τ_k} towards zero but not exactly to zero if $||\boldsymbol{\beta}_{l,\tau_K}||_2 = 0$ and shrink d_{l,τ_k} towards zero but not exactly to zero if $||\mathbf{d}_{l,\tau_K}||_2 = 0$.

Working Likelihood

Bayesian regularization methods require the likelihood as well as the prior specification in addition to the formulation of the regularization. We use the linearly interpolated density as the working likelihood, which can be expressed as

$$\begin{aligned}\hat{f}(y|\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K-1} I\left\{y \in \left[\alpha_{\tau_k} + \sum_{m=1}^{k}\sum_{l=1}^{p} x_{il} d_{l,\tau_m}, \alpha_{\tau_{k+1}} + \sum_{m=1}^{k+1}\sum_{l=1}^{p} x_{il} d_{l,\tau_m}\right)\right\} \times \\ \frac{\tau_{k+1} - \tau_k}{\alpha_{\tau_{k+1}} - \alpha_{\tau_k} + \mathbf{x}_i^T \mathbf{d}_{\tau_{k+1}}} + I\left\{y \in \left(-\infty, \alpha_{\tau_1} + \sum_{l=1}^{p} x_{il} d_{l,\tau_1}\right)\right\} \tau_1 f_1(y) \\ + I\left\{y \in \left[\alpha_{\tau_K} + \sum_{m=1}^{K}\sum_{l=1}^{p} x_{il} d_{l,\tau_m}, \infty\right)\right\} (1 - \tau_K) f_2(y),\end{aligned} \tag{4.3}$$

where $\boldsymbol{\theta} = (\alpha_{\tau_1}, ..., \alpha_{\tau_K}, \mathbf{d}_{\tau_1}, ..., \mathbf{d}_{\tau_K})$, $\mathbf{d}_{\tau_k} = (d_{1,\tau_k}, \ldots, d_{p,\tau_k})$ and $f_1(\cdot)$ and $f_2(\cdot)$ are left and right tail density functions approximated by generalized Pareto distributions; see details in Chapter 2. Note that we do not include the shape and scale parameters in $f_1(\cdot)$ and $f_2(\cdot)$ to simplify the expression.

Prior Specifications

We put normal prior distributions on $\alpha_{\tau_1}, ..., \alpha_{\tau_K}$, Laplace prior distributions on $\boldsymbol{\beta}_{1,\tau_K}, \ldots,$ $\boldsymbol{\beta}_{p,\tau_K}$ and $\mathbf{d}_{1,\tau_K}, \ldots, \mathbf{d}_{p,\tau_K}$ and gamma prior distributions on λ_1^2 and λ_2^2:

$$\begin{aligned}
\pi(\alpha_{\tau_k}|\varsigma^2) &= \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left(-\frac{\alpha_{\tau_k}^2}{2\varsigma^2}\right), \\
\pi(\boldsymbol{\beta}_K|\lambda_1, \lambda_2) &\propto \exp\left\{-\lambda_1 \sum_{l=1}^{p} \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2 - \lambda_2 \sum_{l=1}^{p} \tilde{v}_l ||\mathbf{d}_{l,\tau_K}||_2\right\}. \\
\pi(\lambda_1^2) &\sim \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b), \\
\pi(\lambda_2^2) &\sim \frac{(\lambda_2^2)^{c-1}}{\Gamma(c)\, d^c} \exp(-\lambda_2^2/d),
\end{aligned}$$

where ς^2, a, b, c and d are some constants.

Similar to Li et al. (2010) and Kyung et al. (2010), we use a scale mixture representation of the multivariate Laplace distribution based on equation (2.2) in Andrews and Mallows (1974) (see Appendix A.4) to utilize Gibb's samplers in the MCMC algorithm. The prior distributions of $\boldsymbol{\beta}_{l,\tau_K}$ and $\mathbf{d}_{l,\tau_K}$ for the group LASSO penalty can be expressed as

$$\pi(\boldsymbol{\beta}_{l,\tau_K}|\lambda_1) \propto \exp\left\{-\lambda_1 \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2\right\}$$

and

$$\pi(\mathbf{d}_{l,\tau_K}|\lambda_2) \propto \exp\left\{-\lambda_2 \tilde{v}_l ||\mathbf{d}_{l,\tau_K}||_2\right\}.$$

The scale mixture representation of the multivariate Laplace distribution $\pi(\boldsymbol{\beta}_{l,\tau_K}|\lambda_1)$ is

$$\exp\left\{-\lambda_1 \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2\right\} \propto \int f_{\mathbf{N}}(\boldsymbol{\beta}_{l,\tau_K}|\mathbf{0}, s_l^2 I) f_\Gamma\left(s_l^2 \middle| \frac{K+1}{2}, \frac{2}{\lambda_1^2 \tilde{w}_l^2}\right) ds_l^2,$$

where f_N and f_Γ are the pdf of $N(\mathbf{0}, s_l^2 I)$ and $\Gamma\{(K+1)/2, 2/(\lambda_1 \tilde{w}_l)^2\}$, respectively, and $\Gamma(a, b)$ represents the gamma distribution with shape parameter a and scale parameter b. Similarly, the multivariate Laplace distribution $\pi(\mathbf{d}_{l,\tau_K}|\lambda_2)$ can be expressed as

$$\exp\left\{-\lambda_2 \tilde{v}_l ||\mathbf{d}_{l,\tau_K}||_2\right\} \propto \int f_{\mathbf{N}}(\mathbf{d}_{l,\tau_K}|\mathbf{0}, t_l^2 I) f_\Gamma\left(t_l^2 \middle| \frac{K}{2}, \frac{2}{\lambda_2^2 \tilde{v}_l^2}\right) dt_l^2,$$

where s_l^2 and t_l^2 are latent parameters.

For given s_l^2 and t_l^2, we can rearrange the quantile slopes for the l-th predictor and have the following identity

$$\begin{aligned}
&\sum_{k=1}^{K} \frac{\beta_{l,\tau_k}^2}{s_l^2} + \sum_{k=1}^{K-1} \frac{(\beta_{l,\tau_{k+1}} - \beta_{l,\tau_k})^2}{t_l^2} \\
=&\beta_{l,\tau_1}^2(s_l^{-2} + t_l^{-2}) + \beta_{l,\tau_2}^2(s_l^{-2} + t_l^{-2} + t_l^{-2}) + \cdots + \\
&\beta_{l,\tau_{K-1}}^2(s_l^{-2} + t_l^{-2} + t_l^{-2}) + \beta_{l,\tau_K}^2(s_l^{-2} + t_l^{-2}) - 2\sum_{k=1}^{K-1}(\beta_{l,\tau_{k+1}}\beta_{l,\tau_k})t_l^{-2}.
\end{aligned}$$

Hence, we can see that $\boldsymbol{\beta}_{l,\tau_K}|s_l^2, t_l^2$ is multivariate normal with mean $\mathbf{0}$ and precision matrix Σ_l^{-1}, where $\boldsymbol{\beta}_{l,\tau_K} = (\beta_{l,\tau_1}, ..., \beta_{l,\tau_K})^T$ and Σ_l^{-1} is a tridiagonal matrix with diagonal elements $\{(s_l^{-2} + t_l^{-2}), (s_l^{-2} + 2t_l^{-2}), ..., (s_l^{-2} + 2t_l^{-2}), (s_l^{-2} + t_l^{-2})\}$ and off diagonal elements $\{-t_l^{-2}, ..., -t_l^{-2}\}$. Thus, $(\boldsymbol{\beta}_{1,\tau_K}^T, ..., \boldsymbol{\beta}_{p,\tau_K}^T)^T$ follows multivariate normal with mean $\mathbf{0}$ and precision matrix Σ^{-1}, where $\Sigma^{-1} = diag(\Sigma_1^{-1}, ..., \Sigma_p^{-1})$ is a block diagonal matrix.

Using the latent parameters s_l^2 and t_l^2, we now have the following Bayesian hierarchical model

$$\begin{aligned}
(\boldsymbol{\beta}_{1,\tau_K}^T, ..., \boldsymbol{\beta}_{p,\tau_K}^T)^T|\mathbf{s}, \mathbf{t} &\sim N_{K\times p}(\mathbf{0}, \Sigma), \\
s_l^2|\lambda_{1l}^2 &\sim \Gamma\left(\frac{K+1}{2}, \frac{2}{\lambda_{1l}^2}\right), \\
t_l^2|\lambda_{2l}^2 &\sim \Gamma\left(\frac{K}{2}, \frac{2}{\lambda_{2l}^2}\right),
\end{aligned}$$

where $\mathbf{s} = (s_1, \ldots, s_p)^T$, $\mathbf{t} = (t_1, \ldots, t_p)^T$, $\lambda_{1l} = \tilde{w}_l\lambda_1$ and $\lambda_{2l} = \tilde{v}_l\lambda_2$.

Hierarchical Model

With the approximate conditional density function and prior distributions, the corresponding Bayesian hierarchical model is

$$\begin{aligned}
y|\mathbf{x}_i, \boldsymbol{\theta} &\sim \hat{f}(y|\mathbf{x}_i, \boldsymbol{\theta}), \\
(\alpha_{\tau_1}..., \alpha_{\tau_K})^T = \boldsymbol{\alpha}_K &\sim \prod_{k=1}^{K} \exp\left(-\frac{\alpha_{\tau_k}^2}{2\varsigma^2}\right), \\
(\boldsymbol{\beta}_{1,\tau_K}^T, ..., \boldsymbol{\beta}_{p,\tau_K}^T)^T|\mathbf{s}, \mathbf{t} &\sim N_{K\times p}(\mathbf{0}, \Sigma), \\
\mathbf{s}|\lambda_1^2 &\sim \prod_{l=1}^{p} \frac{(s_l^2)^{\frac{K+1}{2}-1}}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda_{1l}^2}\right)^{\frac{K+1}{2}}} \exp\left(-\frac{\lambda_{1l}^2 s_l^2}{2}\right), \\
\mathbf{t}|\lambda_2^2 &\sim \prod_{l=1}^{p} \frac{(t_l^2)^{\frac{K}{2}-1}}{\Gamma\left(\frac{K}{2}\right)\left(\frac{2}{\lambda_{2l}^2}\right)^{\frac{K}{2}}} \exp\left(-\frac{\lambda_{2l}^2 t_l^2}{2}\right), \\
\lambda_1^2 &\sim \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b), \\
\lambda_2^2 &\sim \frac{(\lambda_2^2)^{c-1}}{\Gamma(c)\, d^c} \exp(-\lambda_2^2/d).
\end{aligned}$$

MCMC Algorithm for Bayesian Fused Adaptive Group LASSO

1. Initialize $\boldsymbol{\alpha}_K$, $\boldsymbol{\beta}_K$, s_l^2, t_l^2, λ_1^2 and λ_2^2.

   (a) The initial values for $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ can be obtained by applying the conventional quantile regression at each τ_k, separately.

   (b) The initial values for s_l^2 and t_l^2 are drawn from GIG$(1/2, 1, ||\boldsymbol{\beta}_{l,\tau_K}^{(0)}||_2^2)$ and GIG$(1/2, 1, ||\mathbf{d}_{l,\tau_K}^{(0)}||_2^2)$, respectively, where $\boldsymbol{\beta}_{l,\tau_K}^{(0)}$ and $\mathbf{d}_{l,\tau_K}^{(0)}$ denote the initial estimates for the l-th predictor and GIG(q, u, v) is the generalized inverse Gaussian distribution with three parameters $u, v \in \mathbb{R}^+$ and $q \in \mathbb{R}$. The density function of GIG(q, u, v) is

$$f(x; u, v, q) = \frac{(u/v)^{q/2}}{2K_q(\sqrt{uv})} x^{(q-1)} e^{-(ux+v/x)/2} I(x > 0),$$

   where K_q is a modified Bessel function of the second kind (Jorgensen, 1982).

(c) The initial values for λ_1^2 and λ_2^2 are set to be 1. (Our simulation shows that the initial values of 0.1, 1 and 10 generate similar results.)

2. Update $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ by using the LIGPD in Chapter 2. The specifications of the proposal distributions for $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ are in Appendix A.1.3.

3. Update s_l^2 and t_l^2 by using Gibb's sampler.

   (a) Update s_l^2 using either generalized inverse Gaussian or inverse Gaussian distribution. The full conditional of s_l^2 is

$$\begin{aligned}
&f(s_l^2|\boldsymbol{\alpha}_K, \boldsymbol{\beta}_K, \mathbf{x}, \mathbf{y}, \mathbf{t}, \mathbf{s}_{-l}, \mathbf{t}, \lambda_1^2, \lambda_2^2) \\
&\propto \left(\frac{1}{2\pi s_l^2}\right)^{K/2} \exp\left(-\frac{||\boldsymbol{\beta}_{l,\tau_K}||_2^2}{2s_l^2}\right) \frac{(s_l^2)^{\frac{K+1}{2}-1}}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda_{1l}^2}\right)^{\frac{K+1}{2}}} \exp\left(-\frac{\lambda_{1l}^2 s_l^2}{2}\right) \\
&\propto (s_l^2)^{\frac{1}{2}-1} \exp\left(-\frac{||\boldsymbol{\beta}_{l,\tau_K}||_2^2}{2s_l^2} - \frac{\lambda_{1l}^2 s_l^2}{2}\right),
\end{aligned}$$

   which is the GIG$(1/2, \lambda_{1l}^2, ||\boldsymbol{\beta}_{l,\tau_K}||_2^2)$. Then, $1/s_l$ follows the GIG$(-1/2, ||\boldsymbol{\beta}_{l,\tau_K}||_2^2, \lambda_{1l}^2)$ =IG$\left(\sqrt{\frac{\lambda_{1l}^2}{||\boldsymbol{\beta}_{l,\tau_K}||_2^2}}, \lambda_{1l}^2\right)$.

   (b) Update t_l^2 by sampling from IG$\left(\sqrt{\frac{\lambda_{2l}^2}{||\mathbf{d}_{l,\tau_K}||_2^2}}, \lambda_{2l}^2\right)$ and taking the reciprocal.

4. Update λ_1^2 and λ_2^2 using Gibb's sampler.

   (a) Update λ_1^2 using $\Gamma\left(\frac{K+1}{2}+a, \left(\frac{1}{b}+\frac{1}{2}\sum_{l=1}^p \tilde{w}_l^2 s_l^2\right)^{-1}\right)$ since the full conditional of λ_1^2 is

$$\begin{aligned}
&f(\lambda_1^2|\boldsymbol{\alpha}_K, \boldsymbol{\beta}_K, \mathbf{x}, \mathbf{y}, \mathbf{s}, \mathbf{t}, \lambda_2^2) \\
&\propto \prod_{l=1}^p \frac{(s_l^2)^{\frac{K+1}{2}-1}}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda_{1l}^2}\right)^{\frac{K+1}{2}}} \exp\left(-\frac{\lambda_{1l}^2 s_l^2}{2}\right) \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b) \\
&\propto (\lambda_1^2)^{\frac{K+1}{2}+a-1} \exp\left\{-\left(\frac{1}{b}+\frac{1}{2}\sum_{l=1}^p \tilde{w}_l^2 s_l^2\right)\lambda_1^2\right\}.
\end{aligned}$$

   (b) Update λ_2^2 by sampling from $\Gamma\left(\frac{K}{2}+c, \left(\frac{1}{d}+\frac{1}{2}\sum_{l=1}^p \tilde{v}_l^2 t_l^2\right)^{-1}\right)$.

5. Repeat Steps 2-4 until the MCMC chain converges. We estimate each parameter by using the mean of the corresponding MCMC samples.

4.2.3 Bayesian Fused Adaptive LASSO

We propose a Bayesian fused adaptive LASSO estimator using Laplace priors on individual quantile slopes and interquantile differences. For convenience, we explain the proposed method in the frequentist framework and then discuss the proposed Bayesian method in detail.

Our second proposed method imposes a weighted fused LASSO penalty on the quantile slope coefficients and interquantile slope differences. In the frequentist framework, the fused adaptive LASSO estimator can be obtained by

$$\hat{\boldsymbol{\beta}}_K = \arg\min \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}\{y_i - \alpha_{\tau_k} - \mathbf{x}_i^T\boldsymbol{\beta}_k\} + \lambda_1 \sum_{k=1}^{K}\sum_{l=1}^{p} \tilde{w}_{k,l}|\beta_{l,\tau_k}| + \lambda_2 \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}|d_{l,\tau_{k+1}}|, \tag{4.4}$$

where λ_1 and λ_2 are regularization parameters, $\tilde{w}_{k,l}$ and $\tilde{v}_{k,l}$ are the adaptive weights for β_{l,τ_k} and d_{l,τ_k}, respectively. The adaptive weights are obtained by using the estimates $\hat{\beta}_{l,\tau_k}$ from the conventional quantile regression, and $\tilde{w}_{k,l} = |\hat{\beta}_{l,\tau_k}|^{-1}$ and $\tilde{v}_{k,l} = |\tilde{\beta}_{l,\tau_{k+1}} - \tilde{\beta}_{l,\tau_k}|^{-1}$.

Similar as in Section 4.2.2, with adaptive weights zero β_{l,τ_k} and d_{l,τ_k} will receive larger penalty and be shrunk to zero faster. In contrast to the fused adaptive group LASSO, the second method shrinks the quantile slope and the interquantile slope differences at each quantile level separately. In addition, quantile coefficients and interquantile differences from the same covariate will receive component-wise adaptive weights instead of common group-wise weights as in Section 4.2.2. Consequently, the frequentist fused adaptive LASSO method may lead to local sparsity of quantile slopes (e.g. a predictor has zero coefficients in some quantile region but non-zero at other regions) and local constancy (eg. a predictor has constant effect in some quantile region but non-constant in other regions).

We expect that our Bayesian fused adaptive LASSO estimator shrinks β_{l,τ_k} towards zero if $\beta_{l,\tau_k} = 0$ and d_{l,τ_k} towards zero if $d_{l,\tau_k} = 0$ but β_{l,τ_k} and d_{l,τ_k} cannot be exactly zero. Compared to the Bayesian fused adaptive group LASSO method, the number of adaptive weights and latent variables are K times more in the second method, which

poses some computational challenges especially for high dimensional data.

Prior Specifications

We put normal prior distributions on $\alpha_{\tau_1}, ..., \alpha_{\tau_K}$, Laplace priors on $\beta_{1,\tau_1}, \ldots, \beta_{p,\tau_1}, \ldots, \beta_{1,\tau_K}, \ldots, \beta_{p,\tau_K}$ and $d_{1,\tau_2}, \ldots, d_{p,\tau_2}, \ldots, d_{1,\tau_K}, \ldots, d_{p,\tau_K}$, and gamma prior distributions on λ_1^2 and λ_2^2:

$$\begin{aligned}
\pi(\alpha_{\tau_k}|\varsigma^2) &= \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left(-\frac{\alpha_{\tau_k}^2}{2\varsigma^2}\right), \\
\pi(\boldsymbol{\beta}_K|\lambda_1, \lambda_2) &\propto \exp\left\{-\lambda_1 \sum_{k=1}^{K}\sum_{l=1}^{p} \tilde{w}_{k,l}|\beta_{l,\tau_k}| - \lambda_2 \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}|d_{l,\tau_{k+1}}|\right\}, \\
\pi(\lambda_1^2) &\sim \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b), \\
\pi(\lambda_2^2) &\sim \frac{(\lambda_2^2)^{c-1}}{\Gamma(c)\, d^c} \exp(-\lambda_2^2/d),
\end{aligned}$$

where ς^2, a, b, c and d are some constants.

Using the scale mixture representation of the univariate Laplace distribution (see appendix A.3), the Laplace priors of $\beta_{1,\tau_1}, \ldots, \beta_{p,\tau_1}, \ldots, \beta_{1,\tau_K}, \ldots, \beta_{p,\tau_K}$ can be expressed as

$$\begin{aligned}
\exp\left(-\lambda_1 \sum_{k=1}^{K}\sum_{l=1}^{p} |\beta_{l,\tau_k}|\right) &\propto \prod_{k=1}^{K}\prod_{l=1}^{p} \exp\left(-\lambda_{1kl}|\beta_{l,\tau_k}|\right) \\
&= \prod_{k=1}^{K}\prod_{l=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi s_{k,l}^2}} \exp\left(-\frac{\beta_{l,\tau_k}^2}{2s_{k,l}^2}\right) \frac{\lambda_{1kl}^2}{2} \exp\left(-\frac{\lambda_{1kl}^2}{2} s_{k,l}^2\right) ds_{k,l}^2,
\end{aligned}$$

where $\lambda_{1kl} = \lambda_1 \tilde{w}_{k,l}$ and $s_{k,l}^2$ are the mixing parameters. The Laplace priors of $d_{1,\tau_2}, \ldots, d_{p,\tau_2}$, $\ldots$, $d_{1,\tau_K}, \ldots, d_{p,\tau_K}$ can be expressed as

$$\begin{aligned}
\exp\left(-\lambda \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}|d_{l,\tau_{k+1}}|\right) &\propto \prod_{k=1}^{K-1}\prod_{l=1}^{p} \exp\left(-\lambda_{2kl}|d_{l,\tau_{k+1}}|\right) \\
&= \prod_{k=1}^{K-1}\prod_{l=1}^{p} \int_0^\infty \frac{1}{\sqrt{2\pi t_{k,l}^2}} \exp\left(-\frac{d_{l,\tau_{k+1}}^2}{2t_{k,l}^2}\right) \frac{\lambda_{2kl}^2}{2} \exp\left(-\frac{\lambda_{2kl}^2}{2} t_{k,l}^2\right) dt_{k,l}^2,
\end{aligned}$$

where $\lambda_{2kl} = \lambda\tilde{v}_{k,l}$ and $t^2_{k,l}$ are the mixing parameters.

Then we can rearrange the quantile slopes for the l-th predictor and have the following identity

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{\beta^2_{l,\tau_k}}{s^2_{k,l}} + \sum_{k=1}^{K-1}\frac{(\beta_{l,\tau_{k+1}} - \beta_{l,\tau_k})^2}{t^2_{k,l}} \\
=&\beta^2_{l,\tau_1}(s^{-2}_{1,l} + t^{-2}_{1,l}) + \beta^2_{l,\tau_2}(s^{-2}_{2,l} + t^{-2}_{1,l} + t^{-2}_{2,l}) + \cdots + \\
&\beta^2_{l,\tau_{K-1}}(s^{-2}_{K-1,l} + t^{-2}_{K-2,l} + t^{-2}_{K-1,l}) + \beta^2_{l,\tau_k}(s^{-2}_{k,l} + t^{-2}_{k,l}) - 2\sum_{k=1}^{K-1}(\beta_{l,\tau_{k+1}}\beta_{l,\tau_k})t^{-2}_{k,l}.
\end{aligned}$$

Hence, we can see that $\boldsymbol{\beta}_{l,\tau_K}|s_{1,l},...,s_{k,l},t_{1,l},...,t_{K-1,l}$ is multivariate normal with mean $\mathbf{0}$ and precision matrix Σ_l^{-1}, where Σ_l^{-1} is a tridiagonal matrix with diagonal elements $\{(s^{-2}_{1l}+t^{-2}_{1l}),(s^{-2}_{2l}+t^{-2}_{1l}+t^{-2}_{2l}),...,(s^{-2}_{K-1,l}+t^{-2}_{K-2,l}+t^{-2}_{K-1,l}),(s^{-2}_{k,l}+t^{-2}_{K-1,l})\}$ and off diagonal elements $\{t^{-2}_{k,l}, k = 1,...,K-1\}$. Thus, $(\boldsymbol{\beta}^T_{1,\tau_K},...,\boldsymbol{\beta}^T_{p,\tau_K})^T$ follows multivariate normal with mean $\mathbf{0}$ and precision matrix Σ^{-1}, where $\Sigma^{-1} = diag(\Sigma_1^{-1},...,\Sigma_p^{-1})$.

Hierarchical Model

This leads to the following hierarchical model:

$$\begin{aligned}
y|\mathbf{x}_i &\sim f_i(y|\mathbf{x}_i,\boldsymbol{\theta},\lambda), \\
(\alpha_{\tau_1}...,\alpha_{\tau_K})^T &\sim \prod_{k=1}^{K}\exp\left(-\frac{\alpha^2_{\tau_k}}{2\varsigma^2}\right), \\
(\boldsymbol{\beta}^T_{1,\tau_K},...,\boldsymbol{\beta}^T_{p,\tau_K})^T|\mathbf{s},\mathbf{t} &\sim N_{K\times p}(\mathbf{0},\Sigma), \\
\mathbf{s}|\lambda_1^2 &\sim \prod_{k=1}^{K}\prod_{l=1}^{p}\frac{\lambda^2_{1kl}}{2}\exp\left(-\frac{\lambda^2_{1kl}}{2}s^2_{k,l}\right), \\
\mathbf{t}|\lambda_2^2 &\sim \prod_{k=1}^{K-1}\prod_{l=1}^{p}\frac{\lambda^2_{2kl}}{2}\exp\left(-\frac{\lambda^2_{2kl}}{2}t^2_{k,l}\right), \\
\lambda_1^2 &\propto \lambda_1^{2(a-1)}\exp(-\lambda_1^2/b), \\
\lambda_2^2 &\propto \lambda_2^{2(c-1)}\exp(-\lambda_2^2/d),
\end{aligned}$$

where $\mathbf{s} = (s^2_{11},\ldots,s^2_{1p},...,s^2_{K1},\ldots,s^2_{Kp})^T$ and $\mathbf{t} = (t^2_{11},\ldots,t^2_{1p},...,t^2_{K-1,1},\ldots,t^2_{K-1,p})^T$.

MCMC Algorithm for Bayesian Fused Adaptive LASSO

1. Initialize $\boldsymbol{\alpha}_K$, $\boldsymbol{\beta}_K$, $s_{k,l}^2$, $t_{k,l}^2$, λ_1^2 and λ_2^2.

   (a) The initial values for $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ can be obtained by applying the conventional quantile regression at each $\tau_k, k = 1, \ldots, K$, separately.

   (b) The initial values for $s_{k,l}^2$ and $t_{k,l}^2$ are drawn from $\text{GIG}(1/2, 1, |\beta_{l,\tau_k}^{(0)}|^2)$ and $\text{GIG}(1/2, 1, |d_{l,\tau_k}^{(0)}|^2)$, respectively, where $\beta_{l,\tau_k}^{(0)}$ and $d_{l,\tau_k}^{(0)}$ are the initial estimates for the l-th predictor at the τ_k-th quantile level.

   (c) The initial values for λ_1^2 and λ_2^2 are set to be 1. (Our simulation shows that the initial values of 0.1, 1 and 10 generate similar results.)

2. Update $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ by using the LIGPD.

3. Update $s_{k,l}^2$ and $t_{k,l}^2$ using Gibb's sampler.

   (a) Update $s_{k,l}^2$ using the $\text{GIG}(1/2, \lambda_{1l}^2, |\beta_{l,\tau_k}|^2)$ or update $1/s_{k,l}^2$ using the $\text{GIG}(-1/2, |\beta_{l,\tau_k}|^2, \lambda_{1l}^2) = \text{IG}\left(\sqrt{\frac{\lambda_{1l}^2}{|\beta_{l,\tau_k}|^2}}, \lambda_{1l}^2\right)$.

   (b) Update $t_{k,l}^2$ by sampling from $\text{IG}\left(\sqrt{\frac{\lambda_{2l}^2}{|d_{l,\tau_k}|^2}}, \lambda_{2l}^2\right)$ and taking the reciprocal.

4. Update λ_1^2 and λ_2^2 using Gibb's sampler.

   (a) Update λ_1^2 using $\Gamma\left(Kp + a, \left(1/b + \sum_{k=1}^{K}\sum_{l=1}^{p} \tilde{w}_{k,l}^2 s_{k,l}^2/2\right)^{-1}\right)$.

   (b) Update λ_2^2 using $\Gamma\left((K-1)p + c, \left(1/d + \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}^2 t_{k,l}^2/2\right)^{-1}\right)$.

5. Repeat Steps 2-4 until the MCMC chain converges. We estimate each parameter by using the mean of the corresponding MCMC samples.

4.2.4 Bayesian Fused Adaptive LASSO II

The third proposed method imposes a weighted group LASSO penalty on the quantile slope coefficients and a weighted LASSO penalty on interquantile slope differences. In

the frequentist framework, the fused adaptive LASSO II estimator can be obtained by

$$\hat{\boldsymbol{\beta}}_K = \arg\min \sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}\{y_i - \alpha_{\tau_k} - \mathbf{x}_i^T\boldsymbol{\beta}_k\} + \lambda_1 \sum_{l=1}^{p} \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2 + \lambda_2 \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}|d_{l,\tau_{k+1}}|, \tag{4.5}$$

where $\tilde{w}_l$ and $\tilde{v}_{k,l}$ are the adaptive weights for $\boldsymbol{\beta}_{l,\tau_K} = (\beta_{l,\tau_1}, \ldots, \beta_{l,\tau_K})^T$ and d_{l,τ_k}, respectively. The adaptive weights are obtained by using the estimates $\tilde{\boldsymbol{\beta}}_k$ from the conventional quantile regression, and we define $\tilde{w}_l = ||\tilde{\boldsymbol{\beta}}_{l,\tau_K}||_2^{-1}$ and $\tilde{v}_{k,l} = |\tilde{\beta}_{l,\tau_{k+1}} - \tilde{\beta}_{l,\tau_k}|^{-1}$. The group LASSO penalty on $\boldsymbol{\beta}_{l,\tau_K}$ will eliminate the predictors that have no impact across all quantile levels. The LASSO penalty on d_{l,τ_k} will shrink the interquantile slope differences at each quantile level separately and thus can lead to local constant quantile slopes.

The Bayesian fused adaptive LASSO II estimator can shrink β_{l,τ_k} towards zero but not exactly to zero if $||\boldsymbol{\beta}_{l,\tau_K}||_2 = 0$ and d_{l,τ_k} towards zero but not exactly to zero if $d_{l,\tau_k} = 0$.

Prior Specifications

We put normal prior distributions on $\alpha_{\tau_1}, ..., \alpha_{\tau_K}$, spherical multivariate Laplace prior distributions on $\boldsymbol{\beta}_{1,\tau_K}, \ldots, \boldsymbol{\beta}_{p,\tau_K}$ and univariate Laplace prior distributions on $d_{1,\tau_2}, \ldots, d_{p,\tau_2}, \ldots, d_{1,\tau_K}, \ldots, d_{p,\tau_K}$, and gamma prior distributions on λ_1^2 and λ_2^2:

$$\begin{aligned}
\pi(\alpha_{\tau_k}|\varsigma^2) &= \frac{1}{\sqrt{2\pi\varsigma^2}} \exp\left(-\frac{\alpha_{\tau_k}^2}{2\varsigma^2}\right), \\
\pi(\boldsymbol{\beta}_K|\lambda_1,\lambda_2) &\propto \exp\left\{-\lambda_1 \sum_{l=1}^{p} \tilde{w}_l ||\boldsymbol{\beta}_{l,\tau_K}||_2 - \lambda_2 \sum_{k=1}^{K-1}\sum_{l=1}^{p} \tilde{v}_{k,l}|d_{l,\tau_{k+1}}|\right\}, \\
\pi(\lambda_1^2) &\sim \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b), \\
\pi(\lambda_2^2) &\sim \frac{(\lambda_2^2)^{c-1}}{\Gamma(c)\, d^c} \exp(-\lambda_2^2/d),
\end{aligned}$$

where ς^2, a, b, c and d are some constants.

For given latent parameters s_l^2 and $t_{k,l}^2$, we can rearrange the quantile slopes for the

l-th predictor and have the following identity

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{\beta_{l,\tau_k}^2}{s_l^2}+\sum_{k=1}^{K-1}\frac{(\beta_{l,\tau_{k+1}}-\beta_{l,\tau_k})^2}{t_{k,l}^2}\\
=&\beta_{l,\tau_1}^2(s_l^{-2}+t_{1,l}^{-2})+\beta_{l,\tau_2}^2(s_l^{-2}+t_{1,l}^{-2}+t_{2l}^{-2})+\cdots+\\
&\beta_{l,\tau_{K-1}}^2(s_l^{-2}+t_{K-2,l}^{-2}+t_{K-1,l}^{-2})+\beta_{l,\tau_k}^2(s_l^{-2}+t_{K-1,l}^{-2})-2\sum_{k=1}^{K-1}(\beta_{l,\tau_{k+1}}\beta_{l,\tau_k})t_{k,l}^{-2}.
\end{aligned}$$

Hence, we can see that $\boldsymbol{\beta}_{l,\tau_K}|s_l^2,t_{k,l}^2$ is multivariate normal with mean $\mathbf{0}$ and precision matrix Σ_l^{-1}, where $\boldsymbol{\beta}_{l,\tau_K}=(\beta_{l,\tau_1},...,\beta_{l,\tau_K})$ and Σ_l^{-1} is a tridiagonal matrix with diagonal elements $\{(s_l^{-2}+t_{1,l}^{-2}),(s_l^{-2}+t_{1,l}^{-2}+t_{2,l}^{-2}),...,(s_l^{-2}+t_{K-2,l}^{-2}+t_{K-1,l}^{-2}),(s_l^{-2}+t_{K-1,l}^{-2})\}$ and off diagonal elements $\{-t_{1l}^{-2},...,-t_{K-1,l}^{-2}\}$. Thus, after rearranging $\boldsymbol{\beta}_K$, $(\boldsymbol{\beta}_{1,\tau_K}^T,...,\boldsymbol{\beta}_{p,\tau_K}^T)^T$ follows multivariate normal with mean $\mathbf{0}$ and precision matrix Σ^{-1}, where $\Sigma^{-1}=diag(\Sigma_1^{-1},...,\Sigma_p^{-1})$. Hence, Bayesian hierarchical model for $\boldsymbol{\beta}_K$ is

$$\begin{aligned}
(\boldsymbol{\beta}_{1,\tau_K}^T,...,\boldsymbol{\beta}_{p,\tau_K}^T)^T|\mathbf{s},\mathbf{t}&\sim N_{K\times p}(\mathbf{0},\Sigma),\\
s_l^2|\lambda_{1l}^2&\sim\Gamma\left(\frac{K+1}{2},\frac{2}{\lambda_{1l}^2}\right),\\
t_{k,l}^2|\lambda_{2kl}^2&\sim\exp\left(\frac{2}{\lambda_{2kl}^2}\right),
\end{aligned}$$

where $\mathbf{s}=(s_1^2,\ldots,s_p^2)^T$, $\mathbf{t}=(t_{1,1}^2,\ldots,t_{1,p}^2,...,t_{K-1,1}^2,\ldots,t_{K-1,p}^2)^T$, $\lambda_{1l}=\tilde{w}_l\lambda_1$ and $\lambda_{2kl}=\tilde{v}_{k,l}\lambda_2$.

Hierarchical Model

With the approximate conditional density function and prior specifications, the corresponding Bayesian hierarchical model can be expressed as

$$
\begin{aligned}
y|\mathbf{x}_i, \boldsymbol{\theta} &\sim \hat{f}(y|\mathbf{x}_i, \boldsymbol{\theta}), \\
\alpha_{\tau_1}, ..., \alpha_{\tau_K} = \boldsymbol{\alpha}_K &\sim \prod_{k=1}^{K} \exp\left(-\frac{\alpha_{\tau_k}^2}{2\varsigma^2}\right), \\
(\boldsymbol{\beta}_{1,\tau_K}^T, ..., \boldsymbol{\beta}_{p,\tau_K}^T)^T|\mathbf{s}, \mathbf{t} &\sim N_{K\times p}(\mathbf{0}, \Sigma), \\
\mathbf{s}|\lambda_{1l}^2 &\sim \prod_{l=1}^{p} \frac{(s_l^2)^{\frac{K+1}{2}-1}}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda_{1l}^2}\right)^{\frac{K+1}{2}}} \exp\left(-\frac{\lambda_{1l}^2 s_l^2}{2}\right), \\
\mathbf{t}|\lambda_2^2 &\sim \prod_{k=1}^{K-1}\prod_{l=1}^{p} \frac{\lambda_{2kl}^2}{2} \exp\left(-\frac{\lambda_{2kl}^2}{2} t_{k,l}^2\right), \\
\lambda_1^2 &\sim \frac{(\lambda_1^2)^{a-1}}{\Gamma(a)\, b^a} \exp(-\lambda_1^2/b), \\
\lambda_2^2 &\sim \frac{(\lambda_2^2)^{c-1}}{\Gamma(c)\, d^c} \exp(-\lambda_2^2/d).
\end{aligned}
$$

MCMC Algorithm for Bayesian Fused Adaptive LASSO II

1. Initialize $\boldsymbol{\alpha}_K$, $\boldsymbol{\beta}_K$, $s_{k,l}^2$, $t_{k,l}^2$, λ_1^2 and λ_2^2.

   (a) The initial values for $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ can be obtained by applying the conventional quantile regression at each τ_k, separately.

   (b) The initial values for s_l^2 and $t_{k,l}^2$ are drawn from $\text{GIG}(1/2, 1, ||\boldsymbol{\beta}_{l,\tau_K}^{(0)}||_2^2)$ and $\text{GIG}(1/2, 1, |d_{l,\tau_k}^{(0)}|^2)$, respectively, where $\beta_{l,\tau_k}^{(0)}$ and $d_{l,\tau_k}^{(0)}$ are the initial estimates for the l-th predictor at the τ_k-th quantile level.

   (c) The initial values for λ_1^2 and λ_2^2 are 1.

2. Update $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ using the LIGPD in Chapter 2.

3. Update s_l^2 and $t_{k,l}^2$ using Gibb's sampler.

   (a) Update s_l^2 using the $\text{GIG}(1/2, \lambda_{1l}^2, ||\boldsymbol{\beta}_{l,\tau_K}||_2^2)$ or update $1/s_l^2$ using the $\text{GIG}(-1/2, ||\boldsymbol{\beta}_{l,\tau_K}||_2^2, \lambda_{1l}^2)$ $=\text{IG}\left(\sqrt{\frac{\lambda_{1l}^2}{||\boldsymbol{\beta}_{l,\tau_K}||_2^2}}, \lambda_{1l}^2\right)$.

(b) Update $t_{k,l}^2$ by sampling from $\text{IG}\left(\sqrt{\frac{\lambda_{2l}^2}{|d_{l,\tau_k}|^2}}, \lambda_{2l}^2\right)$ and taking the reciprocal.

4. Update λ_1^2 and λ_2^2 using Gibb's sampler.

   (a) Update λ_1^2 using $\Gamma\left(\frac{K+1}{2}+a, \left(\frac{1}{b}+\frac{1}{2}\sum_{l=1}^{p}\tilde{w}_l^2 s_l^2\right)^{-1}\right)$.

   (b) Update λ_2^2 using $\Gamma\left((K-1)p+c, \left(1/d+\sum_{k=1}^{K-1}\sum_{l=1}^{p}\tilde{v}_{k,l}^2 t_{k,l}^2/2\right)^{-1}\right)$.

5. Repeat Steps 2-4 until the MCMC chain converges. We estimate each parameter by using the mean of the corresponding MCMC samples.

4.3 Simulation Studies

We carried out simulation studies to compare the performance of the three proposed methods and the conventional quantile regression. We refer to the conventional quantile regression method as QR, the Bayesian fused adaptive group LASSO as BFGLGL, the Bayesian fused adaptive LASSO as BFLL and the Bayesian fused adaptive LASSO II as BFGLL.

For each scenario, the simulation is repeated 500 times and a grid of nine equally spaced quantile levels, $(0.1, 0.2, \ldots, 0.9)$, is considered. We evaluate the three proposed methods using the mean of integrated squared error (MISE) and the mean squared errors (MSE) of quantile slopes. The MISE is defined as the average of ISE over 500 simulations, where

$$ISE = \frac{1}{n}\sum_{i=1}^{n}\left\{(\alpha_{\tau_k}+\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_k})-(\hat{\alpha}_{\tau_k}+\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau_k})\right\}^2.$$

The MISE evaluates the overall estimation efficiency of an estimator for estimating the τ_k-th conditional quantile given $\mathbf{x}_i, i = 1, \ldots, n$. We measure the individual slope estimation efficiency by the mean squared errors. The MSE of a quantile slope $\beta_{l,\tau_k}, k = 1, \ldots, K, l = 1, \ldots, p$ is defined as

$$\frac{1}{500}\sum_{m=1}^{500}(\hat{\beta}_{l,\tau_k}-\beta_{l,\tau_k})^2,$$

where $\hat{\beta}_{l,\tau_k}$ and β_{l,τ_k} are the estimated and true τ_k-th quantile slope associated with the l-th predictor, respectively.

4.3.1 Simulation Study I

We consider the following model for generating the simulation data

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + h(x_{i1})e_i, i = 1, \ldots, 200, \tag{4.6}$$

where $\beta_0 = \beta_1 = 1$, $\beta_2 = \beta_3 = 0$, $x_{ij} \overset{i.i.d.}{\sim} U(0,1)$, $e_i \overset{i.i.d.}{\sim} N(0,1)$ and $h(x_{i1}) = 1, 1 + I(\tau \geq 0.5)x_{i1}$, or $1 + x_{i1}$. The corresponding conditional quantile function is

$$Q_\tau(Y_i|\mathbf{x}_i) = \beta_{0,\tau} + \beta_{1,\tau} x_{i1} + \beta_{2,\tau} x_{i2} + \beta_{3,\tau} x_{i3},$$

where $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{2,\tau} = \beta_{3,\tau} = 0$ and

$$\beta_{1,\tau} = \begin{cases} 1 & \text{if } h(x_{i1}) = 1, \\ 1 + I(\tau \geq 0.5)\Phi^{-1}(\tau) & \text{if } h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}, \\ 1 + \Phi^{-1}(\tau) & \text{if } h(x_{i1}) = 1 + x_{i1}. \end{cases}$$

If $h(x_{i1}) = 1$, $\beta_{1,\tau}$ are constant across all quantile levels. If $h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}$, $\beta_{1,\tau}$ remain constant for $\tau < 0.5$ and vary with τ otherwise. If $h(x_{i1}) = 1 + x_{i1}$, $\beta_{1,\tau}$ vary with τ across the entire quantile region.

We put independent normal priors $N(\mathbf{0}, 100)$ on $\alpha_{\tau_1}, \ldots, \alpha_{\tau_K}$ and we use $\Gamma(1, 0.1)$ prior on λ_1^2 and λ_2^2. As discussed in Kyung et al. (2010), the prior distribution $\Gamma(1, 0.1)$ is relatively flat and gives high posterior density near the MLE. With 100,000 MCMC samples, we discard the first 25,000 samples as burn-in and estimate $\beta_{\tau,j}$ for $j = 0, 1, 2, 3$ by the posterior means of the thinned sample using every 100th MCMC sample.

Simulation Results

Table 4.1 summarizes the mean of integrated squared errors (MISE) of the QR and three proposed methods across 500 data sets. Figures 4.1 – 4.3 plot the mean squared errors (MSEs) of the QR and three proposed methods and 95% pointwise confidence bands of the MSE of the QR estimator. Figure 4.4 shows the box plots of $\hat{\beta}_{1,\tau}$ and Table 4.2 summarizes the biases of $\hat{\beta}_{l,\tau}, l = 0, 1, 2, 3$ when $h(x_{i1}) = 1$.

Figure 4.1 shows that the proposed estimators have similar efficiency at central quantiles and slightly higher efficiency at tails than the QR estimator for estimating the con-

stant non-zero quantile slope $\beta_{1,\tau}$ when $h(x_{i1}) = 1$. Figure 4.4 shows that the proposed methods slightly overshrink $\beta_{1,\tau}$ towards zero since all three proposed methods penalize quantile slopes. The variances of $\hat{\beta}_{1,\tau}$ for the BFLL, BFGLL and BFGLGL estimators are smaller than for the QR estimator, especially at tails. Thus, the proposed methods yield higher efficiency for estimating $\beta_{1,\tau}$ at tails even though the BFLL, BFGLL and BFGLGL estimators have slightly higher biases than the QR in Table 4.2. When $h(x_{i1}) = 1$, we expect that BFGLGL performs the best since it can capture global sparsity and constancy in the quantile slope. However, BFGLGL shrinks $\beta_{1,\tau}$ towards zero more than necessary. Instead, the BFGLL estimator has smaller biases and MSEs than the other two proposed methods.

Figures 4.2 – 4.3 show that the QR estimator has similar or slightly higher efficiency than the proposed methods for estimating the non-constant and non-zero quantile slope $\beta_{1,\tau}$ when $h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}$ or $1 + x_{i1}$. When $h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}$, we expect that the BFLL and BFGLL methods performs better than BFGLGL since the BFLL and BFGLL estimators can capture non-zero local constancy of $\beta_{1,\tau}$. However, we do not observe the advantage of using the proposed methods. Figure 4.5 shows that, at $\tau < 0.5$, the three proposed methods have similar variances and biases of $\hat{\beta}_{1,\tau}$ compared to the QR and, at $\tau > 0.5$, the three proposed methods have smaller variances and higher biases for estimating $\hat{\beta}_{1,\tau}$ than QR due to shrinkage towards constant. When $h(x_{i1}) = 1 + x_{i1}$, the proposed methods also have higher biases and smaller variances for estimating $\beta_{1,\tau}$ and thus we observe no improvement over the QR for estimating $\beta_{1,\tau}$.

Across all three scenarios considered, the proposed methods yield significantly smaller MISEs than the QR method because the proposed methods have higher efficiency for estimating the zero quantile slopes $\beta_{2,\tau}$ and $\beta_{3,\tau}$. The BFGLL estimator has smaller biases and MSEs than the other two proposed methods for estimating the non-zero constant and sparse quantile slopes at the lower and upper quantiles. Hence, the BFGLL method shows smaller MISEs at tails than the other methods. In addition, the higher efficiency of the proposed methods for estimating quantile intercepts can be explained by the compounding effects of joint-modeling of LIGPD and higher efficiency for estimating other quantile slope coefficients.

Table 4.1: A thousand times the mean of integrated squared errors ($10^3\times$ MISE) across 500 simulations for data generated from model (4.6) from the conventional quantile regression method (QR), the Bayesian fused adaptive LASSO (BFLL), the Bayesian fused adaptive LASSO II (BFGLL) and the Bayesian fused adaptive group LASSO (BFGLGL). The values in the parentheses are the standard errors of $10^3\times$ MISE.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
					$h(x_{i1}) = 1$				
BFLL	48.64	32.7	27.01	24.03	24	24.77	27.76	32.64	46.29
	(1.87)	(1.29)	(1.02)	(0.89)	(0.87)	(0.89)	(1.02)	(1.20)	(1.70)
BFGLL	38.75	30.17	25.41	23.37	23.61	24.15	25.39	29.09	35.28
	(1.53)	(1.23)	(0.95)	(0.83)	(0.84)	(0.85)	(0.86)	(0.99)	(1.22)
BFGLGL	47.36	34.67	27.45	23.4	23.24	24.47	27.28	34.64	46.63
	(1.87)	(1.38)	(0.94)	(0.78)	(0.78)	(0.85)	(0.93)	(1.18)	(1.74)
QR	60.09	41.49	33.16	30.46	32.26	33.37	35.32	42.18	59.27
	(2.01)	(1.37)	(1.08)	(0.96)	(0.99)	(0.99)	(1.11)	(1.33)	(2.05)
					$h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}$				
BFLL	46.78	32.57	27.83	27.21	40.56	48.82	59.44	74.5	117.54
	(1.75)	(1.27)	(1.08)	(1.10)	(1.70)	(1.83)	(2.26)	(2.66)	(4.26)
BFGLL	40.15	31.72	27.95	26.71	36.58	43.63	51.54	65.78	96.23
	(1.51)	(1.24)	(1.03)	(1.00)	(1.53)	(1.59)	(1.89)	(2.26)	(3.33)
BFGLGL	47.39	34.1	28.23	26.94	41.37	49.12	60.17	79.26	118.06
	(1.83)	(1.26)	(1.01)	(1.01)	(1.73)	(1.86)	(2.19)	(2.73)	(4.16)
QR	60.19	41.53	33.27	33.08	50.97	68.58	76.82	93.13	127.83
	(2.00)	(1.37)	(1.10)	(1.15)	(1.84)	(2.03)	(2.58)	(3.00)	(4.51)
					$h(x_{i1}) = 1 + x_{i1}$				
BFLL	107.33	67.27	54.38	49.66	49.95	52.43	63.97	79.26	122.87
	(4.34)	(2.60)	(2.08)	(1.83)	(1.79)	(1.82)	(2.25)	(2.68)	(4.20)
BFGLL	104.18	73.98	59.11	52.25	52.58	55.83	65.14	82.42	115.4
	(3.81)	(2.97)	(2.25)	(1.90)	(1.90)	(2.02)	(2.34)	(2.75)	(3.91)
BFGLGL	114.64	77.66	60.54	51.54	51.03	54.88	66.53	85.79	130.85
	(4.48)	(3.00)	(2.19)	(1.80)	(1.82)	(1.94)	(2.30)	(2.82)	(4.56)
QR	129.48	88.79	70.41	66.69	68.67	72.3	77.15	93.13	127.83
	(4.24)	(3.03)	(2.33)	(2.17)	(2.21)	(2.25)	(2.63)	(3.00)	(4.51)

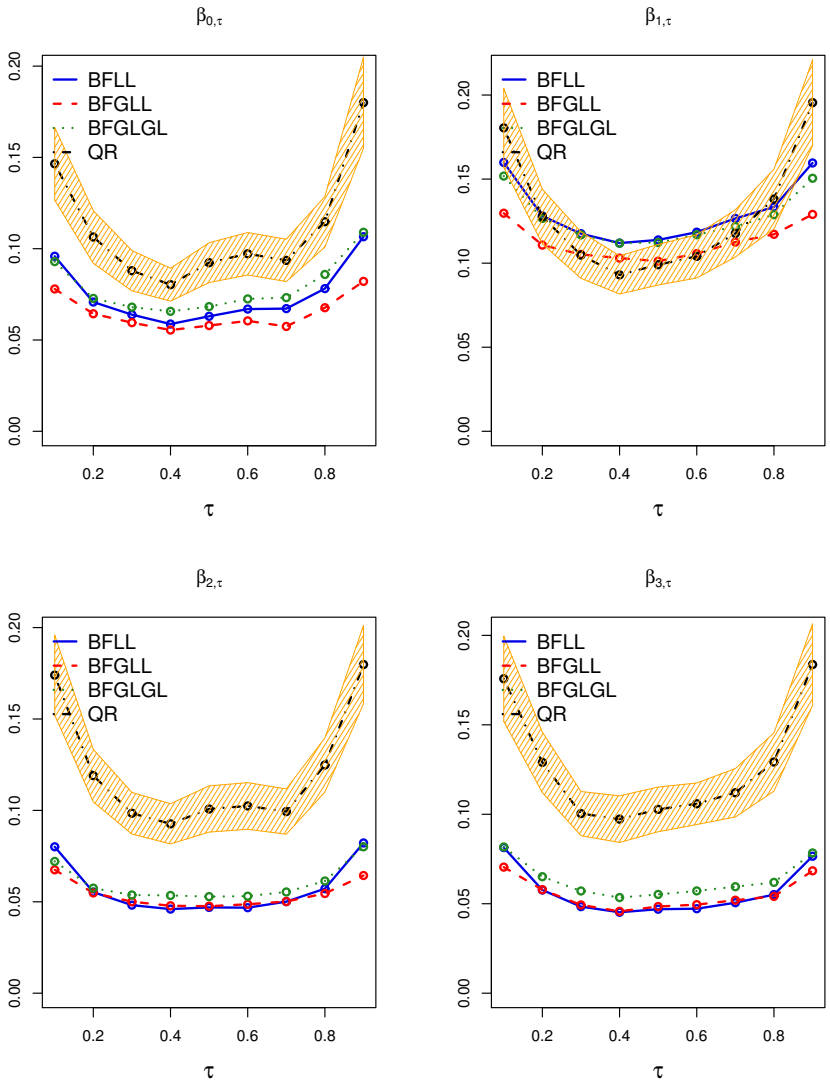


Figure 4.1: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.6) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = 1$ and $\beta_{2,\tau} = \beta_{3,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

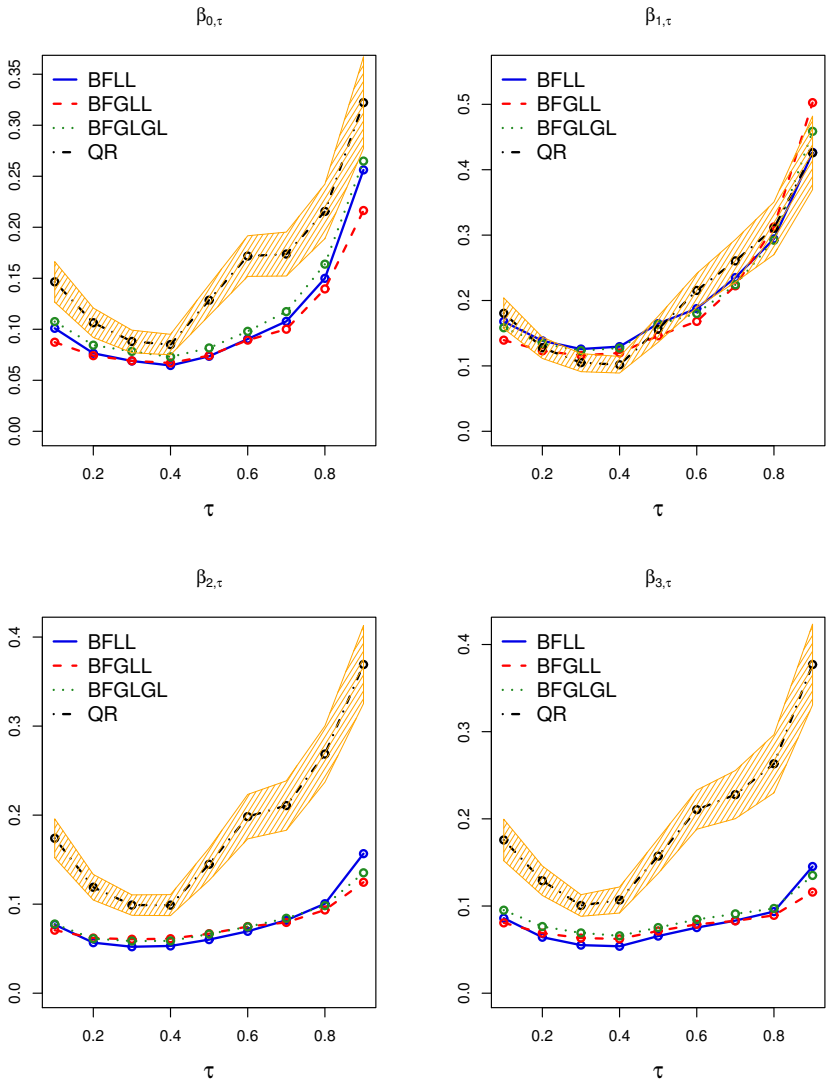


Figure 4.2: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.6) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = 1 + I(\tau \geq 0.5)\Phi^{-1}(\tau)$ and $\beta_{2,\tau} = \beta_{3,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

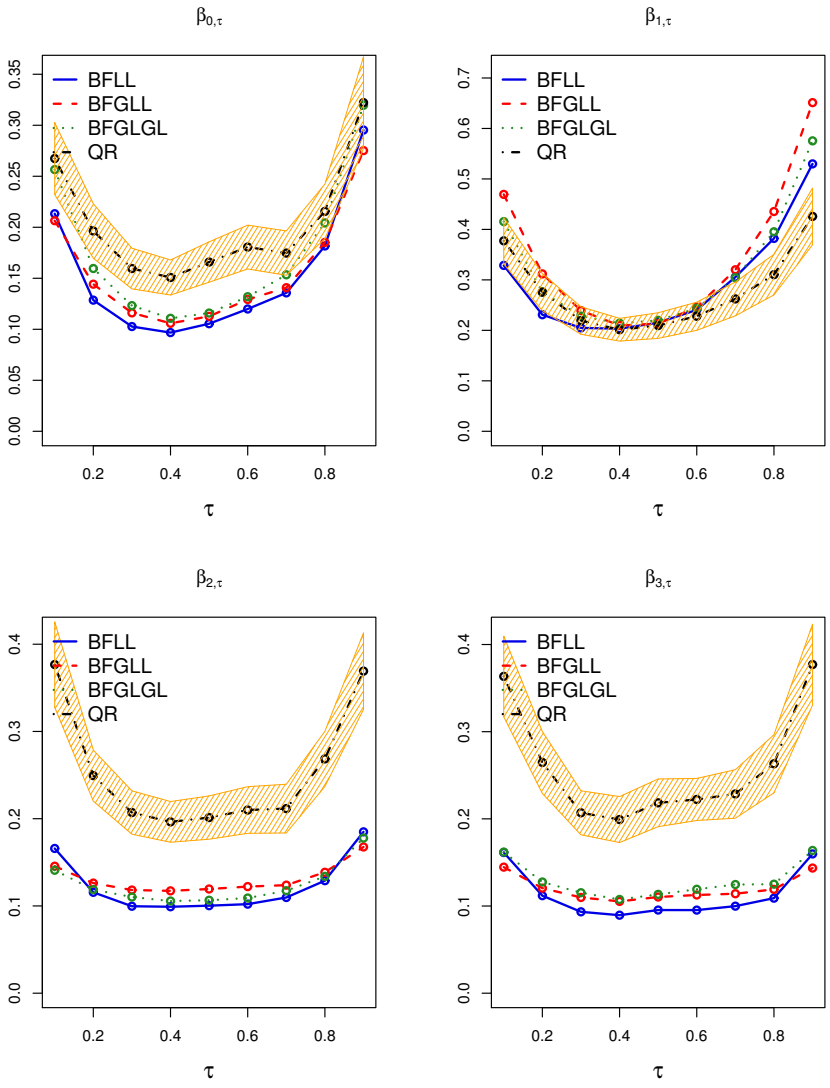


Figure 4.3: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.6) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = 1 + \Phi^{-1}(\tau)$ and $\beta_{2,\tau} = \beta_{3,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

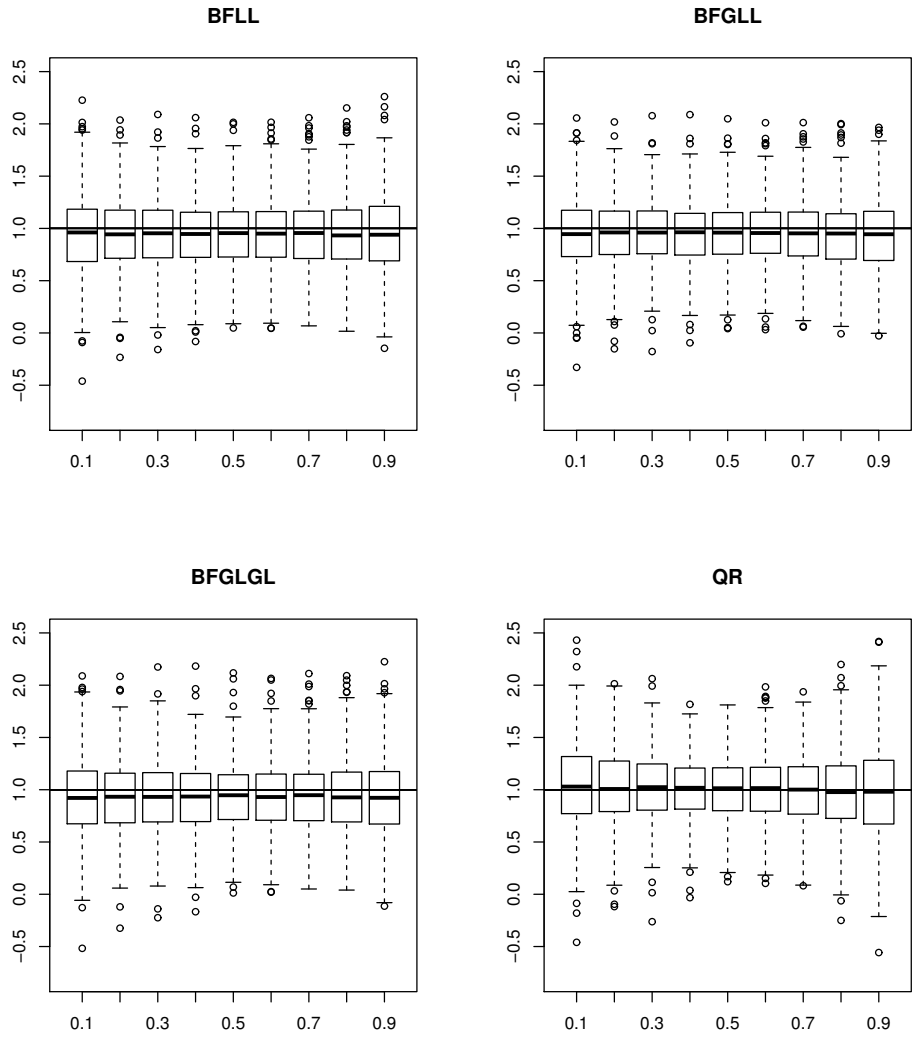


Figure 4.4: Boxplots of $\hat{\beta}_{1,\tau}$ from BFLL, BFGLL, BFGLGL and QR estimators in model (4.6) with $h(x_{i1}) = 1$. The x-axis represents the quantile levels and the y-axis represents the estimated quantile slope for $\beta_{1,\tau}$. The horizontal lines indicate the true $\beta_{1,\tau}$.

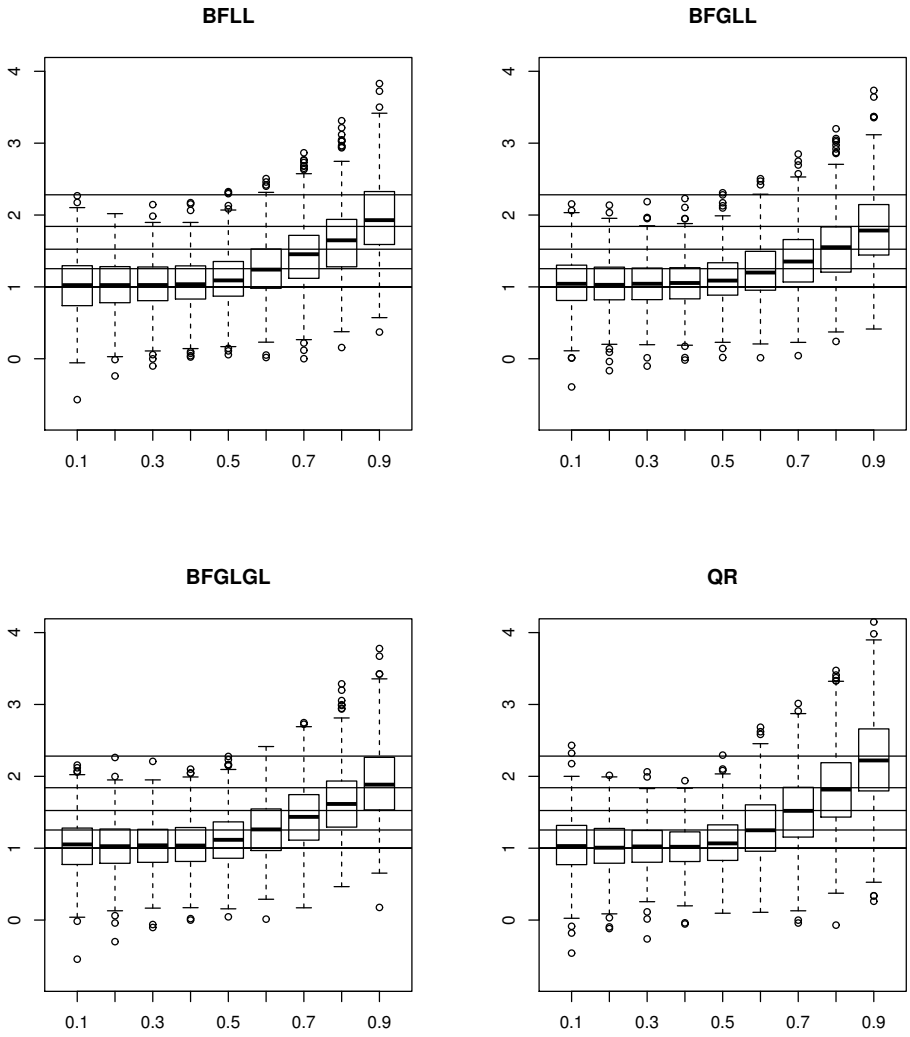


Figure 4.5: Boxplots of $\hat{\beta}_{1,\tau}$ from BFLL, BFGLL, BFGLGL and QR estimators in model (4.6) with $h(x_{i1}) = 1 + I(\tau \geq 0.5)x_{i1}$. The x-axis represents the quantile levels and the y-axis represents the estimated quantile slope for $\beta_{1,\tau}$. The horizontal lines indicate the true $\beta_{1,\tau}$. For example, if the center of the last box plot in each graph is at the top horizontal line, $\hat{\beta}_{0.9,1}$ is unbiased.

Table 4.2: $10^2\times$ bias of $\hat{\beta}_{l,\tau}, l = 0, 1, 2, 3$ from the QR, BFLL, BFGLL and BFGLGL estimators in model (4.6) with $h(x_{i1}) = 1$.

Method	τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
BFLL	$\beta_{0,\tau}$	-2.63	-0.68	0.28	1.01	2.59	4.53	6.15	7.64	9.90
	$\beta_{1,\tau}$	-7.11	-6.14	-5.48	-5.27	-5.11	-5.19	-5.38	-5.95	-6.42
	$\beta_{2,\tau}$	-1.91	-2.05	-1.69	-0.72	-0.62	-0.14	0.71	2.09	2.96
	$\beta_{3,\tau}$	-1.08	-0.36	-0.25	0.36	0.68	0.67	0.35	0.44	0.41
BFGLL	$\beta_{0,\tau}$	-0.45	-0.43	0.13	0.77	1.43	2.70	3.63	4.85	5.06
	$\beta_{1,\tau}$	-4.58	-4.11	-3.73	-4.05	-3.81	-3.99	-4.31	-5.32	-6.24
	$\beta_{2,\tau}$	-1.41	-1.62	-1.66	-1.17	-0.99	-0.73	0.05	1.17	2.44
	$\beta_{3,\tau}$	1.26	1.34	1.31	1.21	1.59	1.40	0.71	-0.06	-0.31
BFGLGL	$\beta_{0,\tau}$	-2.82	-0.26	1.10	2.04	3.70	5.80	7.55	9.41	11.37
	$\beta_{1,\tau}$	-7.69	-7.69	-7.15	-6.74	-6.42	-6.46	-6.62	-7.07	-7.25
	$\beta_{2,\tau}$	-2.59	-2.91	-2.49	-1.64	-1.29	-0.67	0.25	1.46	2.07
	$\beta_{3,\tau}$	-0.48	-0.57	-0.50	0.13	0.60	0.71	0.54	0.51	0.58
QR	$\beta_{0,\tau}$	-0.86	-0.08	-0.58	-0.58	0.04	-0.68	-0.00	1.30	-0.30
	$\beta_{1,\tau}$	2.68	2.85	3.55	2.00	2.06	2.12	-0.35	-1.72	-3.34
	$\beta_{2,\tau}$	-0.86	-1.64	-2.18	-1.31	-2.26	-1.47	-0.19	1.14	5.07
	$\beta_{3,\tau}$	1.57	0.03	1.12	0.70	0.50	0.59	-0.15	-2.07	-2.34

4.3.2 Simulation Study II

We consider the following model for generating the simulation data

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + h(\mathbf{x})e_i, i = 1, \ldots, 200, \tag{4.7}$$

where $\beta_0 = \beta_1 = \beta_2 = 0$, $x_{ij} \sim U(0,1)$, $x_{ij} \overset{i.i.d.}{\sim} U(0,1)$, $e_i \overset{i.i.d.}{\sim} N(0,1)$ and $h(\mathbf{x}_i) = 1 + I(\tau \geq 0.5)x_{i1}$ or $1 + I(\tau \geq 0.5)x_{i1} + I(\tau \geq 0.5)x_{i2}$. The corresponding conditional quantile function is

$$Q_\tau(Y_i|\mathbf{x}_i) = \beta_{0,\tau} + \beta_{1,\tau} x_{i1} + \beta_{2,\tau} x_{i2},$$

where $\beta_{0,\tau} = \Phi^{-1}(\tau)$, $\beta_{1,\tau} = I(\tau \geq 0.5)\Phi^{-1}(\tau)$ and

$$\beta_{2,\tau} = \begin{cases} 0 & \text{if } h(\mathbf{x}) = 1 + I(\tau \geq 0.5)x_{i1}, \\ I(\tau \geq 0.5)\Phi^{-1}(\tau) & \text{if } h(\mathbf{x}) = 1 + I(\tau \geq 0.5)x_{i1} + I(\tau \geq 0.5)x_{i2}. \end{cases}$$

We put independent normal priors $N(\mathbf{0}, 100)$ on $\alpha_{\tau_1}, \ldots, \alpha_{\tau_K}$ and we use $\Gamma(1, 0.1)$ prior on λ_1^2 and λ_2^2. With 100,000 MCMC samples, we discard the first 25,000 samples as burn-in and estimate $\beta_{\tau,j}$ for $j = 0, 1, 2$ by the posterior means of the thinned sample using every 100th MCMC sample.

Table 4.3 summarizes the mean of integrated squared errors (MISE) of the QR and three proposed methods. Figures 4.6 – 4.7 plot the mean squared errors (MSEs) of the QR and three proposed methods and 95% pointwise confidence bands of the MSE of the QR estimator.

For model (4.7) with $h(\mathbf{x}_i) = 1 + I(\tau \geq 0.5)x_{i1}$, the proposed methods yield significantly smaller MISE than the QR method because the proposed estimators have higher efficiency for estimating the zero quantile slope $\beta_{2,\tau}$. Since the BFGLL and BFGLGL methods regularize $\beta_{1,\tau}$ with a group LASSO penalty, these two proposed methods are not able to capture local sparsity of $\beta_{1,\tau}$ at $\tau \leq 0.5$. Generally speaking, the BFLL estimator have the smallest MSEs for estimating $\beta_{1,\tau}$ at the lower and central quantiles.

For model (4.7) with $h(\mathbf{x}) = 1 + I(\tau \geq 0.5)x_{i1} + I(\tau \geq 0.5)x_{i2}$, the BFLL estimator has smaller MISE than the others because it better captures the local sparsity of $\beta_{1,\tau}$ and $\beta_{2,\tau}$. The BFGLL and BFGLGL generally give more efficient estimation than the QR, possibly due to the joint-modeling of multiple quantiles using LIGPD and the quantile

Table 4.3: A thousand times the mean of integrated squared errors ($10^3\times$ MISE) for data generated from model (4.7) from the conventional quantile regression method (QR), the Bayesian fused adaptive LASSO (BFLL), the Bayesian fused adaptive LASSO II (BFGLL) and the Bayesian fused adaptive group LASSO (BFGLGL). The values in the parentheses are the standard errors of $10^3\times$ MISE.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$h(\mathbf{x}_i) = 1 + I(\tau \geq 0.5)x_{i1}$									
BFLL	19.55	13.39	11.02	9.77	15.4	25.09	33.64	51.38	86.17
	(0.98)	(0.73)	(0.56)	(0.53)	(0.96)	(1.24)	(1.49)	(1.86)	(2.96)
BFGLL	31.18	24.31	22.23	22.26	31.68	38.59	43.54	55.91	80.84
	(1.30)	(1.03)	(0.91)	(0.99)	(1.60)	(1.64)	(1.93)	(2.21)	(2.83)
BFGLGL	34.26	24.83	21.94	21.6	34.38	40.48	48.17	61.72	86.07
	(1.40)	(1.09)	(0.91)	(0.94)	(1.62)	(1.75)	(2.07)	(2.46)	(3.05)
QR	43.94	30.89	25.63	25.07	37.75	54.44	59.98	66.74	90.58
	(1.71)	(1.16)	(1.02)	(1.06)	(1.59)	(2.08)	(2.23)	(2.62)	(3.38)
$h(\mathbf{x}_i) = 1 + I(\tau \geq 0.5)(x_{i1} + x_{i2})$									
BFLL	20.4	13.33	10.96	10.13	24.81	42.95	62.39	102.23	183.92
	(0.96)	(0.66)	(0.56)	(0.63)	(1.96)	(2.00)	(2.37)	(3.12)	(5.29)
BFGLL	33.09	24.85	22.96	24.09	47.46	62.17	73.73	101.38	159.02
	(1.25)	(1.00)	(0.98)	(1.18)	(2.66)	(2.64)	(3.00)	(3.79)	(5.34)
BFGLGL	35.72	25.3	22.88	25.57	60.24	71.05	88.3	115.84	169.57
	(1.35)	(1.04)	(0.94)	(1.27)	(3.04)	(3.23)	(3.71)	(4.53)	(5.50)
QR	44.22	31.09	25.9	27.27	57.29	92.13	103.6	121.67	167.51
	(1.64)	(1.16)	(1.02)	(1.40)	(2.68)	(3.28)	(3.90)	(4.85)	(6.36)

and interquantile shrinkage.

As expected, the BFLL estimator performs the best for capturing the local sparsity and constancy of quantile slopes. If there exists local sparsity of quantile slopes, the regularization parameter λ_1^2 in the BFLL estimator increases. A high value of λ_1^2 may result in the overshrinkage of the quantile slope towards zero at the quantile levels with no sparsity. At the upper quantiles, the BFLL overshrinks locally sparse quantile slopes towards zero and it results in higher biases and MISE than the other two proposed methods.

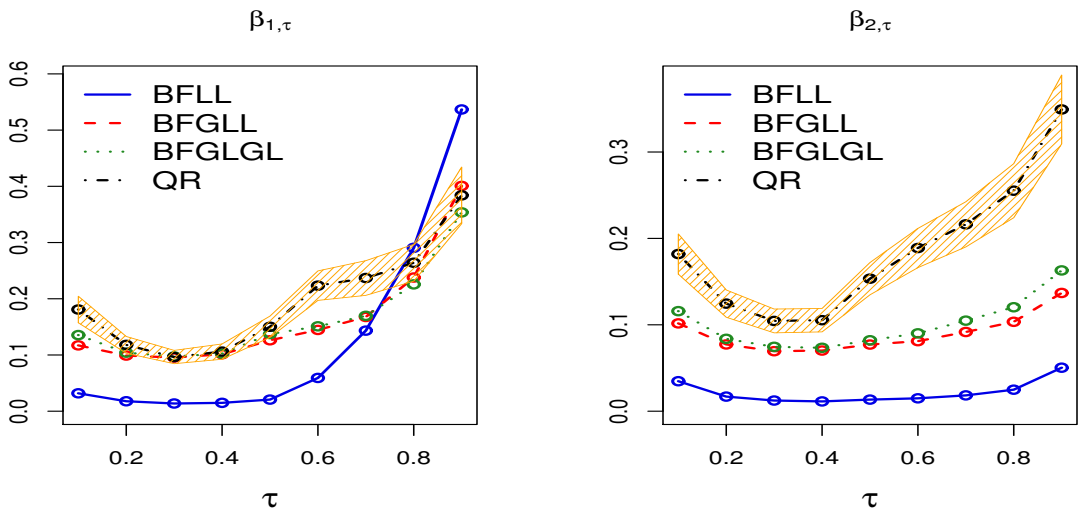


Figure 4.6: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.7) with $\beta_{0,\tau} = \Phi^{-1}(\tau)$, $\beta_{1,\tau} = I(\tau \geq 0.5)\Phi^{-1}(\tau)$ and $\beta_{2,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

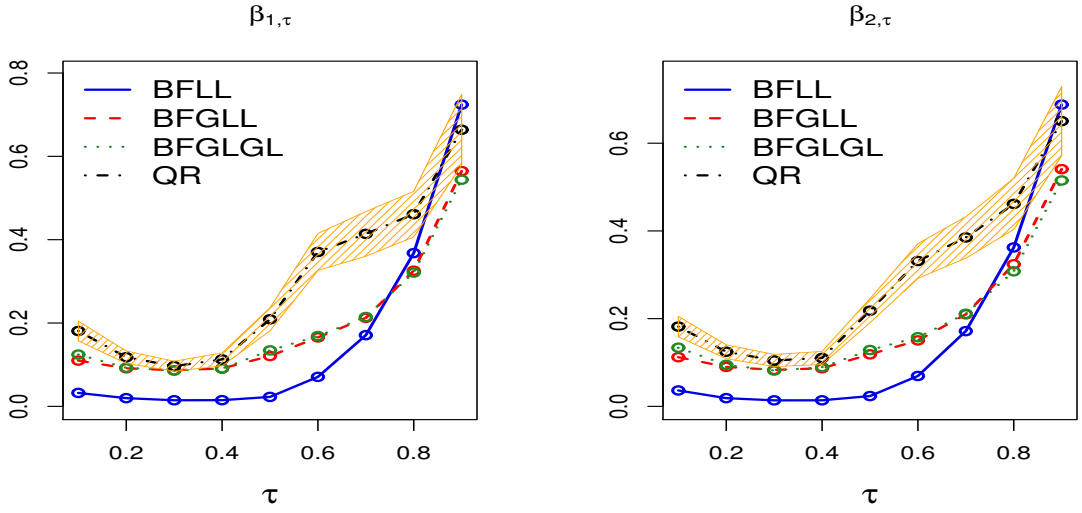


Figure 4.7: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.7) with $\beta_{0,\tau} = \Phi^{-1}(\tau)$ and $\beta_{1,\tau} = \beta_{2,\tau} = I(\tau \geq 0.5)\Phi^{-1}(\tau)$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

4.3.3 Simulation Study III

We consider the following model for generating the simulation data

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + h(x_{i3})e_i, i = 1, \ldots, 200, \quad (4.8)$$

where $\beta_0 = \beta_1 = \beta_2 = 1$, $\beta_3 = 2$, $\beta_4 = \beta_5 = \beta_0 = 0$, $x_{ij} \overset{i.i.d.}{\sim} U(0,1)$, $e_i \overset{i.i.d.}{\sim} N(0,1)$ and $h(x_{i3}) = 1, 1 + I(\tau \geq 0.5)x_{i3}$, or $1 + x_{i3}$. The corresponding conditional quantile function is

$$Q_\tau(Y_i|\mathbf{x}_i) = \beta_{0,\tau} + \beta_{1,\tau} x_{i1} + \beta_{2,\tau} x_{i2} + \beta_{3,\tau} x_{i3} + \beta_{4,\tau} x_{i4} + \beta_{5,\tau} x_{i5} + \beta_{6,\tau} x_{i6},$$

where $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = \beta_{2,\tau} = 1$, $\beta_{4,\tau} = \beta_{5,\tau} = \beta_{6,\tau} = 0$ and

$$\beta_{3,\tau} = \begin{cases} 2 & \text{if } h(x_{i3}) = 1, \\ 2 + I(\tau \geq 0.5)\Phi^{-1}(\tau) & \text{if } h(x_{i3}) = 1 + I(\tau \geq 0.5)x_{i3}, \\ 2 + \Phi^{-1}(\tau) & \text{if } h(x_{i3}) = 1 + x_{i3}. \end{cases}$$

We put independent normal priors $N(\mathbf{0}, 100)$ on $\alpha_{\tau_1}, \ldots, \alpha_{\tau_K}$ and we use $\Gamma(1, 0.1)$ prior on λ_1^2 and λ_2^2. With 100,000 MCMC samples, we discard the first 25,000 samples as burn-in and estimate $\beta_{\tau,j}$ for $j = 0, 1, \ldots, 6$ by the posterior means of the thinned sample using every 100th MCMC sample.

Table 4.4 summarizes the mean of integrated squared errors (MISE) of the QR and three proposed methods. Figures 4.8 – 4.10 plot the mean squared errors (MSEs) of the QR and three proposed methods and 95% pointwise confidence bands of the MSE of the QR estimator.

Figures 4.8 – 4.10 show that the BFGLL estimator has smaller MSE than the BFLL and BFGLGL for estimating the non-zero constant quantile slopes $\beta_{1,\tau}$ and $\beta_{2,\tau}$, especially at tails. The proposed methods have advantage over the QR for estimating $\beta_{1,\tau}$ and $\beta_{2,\tau}$. For estimating $\beta_{3,\tau}$ when $h(x_{i3}) = 1 + I(\tau \geq 0.5)x_{i3}$ or $1 + x_{i3}$, the proposed methods have similar or higher MSE than the QR at the quantile levels with no constancy of quantile slopes. The proposed methods tend to shrink $\beta_{3,\tau}$ towards constants and this leads to higher MSEs than the QR. This explains the slightly higher MISE of the BFLL and BFGLGL at $\tau = 0.1$ and 0.9 when $h(x_{i3}) = 1 + I(\tau \geq 0.5)x_{i3}$ or $1 + x_{i3}$.

All three proposed methods generally give smaller MISE mainly because of their higher efficiency for estimating zero quantile slopes $\beta_{4,\tau}$, $\beta_{5,\tau}$ and $\beta_{6,\tau}$. As we have seen in the simulation study I, the BFGLL estimator has smaller biases than the BFLL and BFGLGL at the upper and lower quantiles and thus the BFGLL estimator has the smallest MISE, especially at tails.

4.4 Summary

Compared to the conventional quantile regression, our proposed methods yield higher efficiency for estimating quantile slope coefficients when they are sparse or constant in some quantile regions. As expected, the proposed methods do not show clear efficiency gain over the QR for estimating non-zero and non-constant quantile slope coefficients.

For estimating zero quantile slopes or non-zero constant quantile slopes, all three proposed methods are more efficient than the QR. One may expect BFGLGL to perform the best for estimating non-zero constant slopes but BFGLGL tends to overshrink towards zero. Instead, BFGLL has smaller MSEs for estimating non-zero constant slopes than BFGLGL. In addition, BFGLL also has less biases at tails and shows the best overall estimation when the model contains non-zero constant quantile slopes.

For estimating quantile slopes with local sparsity, the proposed methods have higher efficiency than QR, possibly due to the compounding effect of joint-modeling of multiple quantiles and regularization. Among the three proposed methods, only BFLL can capture local sparsity of quantile slopes. Consequently, compared to BFGLL, BFGLGL and QR, BFLL shows higher estimation efficiency at quantile levels with local sparsity.

In sum, Figure 4.11 summarizes the features of each proposed method based on our simulation observations. We divide quantile slopes into five different categories (sparse / constant / locally sparse/ locally constant / non-constant). The three proposed methods work well for estimating sparse quantile slopes. The BFGLL method performs the best for estimating non-zero constant quantile slopes and the other two proposed methods work better than QR at tails. The BFLL method is best suited for capturing local sparsity in quantile slopes. The three methods do not show improvement over QR for estimating non-zero locally constant quantile slopes.

Table 4.4: A thousand times the mean of integrated squared errors ($10^3\times$ MISE) for data generated from model (4.8) from the conventional quantile regression method (QR), the Bayesian fused adaptive LASSO (BFLL), the Bayesian fused adaptive LASSO II (BFGLL) and the Bayesian fused adaptive group LASSO (BFGLGL). The values in the parentheses are the standard errors of $10^3\times$ MISE.

τ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$h(x_{i3}) = 1$									
BFLL	105.87	60.55	47.36	39.57	37.46	38.6	43.76	53.65	88.18
	(3.21)	(1.69)	(1.29)	(1.11)	(1.00)	(1.01)	(1.16)	(1.40)	(2.55)
BFGLL	55.98	43.77	38.62	35.44	35.15	35.27	37.62	41.46	54.43
	(1.54)	(1.38)	(1.06)	(0.91)	(0.92)	(0.92)	(1.04)	(1.12)	(1.44)
BFGLGL	108.75	65.07	48.6	40.27	37.08	38.07	44.39	56.83	96.35
	(2.66)	(1.29)	(1.08)	(0.95)	(0.91)	(0.91)	(1.06)	(1.30)	(2.87)
QR	102.2	69.91	61.21	54.98	53.12	55.5	58.75	69.5	103.76
	(2.33)	(1.64)	(1.48)	(1.37)	(1.36)	(1.29)	(1.38)	(1.60)	(2.45)
$h(x_{i3}) = 1 + I(\tau \geq 0.5)x_{i3}$									
BFLL	92.66	56.44	47.89	48.18	68.14	78.83	98.18	131.75	255.59
	(2.62)	(1.57)	(1.33)	(1.41)	(1.97)	(2.21)	(2.71)	(3.39)	(7.48)
BFGLL	62.33	50.62	46.72	47	60.19	63.33	71.29	86.37	134.42
	(1.56)	(1.38)	(1.19)	(1.23)	(1.74)	(1.76)	(1.89)	(2.24)	(3.27)
BFGLGL	100.26	58.46	47.95	49.03	75.29	87.1	107.31	142.58	260.08
	(2.53)	(1.36)	(1.31)	(1.42)	(2.10)	(2.28)	(2.80)	(3.42)	(7.11)
QR	102.12	70.47	62.18	61.73	83.3	106.95	124.18	149.07	220.11
	(2.31)	(1.67)	(1.58)	(1.69)	(2.33)	(2.35)	(3.03)	(3.65)	(5.24)
$h(x_{i3}) = 1 + x_{i3}$									
BFLL	243.64	133.89	102.85	86.42	82.7	88.85	104.81	137.99	260.83
	(7.24)	(3.67)	(2.71)	(2.38)	(2.16)	(2.26)	(2.65)	(3.36)	(7.02)
BFGLL	155.95	114.34	93.17	81.9	79.65	82.75	92.49	110.73	166.19
	(3.84)	(3.61)	(2.66)	(2.21)	(2.21)	(2.30)	(2.52)	(2.87)	(4.01)
BFGLGL	240.22	135.11	98.24	78.83	78.17	86.1	104.48	143.19	283.49
	(5.71)	(2.80)	(2.22)	(1.88)	(1.84)	(2.03)	(2.41)	(3.31)	(7.75)
QR	217.92	149.8	130.34	116.42	113.63	118.81	125.76	149.61	220.39
	(5.09)	(3.74)	(3.10)	(2.90)	(2.87)	(2.83)	(3.10)	(3.62)	(5.29)

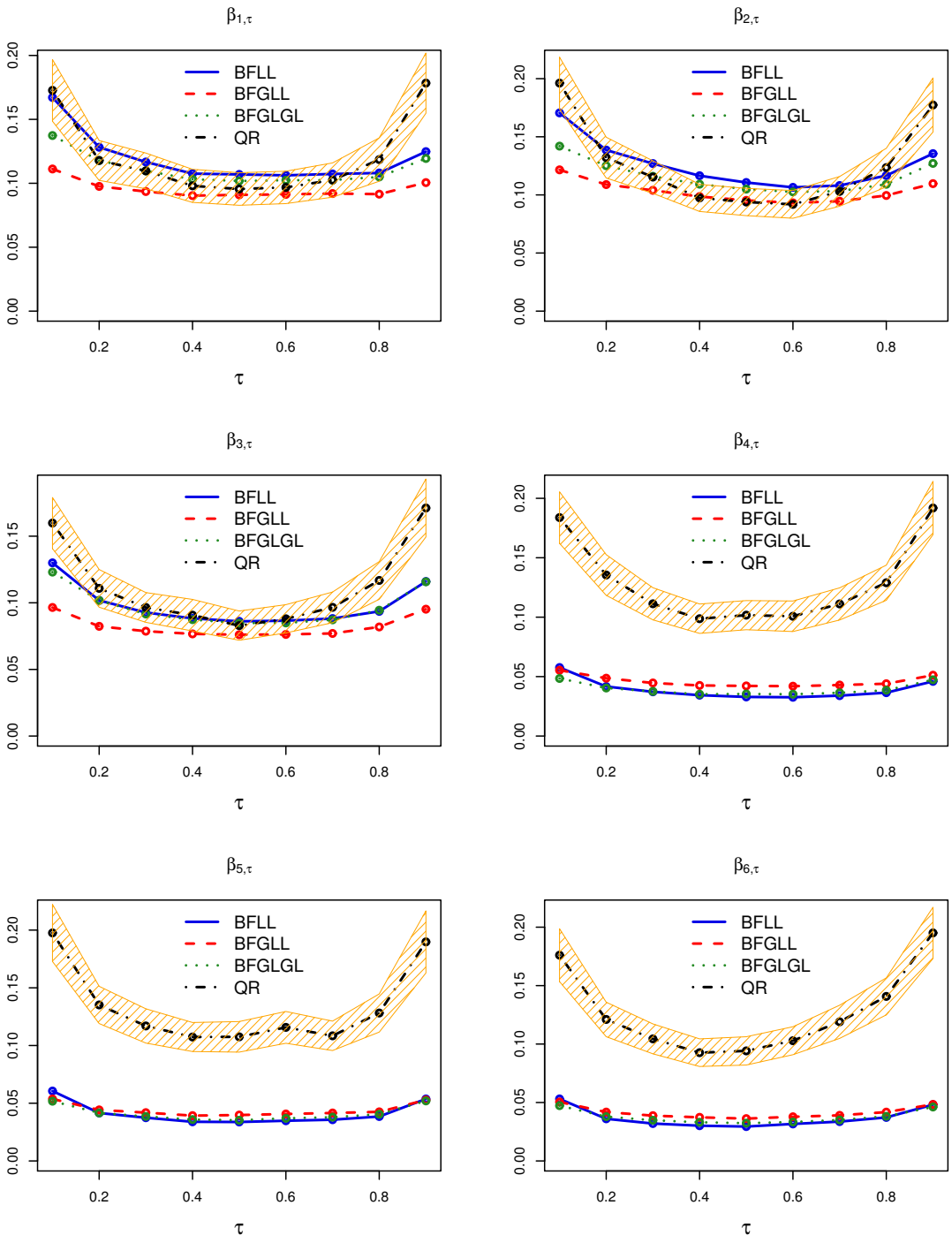


Figure 4.8: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.8) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = \beta_{2,\tau} = 1$, $\beta_{3,\tau} = 2$ and $\beta_{4,\tau} = \beta_{5,\tau} = \beta_{6,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

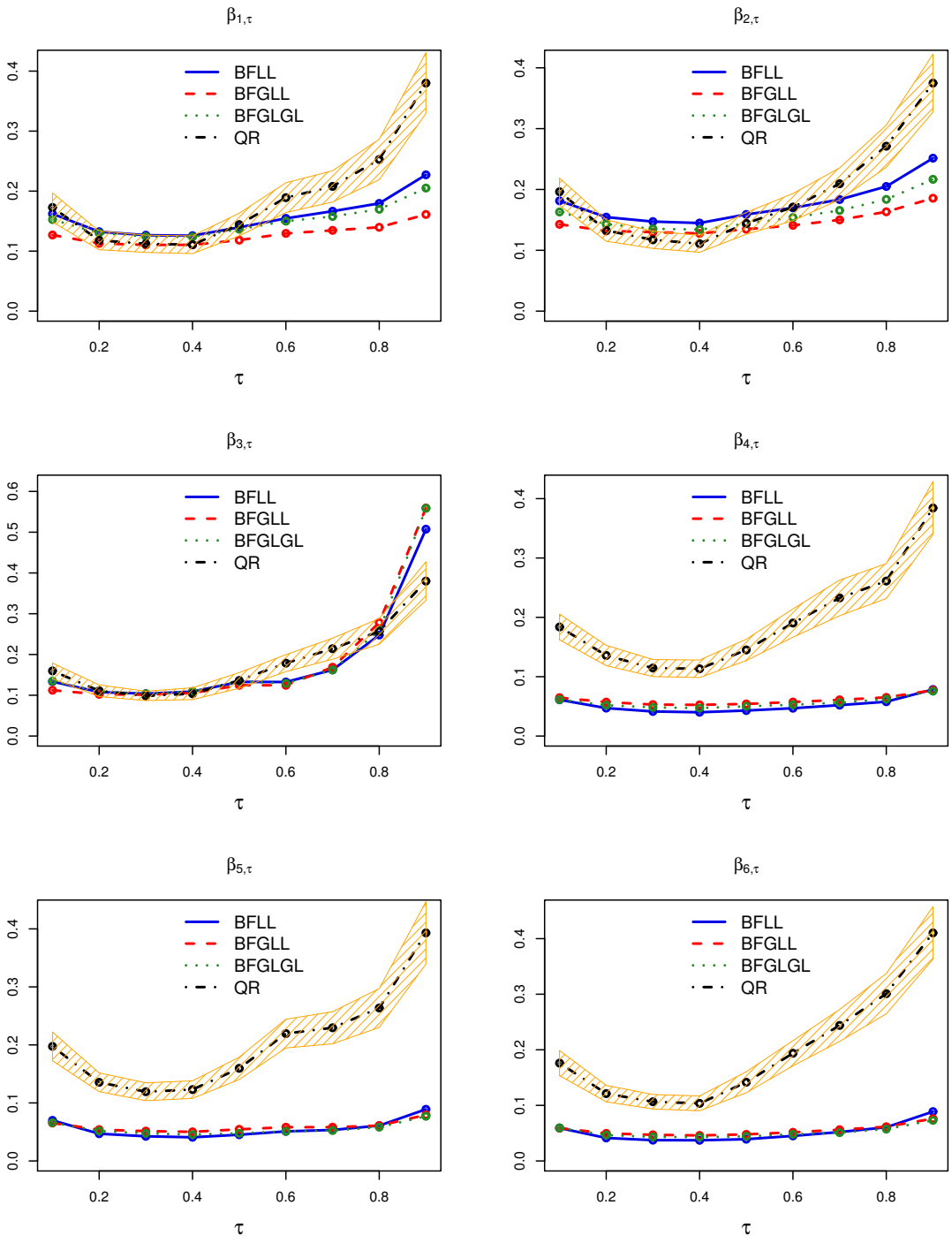


Figure 4.9: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.8) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = \beta_{2,\tau} = 1$, $\beta_{3,\tau} = 2 + I(\tau \geq 0.5)\Phi^{-1}(\tau)$ and $\beta_{4,\tau} = \beta_{5,\tau} = \beta_{6,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.

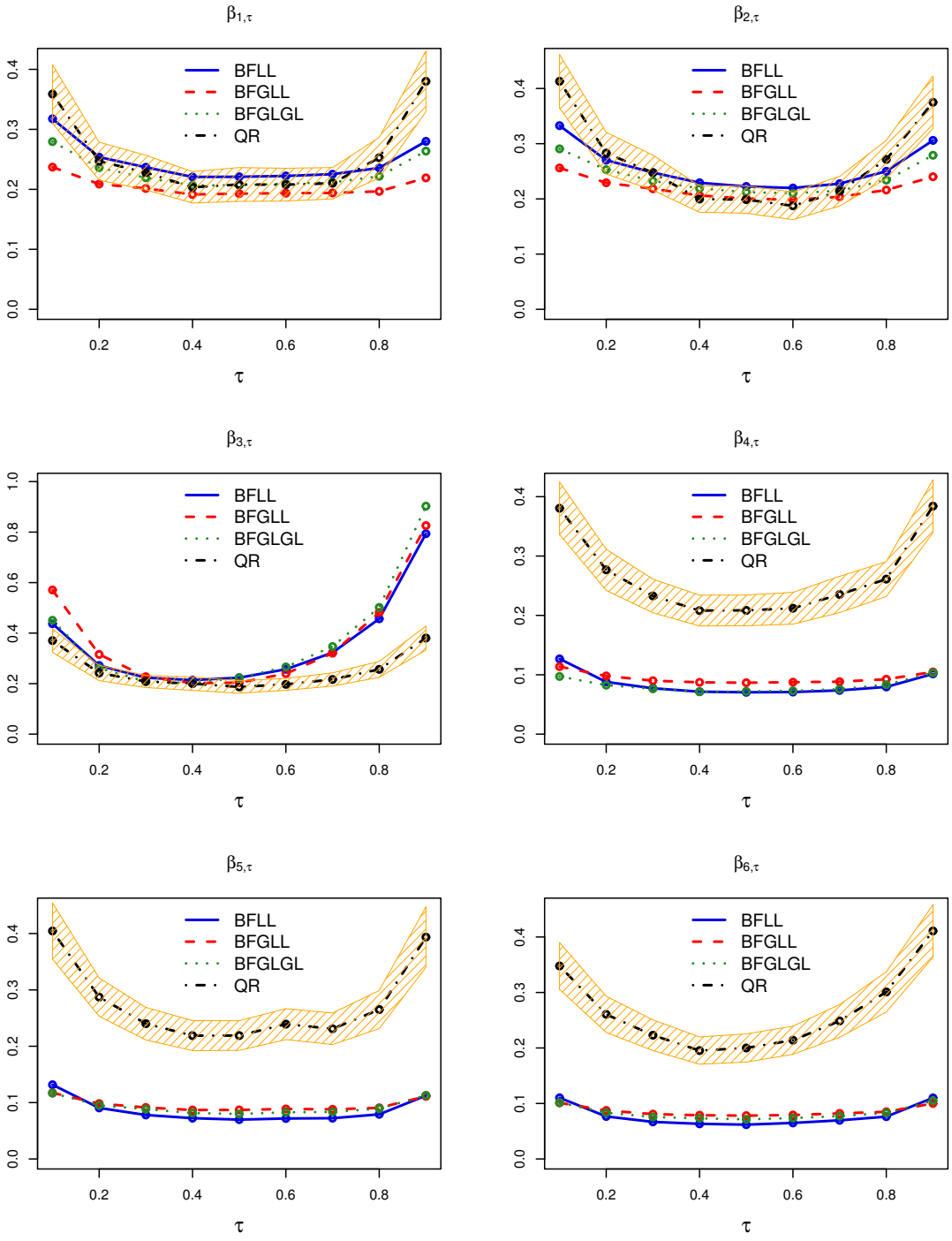
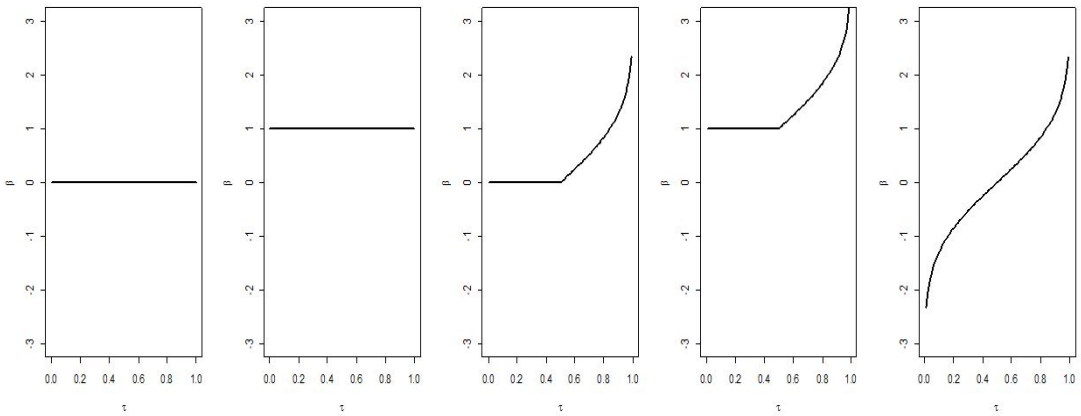


Figure 4.10: Mean squared errors of BFLL, BFGLL, BFGLGL and QR estimators in model (4.8) with $\beta_{0,\tau} = 1 + \Phi^{-1}(\tau)$, $\beta_{1,\tau} = \beta_{2,\tau} = 1$, $\beta_{3,\tau} = 2 + \Phi^{-1}(\tau)$ and $\beta_{4,\tau} = \beta_{5,\tau} = \beta_{6,\tau} = 0$. The shaded bands represent 95% pointwise confidence bands of the MSE of the QR estimator.



BFLL	✓	✓	✓*	△	×
BFGLL	✓	✓*	✓	△	×
BFGLGL	✓	✓	✓	△	×

Figure 4.11: Comparison of BFLL, BFGLL and BFGLGL with QR for five different cases of quantile slopes (sparse / constant / locally sparse/ locally constant / non-constant). The x-axis represents quantile levels and the y-axis represents the true quantile coefficients. A check symbol ✓ means higher efficiency of the proposed method than QR, ✓* indicates the best method among the three penalization methods, Δ means similar efficiency to QR and ✗ indicates slightly lower efficiency than QR.

REFERENCES

Abrevaya, J., and C. M. Dahl. 2008. The Effects of Birth Inputs on Birthweight. *Journal of Business and Economic Statistics* **26**:379–397.

Akaike, H. 1974. A New Look at the Statistical Model Identification. *Automatic Control, IEEE Transactions on* **19**(6):716–723.

Andrews, D. F., and C. L. Mallows. 1974. Scale Mixtures of Normal Distributions. *Journal of the Royal Statistical Society Series B (Methodological)* **36**(1):99–102.

Bakin, S. 1999. *Adaptive Regression and Model Selection in Data Mining Problems*. Ph.D. thesis, Australian National University, Canberra.

Balkema, A. A., and L. de Haan. 1974. Residual Life Time at Great Age. *The Annals of Probability* **2**(5):792–804.

Berger, J. O., B. Liseo, and R. L. Wolpert. 1999. Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **14**(1):1–28.

Bondell, H. D., B. J. Reich, and H. Wang. 2010. Noncrossing Quantile Regression Curve Estimation. *Biometrika* **97**(4):825–838.

Breiman, L. 1995. Better Subset Regression Using the Nonnegative Garrote. *Technometrics* **37**(4).

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag, London.

Davison, A. C., and R. L. Smith. 1990. Models for Exceedances over High Thresholds. *Journal of the Royal Statistical Society Series B (Methodological)* **52**(3):393–442.

Fan, J., and R. Li. 2001. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association* **96**:1348–1360.

Feng, Y. 2011. *Bayesian Quantile Linear Regression*. Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois.

Feng, Y., Y. Chen, and X. He. 2014. Bayesian Quantile Regression with Approximate Likelihood. *Bernoulli* p. to appear.

Fisher, R. A., and L. H. C. Tippett. 1928. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society* **24**:180–190.

Gardiner, J. C., Z. Luo, and L. A. Roman. 2009. Fixed Effects, Random Effects and GEE: What are the Differences? *Statistics in Medicine* **28**(2):221–239.

Geraci, M., and M. Bottai. 2007. Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution. *Biostatistics* **8**(1):140–154.

Geraci, M., and M. Bottai. 2013. Linear Quantile Mixed Models. *Statistics and Computing.* Advance online publication. doi:10.1007/s11222-013-9381-9.

Gilks, W. R., N. G. Best, and K. K. C. Tan. 1995. Adaptive Rejection Metropolis Sampling within Gibbs Sampling. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **44**(4):455–472.

Glen, A. G., L. M. Leemis, and J. H. Drew. 2004. Computing the Distribution of the Product of Two Continuous Random Variables. *Computational Statistics and Data Analysis* **44**(3):451 – 464.

Gnedenko, B. 1943. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *The Annals of Mathematics* **44**(3):423–453.

He, X. 1997. Quantile Curves Without Crossing. *The American Statistician* **51**(2):186–192.

Jiang, L., H. D. Bondell, and H. J. Wang. 2014. Interquantile Shrinkage and Variable Selection in Quantile Regression. *Computational Statistics and Data Analysis* **69**:208–219.

Jiang, L., H. J. Wang, and H. D. Bondell. 2013. Interquantile Shrinkage in Regression Models. *Journal of Computational and Graphical Statistics* **22**(4):970–986.

Jiang, X., J. Jiang, and X. Song. 2012. Oracle Model Selection for Nonlinear Models Based on Weighted Composite Quantile Regression. *Statistica Sinica* **22**:1479–1506.

Jorgensen, B. 1982. *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Number. 9 in Lecture notes in statistics. Springer, New York, NY.

Jung, S.-H. 1996. Quasi-Likelihood for Median Regression Models. *Journal of the American Statistical Association* **91**(433):251–257.

Kim, M., and Y. Yang. 2011. Semiparametric Approach to a Random Effects Quantile Regression Model. *Journal of the American Statistical Association* **106**(496):1405–1417.

Koenker, R. 2004. Quantile Regression for Longitudinal Data. *Journal of Multivariate Analysis* **91**(1):74–89.

Koenker, R. 2005. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, R., and G. Bassett. 1978. Regression Quantiles. *Econometrica* **46**(1):33–50.

Kyung, M., J. Gill, M. Ghosh, and G. Casella. 2010. Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis* **5**(2):369–412.

Laird, N. M., and J. H. Ware. 1982. Random-Effects Models for Longitudinal Data. *Biometrics* **38**(4):963–974.

Li, Q., R. Xi, and N. Lin. 2010. Bayesian Regularized Quantile Regression. *Bayesian Analysis* **5**(3):533–556.

Li, Y., and J. Zhu. 2008. L1-Norm Quantile Regression. *Journal of Computational and Graphical Statistics* **17**(1):163–185.

Liang, K.-Y., and S. L. Zeger. 1986. Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* **73**(1):13–22.

Liya, F., and Y. Wang. 2012. Quantile Regression for Longitudinal Data with a Working Correlation Model. *Computational Statistics and Data Analysis* **56**:2526–2538.

Mu, Y., and Y. Wei. 2009. A Dynamic Quantile Regression Transformation Model for Longitudinal Data. *Statistica Sinica* **19**:1137–1153.

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

Perlman, A. L., X. He, J. Barkmeier, and E. Van Leer. 2005. Bolus Location Associated with Videofluoroscopic and Respirodeglutometric Events. *Journal of Speech, Language, and Hearing Research* **48**:20–33.

Pickands, J. 1975. Statistical Inference using Extreme Order Statistics. *The Annals of Statistics* **3**(1):119–131.

Reich, B. J., H. D. Bondell, and H. Wang. 2010. Flexible Bayesian Quantile Regression for Independent and Clustered Data. *Biostatistics* **11**(2):337–352.

Schwarz, G. 1978. Estimating the Dimension of a Model. *The Annals of Statistics* **6**(2):461–464.

Tang, C. Y., and C. Leng. 2011. Empirical Likelihood and Quantile Regression in Longitudinal Data Analysis. *Biometrika* **98**:1001–1006.

Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**(1):267–288.

Wang, H. 2009. Inference on Quantile Regression for Heteroscedastic Mixed Models. *Statistica Sinica* **19**(3):1247–1261.

Wang, H., G. Li, and G. Jiang. 2007. Robust Regression Shrinkage and Consistent Variable Selection through the LAD-Lasso. *Journal of Business and Economic Statistics* **25**(3):347–355.

Wang, H., J. Zhou, and Y. Li. 2013. Variable Selection for Censored Quantile Regression. *Statistica Sinica* **23**(2013):145–167.

Wei, Y., and X. He. 2006. Conditional Growth Charts (with discussions). *Annals of Statistics* **34**:2069–2097 and 2126–2131.

Wu, Y., and Y. Liu. 2009. Variable Selection in Quantile Regression. *Statistica Sinica* **19**:801–817.

Yang, Y., and X. He. 2012. Bayesian Empirical Likelihood For Quantile Regression. *The Annals of Statistics* **40**(2):1102–1131.

Yu, K., and R. A. Moyeed. 2001. Bayesian Quantile Regression. *Statistics Probability Letters* **54**(4):437–447.

Yu, O., L. Sheppard, T. Lumley, J. Q. Koenig, and G. G. Shapiro. 2000. Effects of Ambient Air Pollution on Symptoms of Asthma in Seattle-area Children Enrolled in the CAMP Study. *Environmental Health Perspectives* **108**(12):1209–14.

Yuan, M., and Y. Lin. 2006. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B* **68**:49–67.

Yuan, Y., and G. S. Yin. 2010. Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data. *Biometrics* **66**(1):105–114.

Zhou, T., and X. He. 2008. Three-step Estimation In Linear Mixed Models with Skew-t Distributions. *Journal of Statistical Planning and Inference* **138**(6):1542–1555.

Zou, H. 2006. The adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101**:1418–1429.

Zou, H., and M. Yuan. 2008. Composite Quantile Regression and the Oracle Model Selection Theory. *The Annals of Statistics* **36**(3):1108–1126.

APPENDIX

Appendix A

A.1 Proposal Distribution for Quantile Monotonicity

A.1.1 For Independent data

We describe the proposal distribution for β_{l,τ_k}, the lth element of $\boldsymbol{\beta}_{\tau_k}$ for the algorithm in Section 2.2.2 to avoid quantile crossing.

First, we calculate lb_i and ub_i, the lower and upper bounds for β^*_{l,τ_k}. We then take the maximum of lb_i and minimum of ub_i over i as the lower and upper bounds for β^*_{l,τ_k}. For $k = 2, \ldots, K-1$, define

$$lb_i = \min \left\{ \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}}, \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}} \right\},$$
$$ub_i = \max \left\{ \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{j,l}}, \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}} \right\},$$

where $x_{i,l}$ is the lth element of $\mathbf{x}_i$. For $k = 1$ and $x_{i,l} > 0$, define

$$lb_i = -\infty \text{ and } ub_i = \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}}.$$

For $k = 1$ and $x_{i,l} < 0$, define

$$lb_i = \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}} \text{ and } ub_i = \infty.$$

For $k = K$ and $x_{i,l} > 0$ define

$$lb_i = \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}} \text{ and } ub_i = \infty.$$

For $k = K$ and $x_{i,l} < 0$ define

$$lb_i = -\infty \text{ and } ub_i = \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{i,m} \beta_{m,\tau_k}}{x_{i,l}}.$$

In addition, for $x_{i,l} = 0$, define $lb_i = -\infty$ and $ub_i = \infty$. Thus, the lower and upper bounds for the candidate β^*_{l,τ_k} are $lb = \max_i(lb_i)$ and $ub = \min_i(ub_i)$, respectively. If k is neither 1 nor K, we draw β^*_{l,τ_k} from $U(lb, ub)$ or a truncated normal distribution with support (lb, ub). If k is either 1 or K, we generate β^*_{l,τ_k} from a normal distribution truncated at lb or ub, respectively.

A.1.2 For Clustered Data

This subsection describes the specifications of the lower and upper bounds for β^*_{l,τ_k} in Step 5 of the MCMC algorithm in Section 3.2.2 to avoid quantile crossing.

First, we calculate lb_{ij} and ub_{ij}, the lower and upper bounds for β^*_{l,τ_k} associated with cluster i for the j-th subject. We then take the maximum of lb_{ij} and minumum of ub_{ij} over i and j as the lower and upper bounds for β^*_{l,τ_k}. For $k = 2, \ldots, K-1$, define

$$lb_{ij} = \min \left\{ \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}}, \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}} \right\},$$
$$ub_{ij} = \max \left\{ \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}}, \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}} \right\},$$

where $x_{ij,l}$ is the lth element of $\mathbf{x}_{ij}$. For $k = 1$ and $x_{ij,l} > 0$, define

$$lb_{ij} = -\infty \text{ and } ub_{ij} = \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}}.$$

For $k = 1$ and $x_{ij,l} < 0$, define

$$lb_{ij} = \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k+1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}} \text{ and } ub_{ij} = \infty.$$

For $k = K$ and $x_{ij,l} > 0$ define

$$lb_{ij} = \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}} \text{ and } ub_{ij} = \infty.$$

For $k = K$ and $x_{ij,l} < 0$ define

$$lb_{ij} = -\infty \text{ and } ub_{ij} = \frac{\mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau_{k-1}} - \sum_{m \neq l} x_{ij,m} \beta_{m,\tau_k}}{x_{ij,l}}.$$

In addition, for $x_{ij,l} = 0$, define $lb_{ij} = -\infty$ and $ub_{ij} = \infty$. Thus, the lower and upper bounds for the candidate β^*_{l,τ_k} are $lb = \max_{i,j}(lb_{ij})$ and $ub = \min_{i,j}(ub_{ij})$, respectively. If k is neither 1 nor K, we draw β^*_{l,τ_k} from $U(lb, ub)$ or a truncated normal distribution with support (lb, ub). If k is either 1 or K, we generate β^*_{l,τ_k} from a normal distribution truncated at lb or ub, respectively.

A.1.3 For Interquantile Differences

This subsection describes the specifications of the lower and upper bounds for $d_{k,l}$, the lth element of $\boldsymbol{\beta}_{\tau_k}$ for the algorithm in Section 4.2.2 to avoid quantile crossing.

To avoid quantile crossing, the conditional quantile functions for the quantile levels $\tau_1, ..., \tau_K$ must satisfy that

$$\alpha_1 + \mathbf{x}_i^T \mathbf{d}_1 < \alpha_2 + \sum_{k=1}^{2} \mathbf{x}_i^T \mathbf{d}_k < \cdots < \alpha_K + \sum_{k=1}^{K} \mathbf{x}_i^T \mathbf{d}_k.$$

Equivalently,

$$\alpha_k < \alpha_{k+1} + \mathbf{x}_i^T \mathbf{d}_{k+1}, \text{ for } k = 1, ..., K-1.$$

We can obtain the upper and lower bounds of the candidates for α_k and d_{kl} using the following procedure. Suppose that we would like to update d_{kl}, $k = 2, ..., K$, $l = 1, ..., p$.

Let d_{kl}^* denote the candidate for d_{kl}. Let lb_i and ub_i denote the lower and upper bounds of d_{kl}^* for the i-th observation. For $x_{il} > 0$, define

$$lb_i = -\frac{\alpha_k - \alpha_{k-1} + \sum_{j \neq l} x_{ij} d_{kj}}{x_{ij}}, \tag{A.1}$$

$$ub_i = \infty. \tag{A.2}$$

For $x_{il} < 0$, define

$$lb_i = -\infty, \tag{A.3}$$

$$ub_i = -\frac{\alpha_k - \alpha_{k-1} + \sum_{j \neq l} x_{ij} d_{kj}}{x_{ij}}. \tag{A.4}$$

For $x_{il} = 0$, define $lb_i = -\infty$ and $ub_i = \infty$. Thus, the lower and upper bounds for the candidate are $lb = \max_{1 \leq i \leq n}(lb_i)$ and $ub = \min_{1 \leq i \leq n}(ub_i)$, respectively. The candidate is drawn from the following truncated normal distribution:

$$f_{TN}(d_{kl}^*; d_{kl}, \varphi, lb, ub) = \frac{\frac{1}{\varphi}\phi\left(\frac{d_{kl}^* - d_{kl}}{\varphi}\right)}{\Phi\left(\frac{ub - d_{kl}}{\varphi}\right) - \Phi\left(\frac{lb - d_{kl}}{\varphi}\right)}, \tag{A.5}$$

where Φ and ϕ are the cumulative distribution and probability density functions of the standard normal distribution, respectively, and φ is a some constant.

Similarly, we can update $d_{1l}, l = 1, ..., p$ and $\alpha_k, k = 1, ..., K$ after finding their lower and upper bounds. Let lb_i and ub_i denote the lower and upper bounds of α_k for the i-th observation and let $lb = \max(lb_i)$ and $ub = \min(ub_i)$. The lower and upper bounds for α_k, $k = 2, ..., K-1$, are $\alpha_{k-1} - \mathbf{x}_i^T \mathbf{d}_k$ and $\alpha_{k+1} + \mathbf{x}_i^T \mathbf{d}_{k+1}$, respectively. We choose the proposal distribution for α_k with the support (lb, ub), for example, $U(lb, ub)$ or $f_{TN}(\cdot; \alpha_k, \varphi, lb, ub)$. We only have the upper bound for α_1, that is, $\alpha_2 + \mathbf{x}_i^T \mathbf{d}_2$ and we have the lower bound for α_K, that is, $\alpha_{K-1} - \mathbf{x}_i^T \mathbf{d}_K$. The proposal distributions for α_1 and α_K are $f_{TN}(\cdot; \alpha_1, \varphi, -\infty, ub)$ and $f_{TN}(\cdot; \alpha_K, \varphi, lb, \infty)$, respectively. There are neither a lower bound nor a upper bound for d_{1l}, $l = 1, ..., p$. The proposal distribution for d_{1l} is the normal distribution with mean d_{1l} and some variance.

A.2 Comparisons of the LID and LIGPD Methods

In this section, we propose different versions of the LID and LIGPD methods and compare new methods with the LID and LIGPD methods. First, we propose two new LIGPD methods: one with updating the scale and shape parameters in tail distributions and one with reparameterization and new constraints on the quantile coefficients. Second, we propose a tail parameter estimation method for the LID method by matching the linearly interpolated density and tail density. Lastly, we compare the performance of different LID and LIGPD estimators via simulation studies.

A.2.1 LIGPD with Updating the Scale and Shape Parameters of Tail Distributions

Our proposed LIGPD method uses the generalized Pareto distribution (GPD) as the tail distributions and the scale and shape parameters of the GPD remain constant in the MCMC algorithm. The scale parameter estimates are determined by the initial estimates of the quantile coefficients and the shape parameter is determined by the maximum likelihood estimation using samples over the lower and upper thresholds. Those thresholds are also determined by the initial estimates of the quantile coefficients.

In constrast, the new LIGPD method updates the scale and shape parameters of the GPD in the MCMC algorithm. The scale parameters are estimated by matching the linearly interpolated density and the GPD tail density at the lower and upper thresholds. We update the shape parameters ξ_1 and ξ_2 by using the Metropolis-Hastings algorithm. The algorithm of the LIGPD method with updating tail density parameters is as follows.

1. Choose the initial values for the quantile coefficients $\boldsymbol{\beta}_K$ using the algorithm in Section 2.2 and obtain the initial values for the scale and shape parameters: ϱ_1, ϱ_2, ξ_1 and ξ_2 by using Step 1-2 in Section 2.3.

2. Update $\boldsymbol{\beta}_K$ by using the Metropolis-Hastings algorithm in Section 2.2.

3. Estimate the scale parameters ϱ_1 and ϱ_2. At the t-th iteration in the MCMC, the scale parameter ϱ_2 for the right tail distributions can be updated by

$$\varrho_2^{(t)} = \{1 - (\tau_{K-1} + \tau_K)/2)\} \sum_{i=1}^{n} \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_K}^{(t)} - \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{K-1}}^{(t)}}{n(\tau_K - \tau_{K-1})},$$

where $\boldsymbol{\beta}_{\tau}^{(t)}$ are the quantile coefficients at the t-th iteration. Similarly, the scale parameter ϱ_1 is updated by

$$\varrho_1^{(t)} = (\tau_1 + \tau_2)/2 \sum_{i=1}^{n} \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_2}^{(t)} - \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_1}^{(t)}}{n(\tau_2 - \tau_1)}.$$

4. Update the shape parameters ξ_1 and ξ_2 using the Metropolis-Hastings algorithm. We assign the priors on ξ_1 and ξ_2, say $\pi(\xi_1)$ and $\pi(\xi_2)$ and the proposal distributions for ξ_1 and ξ_2 are distributions with finite support such as uniform distributions and truncated normal distributions. Suppose ξ_1^* is the candidate at the $(t+1)$-th iteration. Then, Take

$$\xi_1^{(t+1)} = \begin{cases} \xi_1^* & \text{with probability} \quad \phi(\xi_1^{(t)}, \xi_1^*) \\ \xi_1^{(t)} & \text{with probability} \quad 1 - \phi(\xi_1^{(t)}, \xi_1^*), \end{cases}$$

where

$$\phi(\xi_1^{(t)}, \xi_1^*) = \min \left\{ \frac{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K^*, \xi_1^*, \xi_1^{(t)}, \varrho_1^{(t)}, \varrho_2^{(t)}) \pi(\xi_1^*)}{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{x}_i, \boldsymbol{\beta}_K^{(t)}, \xi_1^{(t)}, \xi_2^{(t)}, \varrho_1^{(t)}, \varrho_2^{(t)}) \pi(\xi_1^{(t)})}, 1 \right\},$$

and $\hat{f}_Y(y|\mathbf{x}, \boldsymbol{\beta}_K, \xi_1, \xi_1, \varrho_1, \varrho_2)$ is the approximate density with the left tail density $f_1(y) = g(-y; \varrho_1, \xi_1, -l)$ and the right tail density $f_2(y) = g(y; \varrho_2, \xi_2, u)$ and g is the generalized Pareto density function, $l = \{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_1} + \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_2}\}/2$ and $u = \{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_K} + \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{K-1}}\}/2$ are the lower and upper thresholds, respectively. We can update ξ_2 by using the same algorithm.

5. Repeat Steps 2 and 4 until the MCMC chain converges.

A.2.2 LIGPD with Simpler Constraints to the Quantile Coefficients

The LIGPD and LID methods avoid quantile crossing problem by imposing the following constraints to the quantile coefficient:

$$\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{k-1}} \leq \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_k}, k = 2, ..., K, i = 1, \cdots, n. \tag{A.6}$$

These $(K-1)n$ constraints ensure non-crossing property but also increase the computation time. However, Using the constraint (3) in Bondell et al. (2010), the number of constraints can be reduced from $(K-1)n$ to $(K-1)$ after the transformation of covariates and reparameterization. Hence, the computation can be simplified.

Let x_{ij} be the j-th predictor of observation i and $z_{ij} = \frac{x_{ij}-x_{j,min}}{x_{j,max}-x_{j,min}} \in [0,1]$, where $x_{j,min} = \min(x_{1j}, \cdots, x_{nj})$ and $x_{j,max} = \max(x_{1j}, \cdots, x_{nj})$. The domain of interest reduced to $[0,1]^p$. Then, we can re-write the conditional quantile function in terms of $\mathbf{z}_i = (z_{i1}, \cdots, z_{ip})$ and $\boldsymbol{\alpha}_{\tau_k} = (\alpha_{1k}, \cdots, \alpha_{pk})$.

$$Q_{\tau_k}(Y|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_k} = \mathbf{z}_i^T \boldsymbol{\alpha}_{\tau_k}, \qquad i = 1, \cdots, n, \ k = 1, ..., K.$$

Next, consider the reparametrization by using inter-quantile differences from $\boldsymbol{\alpha}_{\tau_1}, \cdots, \boldsymbol{\alpha}_{\tau_K}$ to $\boldsymbol{\gamma}_{\tau_1}, \cdots, \boldsymbol{\gamma}_{\tau_K}$:

$$\begin{aligned}
&(\gamma_{11}, \cdots, \gamma_{p1}) = (\alpha_{11}, \cdots, \alpha_{p1}), \\
&(\gamma_{1k}, \cdots, \gamma_{pk}) = (\alpha_{1k} - \alpha_{0,k-1}, \cdots, \alpha_{pk} - \alpha_{p,k-1}), \qquad k = 2, \cdots, K.
\end{aligned} \tag{A.7}$$

The constraints can be written in terms of the quantile coefficient $\boldsymbol{\alpha}_{\tau_k}, k = 2, \cdots, K$:

$$\begin{aligned}
& \mathbf{z}_i^T \boldsymbol{\alpha}_{\tau_k} \geq \mathbf{z}_i^T \boldsymbol{\alpha}_{\tau_{k-1}}, \ \forall i \\
\Leftrightarrow \ & \mathbf{z}_i^T \boldsymbol{\gamma}_{\tau_k} \geq 0, \ \forall i \\
\Leftrightarrow \ & \gamma_{1k} + \sum_{j=2}^{p} z_{ij}\gamma_{jk} \geq 0, \ \forall i.
\end{aligned} \tag{A.8}$$

Since $z_{ij}\gamma_{jk} \geq 0 \times (\gamma_{jk})_+ - 1 \times (\gamma_{jk})_-$, the constraint (A.8) will be satisfied if

$$\gamma_{1k} - \sum_{j=2}^{p} (\gamma_{jk})_- \geq 0, \tag{A.9}$$

where $\gamma_{jk} = (\gamma_{jk})_+ - (\gamma_{jk})_-$, $(\gamma_{jk})_+ = I(\gamma_{jk} > 0)\gamma_{jk}$ and $(\gamma_{jk})_- = -I(\gamma_{jk} \leq 0)\gamma_{jk}$. Hence, the number of constraints can be reduced to $(K-1)$ by using the constrinat (A.9).

To incorporate the reparameterization and the constraint (A.8), we propose new Metropolis-Hastings algorithm as follows.

1. Choose the initial values for the inter-quantile differences $\boldsymbol{\gamma}_K^{(0)} = (\boldsymbol{\gamma}_{\tau_1}^{(0)}, \cdots, \boldsymbol{\gamma}_{\tau_K}^{(0)})$.

First, obtain $\boldsymbol{\beta}_k^{(0)}$ using the algorithm in Section 2.2 and obtain the initial values for the scale and shape parameters by using Step 1-2 in Section 2.3. Then, $\boldsymbol{\gamma}_K^{(0)}$ can be calculated by the equation (A.7) and the following equations

$$\alpha_{1k} = \beta_{1k} + \sum_{j=2}^{p} \frac{x_{j,min}\alpha_{jk}}{x_{j,max} - x_{j,min}},$$
$$\alpha_{jk} = (x_{j,max} - x_{j,min})\beta_{jk} \text{ for } j = 2, \cdots, p,$$

where β_{jk} and α_{jk} are the j-th element of $\boldsymbol{\beta}_{\tau_k}$ and $\boldsymbol{\alpha}_{\tau_k}$, respectively.

2. Propose a move at the $(t+1)$-th iteration, $t = 0, 1, 2, \cdots$. Randomly pick a quantile level, τ_k, and select a random number from 1 to p, say l. Let γ_{lk} denote the lth element of $\boldsymbol{\gamma}_{\tau_k}$. The candidate γ_{lk}^* need to satisfy the constraint (A.9).

   For $\gamma_{l1}^*, l = 1, \cdots, p$, we do not have any constraint and we draw γ_{l1}^* from a normal distribution.

   For $\gamma_{lk}^*, l = 2, \cdots, p, k = 2, \cdots, K$, the constraint (A.9) is equivalent to

$$\gamma_{lk}^* \geq -\gamma_{1k}^{(t)} + \sum_{j \neq l} (\gamma_{jk}^{(t)})_-,$$

   where $\gamma_{jk}^{(t)}$ is the MCMC sample at the t-th iteration since

$$\begin{aligned}
&\left\{ (\gamma_{lk}^*)_- \,\middle|\, (\gamma_{lk}^*)_- \in \left[\gamma_{1k}^{(t)} - \sum_{j \neq l} (\gamma_{jk}^{(t)})_-, 0 \right] \right\} \\
\Leftrightarrow &\left\{ (\gamma_{lk}^*)_- \,\middle|\, (\gamma_{lk}^*)_- \in \left[\gamma_{1k}^{(t)} - \sum_{j \neq l} (\gamma_{jk}^{(t)})_-, 0 \right) \right\} \cup \{ (\gamma_{lk}^*)_- | (\gamma_{lk}^*)_- = 0 \} \\
\Leftrightarrow &\left\{ \gamma_{lk}^* \,\middle|\, \gamma_{lk}^* \in \left[-\gamma_{1k}^{(t)} + \sum_{j \neq l} (\gamma_{jk}^{(t)})_-, 0 \right) \right\} \cup \{ \gamma_{lk}^* | \gamma_{lk}^* \geq 0 \} \\
\Leftrightarrow &\left\{ \gamma_{lk}^* \,\middle|\, \gamma_{lk}^* \in \left[-\gamma_{1k}^{(t)} + \sum_{j \neq l} (\gamma_{jk}^{(t)})_-, \infty \right) \right\},
\end{aligned}$$

   and we draw γ_{lk}^* from the truncated normal distribution $N_{[-\gamma_{1k}^{(t)} + \sum_{j \neq l} (\gamma_{jk}^{(t)})_-, \infty)}(\gamma_{lk}^{(t)}, \sigma_{lk}^2)$. Similarly, we draw $\gamma_{lk}^*, l = 2, \cdots, p, k = 1$ from the truncated normal distribution

$N_{[\sum_{j=2}^{p}(\gamma_{jk}^{(t)})-,\infty)}(\gamma_{1k}^{(t)}, \sigma_{l1}^2)$.

The l-th element of $\boldsymbol{\gamma}_{\tau_k}$ will be replaced by γ_{lk}^*, generating the candidate $\boldsymbol{\gamma}_K^*$, which has the same elements as $\boldsymbol{\gamma}_K^{(t)}$ except the l-th element of $\boldsymbol{\gamma}_{\tau_k}^{(t)}$.

3. Take

$$\boldsymbol{\gamma}_K^{(t+1)} = \begin{cases} \boldsymbol{\gamma}_K^* & \text{with probability} \quad \phi(\boldsymbol{\gamma}_K^{(t)}, \boldsymbol{\gamma}_K^*) \\ \boldsymbol{\gamma}_K^{(t)} & \text{with probability} \quad 1 - \phi(\boldsymbol{\gamma}_K^{(t)}, \boldsymbol{\gamma}_K^*), \end{cases}$$

where

$$\phi(\boldsymbol{\gamma}_K^{(t)}, \boldsymbol{\gamma}_K^*) = \min \left\{ \frac{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{z}_i, \boldsymbol{\gamma}_K^*)\pi(\boldsymbol{\gamma}_K^*)}{\prod_{i=1}^n \hat{f}_Y(y_i|\mathbf{z}_i, \boldsymbol{\gamma}_K^{(t)})\pi(\boldsymbol{\gamma}_K^{(t)})}, 1 \right\}$$

and $\pi(\cdot)$ is the prior for $\boldsymbol{\gamma}_K$ and $\hat{f}_Y(y|\mathbf{z}, \boldsymbol{\gamma}_K) = f_Y(y|\mathbf{x}, \boldsymbol{\beta}_K)$ is the approximate density. To obtain $\boldsymbol{\beta}_{\tau_k}, k = 1, \cdots, K$, the following equations can be used,

$$\beta_{1k} = \alpha_{1k} - \sum_{j=2}^{p} \frac{x_{j,min}\alpha_{jk}}{x_{j,max} - x_{j,min}},$$
$$\beta_{jk} = \frac{\alpha_{jk}}{x_{j,max} - x_{j,min}},$$

where $\alpha_{jk} = \sum_{m=1}^{k} \gamma_{jm}$, $j = 1, \cdots, p$.

4. Repeat Steps 2 and 3 until the MCMC chain converges.

A.2.3 Parameter Estimation for Tail Distributions of LID

The LID method uses the truncated normal distributions as tail distributions and the parameters of tail distributions need to be pre-specified. Simulation studies I and II show that bad pre-specified values can result in estimations with high mean squared errors. Thus, we propose a simple method for estimating the parameter of the tail distribution by matching the linearly interpolated density and truncated normal tail density at $\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}$ and $\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_K}$.

The left and right tail distribution of the LID method are the truncated normal distributions $N_{(-\infty,\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1})}(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_1}, \sigma_1^2)$ and $N_{(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_K},\infty)}(\mathbf{x}_i^T\boldsymbol{\beta}_{\tau_K}, \sigma_2^2)$, respectively. Then, the

variance parameter σ_1^2 and σ_2^2 can be estimated by

$$\hat{\sigma}_1 = 2\tau_1 \sum_{i=1}^{n} \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_2} - \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_1}}{n\sqrt{2\pi}(\tau_2 - \tau_1)} \text{ and}$$

$$\hat{\sigma}_2 = 2\{1 - \tau_K) \sum_{i=1}^{n} \frac{\mathbf{x}_i^T \boldsymbol{\beta}_{\tau_K} - \mathbf{x}_i^T \boldsymbol{\beta}_{\tau_{K-1}}}{n\sqrt{2\pi}(\tau_K - \tau_{K-1})}.$$

A.2.4 Simulation Study

To comply with committee members' comments on the tail density approximation method, we investigate the impact of different tail density approximation methods on the estimation efficiency through numerical analysis.

The data were generated from model (2.9) with $x_i \sim LN(0,1)$ and $(\beta_0, \beta_1) = (2, 4)$. We also consider three distributions for generating the random errors: the standard normal, t_3, and χ_1^2.

For each scenario, the simulation is repeated 200 times. In the following, we refer to the LIGPD method with updating the scale and shape parameters of tail distributions as LIGPD II, the LIGPD method with reparamterization and new constraints as LIGPD III, and the proposed LID method with the simple tail distribution parameter estimation method as LID II.

In the simulations, we generated $n = 100, 200$ and 500 observations from model (2.9). For both LID II and LIGPD II methods, we considered a grid of $K = 7, 11$ or 15 equally spaced quantile levels $\tau_k = k/(1+K), k = 1, \cdots, K$. To obtain the posterior distributions of β_{0,τ_k} and $\beta_{1,\tau_k}, k = 1, ..., K$, we put the normal priors $N(\mathbf{0}, \Sigma)$ on $(\beta_{0,\tau_1}, \cdots, \beta_{0,\tau_K})^T$ and $(\beta_{1,\tau_1}, \cdots, \beta_{1,\tau_K})^T$, where $\Sigma = diag(100, ..., 100)$. For LIGPD III method, the priors on $(\gamma_{0,\tau_1}, \cdots, \gamma_{0,\tau_K})^T$ and $(\gamma_{1,\tau_1}, \cdots, \gamma_{1,\tau_K})^T$ are $N(\mathbf{0}, \Sigma_\gamma)$, where Σ_γ is a tridiagonal matrix with large values of diagonal and off-diagonal elements.

Both LID and LIGPD estimates were based on 100,000 samples after the first 50,000 burn-in samples. We estimated the posterior means by thinned samples using every 100th MCMC step.

Tables A.1–A.3 summarize the mean squared errors of the LIGPD II, LIGPD III, and LID II estimators at three quartiles (τ=0.25, 0.5, 0.75), respectively.

In general, the LIGPD estimator has higher efficiency than the LIGPD II. The LIGPD III estimator complicates the algorithm and computation time without show clear ad-

vantage over the LIGPD. Hence, we estimate the shape and scale parameters of the tail densities and then keep them unchanged throughout the MCMC sampling.

A.3 The Scale Mixture Representation of the Laplace Distribution

Bayesian hierarchical models in Chapter 4 rely on a scale mixture representation of the Laplace prior distribution. The Laplace distribution can be written as a scale mixture of normal distributions with an exponential mixing density (Andrews and Mallows, 1974, Example 4.2), that is,

$$\frac{\lambda}{2}\exp\left(-\lambda|\beta|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{\beta^2}{2s}\right)\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}s\right)ds.$$

Proof:

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{\beta^2}{2s}\right)\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}s\right)ds \\
=&\frac{\lambda^2}{2}\int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{1}{2s}(|\beta|-\lambda s)^2\right)\exp\left(-\lambda|\beta|\right)ds \\
=&\frac{\lambda^2}{2}\exp\left(-\lambda|\beta|\right)\int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{1}{2s}(|\beta|-\lambda s)^2\right)ds \\
=&\frac{\lambda}{2}\exp\left(-\lambda|\beta|\right),
\end{aligned}$$

since, by letting $z = (\lambda s)^{-1}$,

$$\begin{aligned}
&\int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{1}{2s}(|\beta|-\lambda s)^2\right)ds \\
=&\int_0^\infty \frac{1}{\lambda z^2}\sqrt{\frac{\lambda z}{2\pi}}\exp\left(-\frac{\lambda\beta^2}{2z}(z-|\beta|^{-1})^2\right)dz \\
=&\lambda^{-1}\int_0^\infty \sqrt{\frac{\lambda}{2\pi z^3}}\exp\left(-\frac{\lambda\beta^2}{2z}(z-|\beta|^{-1})^2\right)dz \\
=&\lambda^{-1}\int_0^\infty f_Z(z)dz = \lambda^{-1},
\end{aligned}$$

Table A.1: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim N(0,1), t_3$ or χ^2_1 from the proposed LIGPD method with updating the scale and shape parameters, where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
					$N(0,1)$				
100	7	0.225	0.130	0.165	0.057	0.224	0.128	0.201	0.080
	11	0.228	0.153	0.196	0.050	0.206	0.140	0.168	0.049
	15	0.321	0.172	0.277	0.091	0.371	0.209	0.261	0.076
200	7	0.090	0.053	0.081	0.039	0.111	0.055	0.115	0.060
	11	0.081	0.051	0.066	0.030	0.082	0.052	0.080	0.029
	15	0.104	0.076	0.088	0.029	0.098	0.067	0.090	0.031
500	7	0.051	0.034	0.053	0.022	0.078	0.034	0.066	0.036
	11	0.043	0.035	0.043	0.016	0.052	0.036	0.046	0.017
	15	0.050	0.038	0.041	0.013	0.054	0.038	0.047	0.015
					t_3				
100	7	0.243	0.145	0.248	0.112	0.300	0.134	0.207	0.079
	11	0.354	0.193	0.346	0.136	0.381	0.178	0.269	0.104
	15	0.554	0.234	0.507	0.236	0.649	0.218	0.407	0.201
200	7	0.122	0.063	0.121	0.055	0.136	0.075	0.126	0.052
	11	0.143	0.081	0.137	0.051	0.142	0.095	0.131	0.037
	15	0.183	0.111	0.186	0.055	0.164	0.102	0.145	0.047
500	7	0.043	0.025	0.041	0.022	0.056	0.035	0.054	0.027
	11	0.042	0.030	0.049	0.021	0.052	0.038	0.050	0.023
	15	0.043	0.030	0.054	0.020	0.045	0.038	0.055	0.018
					χ^2_1				
100	7	0.005	0.036	0.383	0.271	0.005	0.025	0.261	0.195
	11	0.006	0.114	1.133	0.613	0.004	0.093	0.752	0.382
	15	0.015	0.267	2.361	1.155	0.012	0.265	2.135	1.004
200	7	0.004	0.018	0.117	0.096	0.007	0.022	0.136	0.095
	11	0.002	0.025	0.213	0.136	0.002	0.018	0.178	0.116
	15	0.004	0.053	0.457	0.247	0.002	0.041	0.344	0.186
500	7	0.003	0.018	0.095	0.058	0.008	0.038	0.165	0.071
	11	0.002	0.013	0.081	0.048	0.003	0.013	0.084	0.051
	15	0.001	0.014	0.092	0.056	0.001	0.008	0.070	0.050

Table A.2: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim N(0,1), t_3$ or χ^2_1 from the proposed LIGPD method with simpler constraint, where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
					$N(0,1)$				
100	7	0.195	0.127	0.134	0.049	0.180	0.134	0.144	0.061
	11	0.232	0.135	0.147	0.053	0.210	0.152	0.175	0.061
	15	0.323	0.181	0.197	0.072	0.272	0.187	0.247	0.081
200	7	0.096	0.061	0.087	0.040	0.079	0.054	0.070	0.034
	11	0.091	0.058	0.073	0.035	0.076	0.051	0.073	0.033
	15	0.097	0.068	0.084	0.033	0.091	0.062	0.088	0.034
500	7	0.042	0.034	0.053	0.022	0.042	0.031	0.043	0.018
	11	0.046	0.036	0.043	0.019	0.045	0.035	0.043	0.017
	15	0.050	0.035	0.039	0.015	0.051	0.036	0.043	0.014
					t_3				
100	7	0.358	0.180	0.259	0.106	0.248	0.163	0.254	0.080
	11	0.382	0.185	0.297	0.146	0.288	0.193	0.372	0.135
	15	0.427	0.215	0.475	0.216	0.349	0.228	0.604	0.265
200	7	0.144	0.069	0.128	0.064	0.124	0.076	0.118	0.042
	11	0.184	0.083	0.134	0.061	0.145	0.088	0.132	0.046
	15	0.211	0.109	0.175	0.062	0.178	0.112	0.165	0.059
500	7	0.057	0.030	0.048	0.028	0.053	0.035	0.052	0.026
	11	0.051	0.029	0.053	0.026	0.053	0.035	0.048	0.024
	15	0.057	0.033	0.056	0.022	0.058	0.038	0.054	0.023
					χ^2_1				
100	7	0.018	0.110	0.578	0.279	0.007	0.093	0.457	0.216
	11	0.028	0.220	1.379	0.638	0.014	0.189	1.278	0.638
	15	0.033	0.331	2.242	0.986	0.027	0.383	3.524	1.862
200	7	0.008	0.032	0.183	0.124	0.004	0.025	0.133	0.083
	11	0.008	0.051	0.308	0.164	0.004	0.037	0.233	0.129
	15	0.011	0.084	0.523	0.252	0.006	0.064	0.393	0.192
500	7	0.007	0.025	0.098	0.063	0.005	0.027	0.136	0.065
	11	0.004	0.017	0.100	0.062	0.002	0.013	0.086	0.053
	15	0.002	0.021	0.121	0.065	0.001	0.011	0.078	0.052

Table A.3: Mean squared errors of parameter estimators for data generated from model (2.9) with $\epsilon_i \sim N(0,1), t_3$ or χ^2_1 from the proposed LID method with matching linearly interpolated density and truncated normal density, where δ_0=$\beta_{0,.75} - \beta_{0,.5}$ and δ_1=$\beta_{1,.75} - \beta_{1,.5}$.

n	K	$\beta_{0,.25}$	$\beta_{0,.5}$	$\beta_{0,.75}$	δ_0	$\beta_{1,.25}$	$\beta_{1,.5}$	$\beta_{1,.75}$	δ_1
					$N(0,1)$				
100	7	0.180	0.138	0.134	0.039	0.207	0.155	0.183	0.052
	11	0.221	0.147	0.186	0.046	0.230	0.160	0.179	0.048
	15	0.281	0.173	0.220	0.060	0.276	0.187	0.220	0.061
200	7	0.090	0.060	0.083	0.038	0.082	0.055	0.076	0.039
	11	0.094	0.061	0.075	0.029	0.102	0.059	0.083	0.036
	15	0.090	0.062	0.077	0.030	0.083	0.058	0.081	0.032
500	7	0.046	0.035	0.050	0.021	0.050	0.035	0.048	0.020
	11	0.041	0.034	0.045	0.016	0.047	0.034	0.049	0.018
	15	0.046	0.039	0.047	0.015	0.047	0.041	0.057	0.017
					t_3				
100	7	0.315	0.195	0.284	0.103	0.329	0.200	0.269	0.072
	11	0.429	0.257	0.431	0.143	0.478	0.259	0.375	0.098
	15	0.563	0.254	0.503	0.204	0.578	0.255	0.460	0.181
200	7	0.132	0.078	0.127	0.056	0.153	0.102	0.158	0.054
	11	0.159	0.095	0.154	0.053	0.176	0.109	0.147	0.041
	15	0.174	0.108	0.175	0.051	0.184	0.118	0.165	0.051
500	7	0.058	0.032	0.050	0.024	0.064	0.040	0.061	0.029
	11	0.061	0.034	0.057	0.024	0.059	0.042	0.062	0.026
	15	0.053	0.032	0.052	0.019	0.051	0.040	0.059	0.020
					χ^2_1				
100	7	0.013	0.075	0.531	0.292	0.007	0.073	0.402	0.205
	11	0.020	0.181	1.152	0.511	0.012	0.143	0.815	0.353
	15	0.032	0.323	2.287	1.038	0.021	0.263	1.583	0.673
200	7	0.007	0.029	0.167	0.117	0.004	0.023	0.139	0.091
	11	0.007	0.047	0.277	0.149	0.003	0.031	0.220	0.120
	15	0.009	0.073	0.484	0.237	0.006	0.055	0.358	0.179
500	7	0.007	0.029	0.118	0.064	0.006	0.026	0.126	0.064
	11	0.003	0.015	0.086	0.053	0.002	0.013	0.087	0.053
	15	0.003	0.020	0.119	0.066	0.001	0.011	0.082	0.055

where $f_Z(z)$ is the inverse Gaussian density function with mean $|\beta|^{-1}$ and shape parameter λ.

A.4 The Scale Mixture representation of the Spherical Multivariate Laplace Distribution

The K-dimensional spherical multivariate Laplace distribution can be written as a scale mixture of multivariate normal distributions with a gamma mixing density, that is,

$$\exp(-\lambda||\boldsymbol{\beta}_{K,l}||_2) \propto \int_0^\infty N(\mathbf{0}, s^2 I) \times \Gamma\left(\frac{K+1}{2}, \frac{2}{\lambda^2}\right) ds^2.$$

Proof:

$$\begin{aligned}
&\int_0^\infty \left(\frac{1}{2\pi s^2}\right)^{K/2} \exp\left(-\frac{||\boldsymbol{\beta}_K||_2^2}{2s^2}\right) \frac{(s^2)^{\frac{K+1}{2}-1}}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda^2}\right)^{\frac{K+1}{2}}} \exp\left(-\frac{\lambda^2 s^2}{2}\right) ds^2 \\
=&\left(\frac{1}{2\pi}\right)^{K/2} \frac{1}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda^2}\right)^{\frac{K+1}{2}}} \int_0^\infty (s^2)^{\frac{1}{2}-1} \exp\left(-\frac{||\boldsymbol{\beta}_K||_2^2}{2s^2} - \frac{\lambda^2 s^2}{2}\right) ds^2 \\
=&\left(\frac{1}{2\pi}\right)^{K/2} \frac{2}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda^2}\right)^{\frac{K+1}{2}}} \left(\frac{||\boldsymbol{\beta}_K||_2^2}{\lambda^2}\right)^{1/4} \mathbf{K}_{1/2}\left(\sqrt{||\boldsymbol{\beta}_K||_2^2 \lambda^2)}\right) \\
=&\left(\frac{1}{2\pi}\right)^{K/2} \frac{2}{\Gamma\left(\frac{K+1}{2}\right)\left(\frac{2}{\lambda^2}\right)^{\frac{K+1}{2}}} \left(\frac{||\boldsymbol{\beta}_K||_2^2}{\lambda^2}\right)^{1/4} \sqrt{\frac{\pi}{2}} \left(||\boldsymbol{\beta}_K||_2^2 \lambda^2\right)^{-1/4} \exp(-\lambda||\boldsymbol{\beta}_K||_2) \\
=&C_K \exp(-\lambda||\boldsymbol{\beta}_K||_2),
\end{aligned}$$

where $\mathbf{K}_{1/2}(z) = \sqrt{\pi/(2z)}\exp(-z)$ is the modified bessel function of the third kind.

A.5 Inverse Gaussian Distribution

The probability density function of the inverse Gaussian distribution (IG) with mean $\mu \in \mathbb{R}^+$ and shape parameter $\lambda \in \mathbb{R}^+$ is

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right\} I(x>0).$$

A.6 Generalized Inverse Gaussian Distribution

The probability density function of the generalized inverse Gaussian distribution (GIG) with three parameters $a, b \in \mathbb{R}^+$ and $p \in \mathbb{R}$ is

$$f(x;a,b,p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{(p-1)} e^{-(ax+b/x)/2} I(x>0),$$

where K_p is a modified Bessel function of the second kind (Jorgensen, 1982). The GIG($a = \lambda/\mu^2, b = \lambda, p = -1/2$) is the inverse Gaussian with mean μ and shape parameter λ.