Asymmetric Heavy-Tailed Robust Loss Function for Regression, Regularisation and Fast Computation

A thesis presented for the degree of
Doctor of Philosophy

by

Sanna Soomro

Department of Mathematics
College of Engineering, Design and Physical Sciences
Brunel University London
May 2024

Abstract

Robust inference for outliers to statistics and adversarial samples in deep learning is not only 'a new tune of an old song', yet also the current hot research topic in both statistics and artificial intelligence. Asymmetric distributions, including heavy-tailed distributions in regression models and data analysis have been challenging and are required to take into account the robustness of methods. Bayesian inference or Bayesian analysis, as one of the most popular statistics methods, has been widely used in all research fields, including science, social science and engineering recently. Many regression models face challenging issues in high-dimensional and computational problems, which has attracted substantial research in the literature recent years. Therefore, this thesis aims to develop novel Bayesian methods in parametric statistical inference to address these issues via three attempts. The first attempt is to employ Bayesian variable selection with quantile-dependent prior for the fractional polynomial (FP) model, with a medical application in the analysis of blood pressure (BP) amongst United States adults. Whilst the FPs act as a concise and accurate formula for examining smooth relationships between BP measures and risk factors of cardiovascular disease, conditional quantile functions with FPs provide comprehensive relationships, including median and extremely high BP measures. The second attempt is to propose a new asymmetric Huberised loss function taking account of robustness, asymmetry and heavy tails. This motivates the development of robust Bayesian regularisation for a high-dimensional setting. The former has its corresponding probability distribution with the normal scale-mixture property. This leads to a by-product of the research, that is, a new Bayesian Huberised regularised quantile regression, which is derived by adopting the Markov chain Monte Carlo (MCMC) method. Finally, the third attempt is to revisit the work of the previous attempt addressing the computational issue. The MCMC method is a popular technique for full Bayesian probabilistic models, however, it faces the high computational cost when the amount of data increases. Alternative to the MCMC method, variational inference is the approximate-based technique to tackle the computational issue, and is utilised to propose variational Bayesian Huberised Lasso quantile regression and variational Bayesian Huberised adaptive Lasso quantile regression for high-dimensional and computational problems.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisors, Prof. Keming Yu and Dr. Dalia Chakrabarty. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. With whole heartily thanks to Prof. Keming Yu for his exceptional guidance, insightful comments, countless discussions and for being a source of unlimited encouragement in pursuit of achieving my full potential, throughout the four years of research.

I must thank Dr. Diana Roman, Dr. Ben Parker, Dr. Xiaochuan Yang, Dr. Man Tang, Dr. Hongying Meng, Dr. Matthias Maischak and Dr. Veronica Vinciotti who served as monitoring panellists of my progression review during my research study period. Throughout different phases of my research they evaluated my progress and provided me with extremely valuable feedback to evolve as researcher.

Special thanks to the staff and students of the Department of Mathematics, with whom I have had many useful discussions. I particularly acknowledge my office mates, for their feedback sessions and moral support. Thanks should also go to the doctoral researchers from various disciplines who impacted and inspired me. I also had the pleasure of collaborating with Prof. Yan Yu from University of Cincinnati for one of my thesis chapters.

Many thanks to the Engineering and Physical Science Research Council (EPSRC) for the funding I have received (DTP PhD studentship grant no. 2295266). Without this funding I would not have been able to undertake this research.

I am also grateful to my sister for giving her time to proofread my thesis and her useful suggestions.

Finally, I would be remiss to not mentioning all of my friends and family, especially my parents and siblings. Their unwavering belief and encouragement have kept my spirits and motivation high during this process. I would also like to thank my cat for all the entertainment and emotional support.

Author's Publications

1. **Soomro, S.** and Yu, K. (2024). Bayesian Fractional Polynomial Approach to Quantile Regression and Variable Selection with Application in the Analysis of Blood Pressure amongst US Adults. *Accepted by Journal of Applied Statistics*

2. **Soomro, S.**, Yu, K. and Yu, Y. (2024). A New Bayesian Huberised Regularisation and Beyond. *Submission for a journal for consideration of publication*

3. **Soomro, S.** and Yu, K. (2024). Variational Bayesian Huberised Adaptive Lasso. *Working paper*

Contents

Acronyms

AL Average length

ALD Asymmetric Laplace distribution

BP Blood pressure

BQR-EN Bayesian Elastic Net quantile regression

BQR-BL Bayesian Lasso quantile regression

BQR-FP Bayesian quantile regression with fractional polynomials

BQRVS-FP Bayesian quantile regression with fractional polynomials and variable selection

CAVI Coordinate ascent variational inference

CP Coverage probability

CVD Cardiovascular disease

DBP Diastolic blood pressure

EB Empirical Bayesian

EM Expectation-maximisation

FP Fractional polynomial

GIG Generalised inverse Gaussian

HBL Bayesian Huberised Lasso

HBQR-EN Bayesian Huberised Elastic Net quantile regression

HBQR-BL Bayesian Huberised Lasso quantile regression

KL Kulback-Leibner

LAD Least absolute deviation

MAP Maximum a posterior

MAPE Mean absolute prediction error

MCMC Markov chain Monte Carlo

MedSPE Median of squared prediction error

MHPE Mean Huber prediction error

MIP Marginal inclusion probability

MMAD Median of mean absolute deviation

MSPE Mean squared prediction error

NHANES National Health and Nutrition Examination Survey

QR-FP Quantile regression model with fractional polynomials

RMSE Root of mean squared error

SBP Systolic blood pressure

US United States

VB Variational Bayesian

VBAL Variational Bayesian Huberised adaptive Lasso quantile regression

VBL Variational Bayesian Huberised Lasso quantile regression

VI Variational inference

Chapter 1

Background and Motivation

Robust statistics are amongst the most fundamental parts in parametric and non-parametric estimations. They address the problem of making estimates that are insensitive to small changes in the basic assumptions of the statistical models employed, and they take into account that parametric models are at best only approximations to reality. The term robust defines the strength of a model, tests and methodologies according to the user's requirements, whilst reducing the influence of small amount of unusual observations.

The earliest robust-statistics-based inference was studied by Peter Jost Huber in 1964, provided in the paper of Huber (1964) amongst some seminal papers, including Tukey (1960), and Hampel (1968). They laid the foundations of modern robust statistics. For over half century, a large and rich literature on robust statistics has been developed on theories and applications, and it is difficult to summarise in the length of a thesis. However, there are book-length expositions, which provided comprehensive information about robust inference, theories, computational methods and applications. They can be found in Huber (1981) (and its second edition, Huber and Ronchetti (2009)), Hampel et al. (1986), and Maronna et al. (2006). Besides, the recent comprehensive review of robust statistics, provided in Gelman and Vehtari (2021), covered several works in statistics, econometrics, psychometrics, epidemiology, and computer science for causal inference, Bayesian inference, big data, regularisation, and machine learning, amongst others. The field is currently ongoing, and has penetrated mainstream statistics.

Title of this thesis employed the term 'Robust', which focused on developing models that are insensitive to outliers in data, whilst considering the potential impact of asymmetry and heavy tails in data. The title also employed the terms 'Regression', 'Regularisation' and 'Fast Computation' to explore several research studies under a framework of regression

analysis, regularisation and fast computational methods. Particularly, there are research gaps in the framework of quantile regression analysis, thus, the term 'Asymmetric Heavy-Tailed Loss Function'. The research scopes of this thesis are given in details in this chapter. First of all, to explain the concept behind the term 'Loss Function' of the title of this thesis, we begin with a brief overview of some basic concepts of robust statistics, and how they motivate the use of regression analysis.

1.1 Loss Functions

A loss function, also known as a cost function or an error function, is the fundamental part of statistics and machine learning. At its core, a loss function is a simple method of evaluating how well an algorithm models a dataset. In most optimisation problems, one loss function or a combination of different loss functions is treated as an objective function that aims to be minimised. In statistics, a loss function is typically used for a parameter estimation, which is the difference between the estimated value and true value. There are several types of loss functions that are suitable for different applications. We list some loss functions that are closely related to the novel loss function to be presented later in this thesis.

1.1.1 Huber Loss

The Huber loss function is commonly used for robust statistics and defined as

$$L_{\delta}^{\text{Huber}}(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq \delta \\ \delta\left(|x| - \frac{\delta}{2}\right) & |x| > \delta, \end{cases} \tag{1.1}$$

where $\delta > 0$ is a robustness parameter and practically set as $\delta = 1.345$ (Huber (1964)). The behaviour of this loss function is quadratic for small values of x and becomes linear when x exceeds δ in magnitude. Clearly, the Huber loss function is not twice-differentiable that makes it non-smooth. Although Huber (1964) developed the M-estimation method that requires a first derivative only, other optimisation algorithms requiring a second derivative are infeasible for the Huber loss function. For Bayesian inference, Li et al. (2020) stated in Theorem 1 of their paper that the Huber likelihood function does not meet necessary conditions of the normal scale-mixture property. In other words, derivatives of the density function are not continuous in some degree, including the second derivative. As a result, it cannot be expressed as a scale mixture of normal distributions. Thus, the Huber loss

function has the limited scope in applications.

1.1.2 Generalised Huber Losses

Li et al. (2020) proposed two generalised Huber loss functions, which are Soft Huber and Non-convex Huber loss functions. They are attractive alternatives to the Huber loss function because they are analogous to the pseudo Huber loss function (Charbonnier et al. (1997)) and have the normal scale-mixture property resulting in a broader range of frequentist and Bayesian applications. The Soft Huber loss function can be defined as

$$L^{\mathrm{SH}}_{\zeta_1,\zeta_2}(x) = \sqrt{\zeta_1\zeta_2}\left(\sqrt{1+\frac{x^2}{\zeta_2}}-1\right), \tag{1.2}$$

and the Non-convex Huber loss function as

$$L^{\mathrm{NH}}_{\zeta_1,\zeta_2}(x) = \sqrt{\zeta_1\zeta_2}\left(\sqrt{1+\frac{|x|}{\zeta_2}}-1\right), \tag{1.3}$$

where $\zeta_1, \zeta_2 > 0$ are non-negative hyper-parameters. Here, the Soft Huber loss bridges the ℓ_1 (absolute) loss and the ℓ_2 (squared) loss. On the other hand, the Non-convex Huber loss bridges the $\ell_{1/2}$ loss and the ℓ_1 loss. By letting $\eta = \sqrt{\zeta_1\zeta_2}$ and $\rho^2 = \sqrt{\zeta_2/\zeta_1}$, the Soft Huber loss function becomes the hyperbolic loss function, that is,

$$L^{\mathrm{Hyp}}_{\eta,\rho^2}(x) = \sqrt{\eta\left(\eta+\frac{x^2}{\rho^2}\right)}-\eta, \tag{1.4}$$

where $\eta > 0$ is a robustness parameter and $\rho^2 > 0$ is a scale parameter. Park and Casella (2008) used this hyperbolic loss function to formulate the Bayesian Huberised Lasso, which is different from that of Kawakami and Hashimoto (2023) even though they share the same name. In case of a conditional prior, the former used σ of a model as the scale parameter, whilst the latter used ρ^2 of the hyperbolic loss function as the scale parameter, which has proven to be more robust to outliers than that of the former.

1.1.3 Quantile Loss

Asymmetry is an important feature in modelling and has useful properties over symmetry. It can provide information on the entire distribution of a dataset, specifically when it is skewed. The quantile loss function, introduced by Koenker and Bassett (1978), is defined

as

$$L_\tau^{\text{Quantile}}(x) = x(\tau - \text{I}(x < 0)), \tag{1.5}$$

where $\tau \in (0, 1)$ is the quantile level and $\text{I}(x < 0)$ is the indicator function representing the value 1 if x belongs to the set $(-\infty, 0)$, and the value 0 otherwise. Setting $\tau = 0.5$, it results in the absolute loss function as a special case. The quantile loss function is commonly adopted as a minimisation problem for quantile regression problems.

1.2 Quantile Regression

Regression analysis is a technique that quantifies the relationship between a response variable and predictors. Quantile regression, introduced by Koenker and Bassett (1978), is a method to estimate the quantiles of a conditional distribution of a response variable and as such, it permits a more complete portrayal of the relationship between the response variable and predictors.

Given a dataset, $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ and fixed τ, the τ^{th} quantile regression model is represented as

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta}(\tau) + \epsilon(\tau)_i, \quad i = 1, \ldots, n, \tag{1.6}$$

where y_i is the response variable, $\boldsymbol{x}_i$ is the vector of predictors, $\boldsymbol{\beta}(\tau)$ is the vector of unknown parameters of interest, and $\epsilon(\tau)$ is the model error term for the τ^{th} quantile. For the sake of notation simplification, we omit τ from these parameters.

We wish to estimate the unknown parameters, $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}$ for each τ^{th} quantile, which can be done by minimising the quantile loss function over $\boldsymbol{\beta}$:

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} L_\tau^{\text{Quantile}}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}), \tag{1.7}$$

where $L_\tau^{\text{Quantile}}(\cdot)$ is the quantile loss function that is defined in Equation (1.5).

Minimising Equation (1.7) is the same as maximising a likelihood function. An asymmetric Laplace distribution (ALD) is employed, which is the common choice for the quantile regression analysis (Yu and Moyeed (2001), and Yu et al. (2003)). We assume that $\epsilon_i \sim \text{AL}(0, \sigma, \tau)$, $i = 1, \ldots, n$, where the $\text{AL}(\cdot)$ is the ALD with its density

$$f^{\text{AL}}(\epsilon_i) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{L_\tau^{\text{Quantile}}(\epsilon_i)}{\sigma}\right\}. \tag{1.8}$$

Before presenting the scale mixture of normal representation of the ALD, we briefly describe the general concept of scale mixture of normals. Suppose that a random variable X has a probability density function $f(x|\Theta)$ and unknown parameter Θ that satisfies

$$f(x|\Theta) = \int \phi(x|\mu, \sigma)\, \pi(\sigma|\Theta)\, d\sigma, \tag{1.9}$$

where $\phi(\cdot)$ is the mixing distribution and $\pi(\cdot)$ is some density function that is defined on $(0, \infty)$, then X or its $f(x|\Theta)$ is a scale mixture of normal distribution. It has many applications in statistics, finance and particularly in Bayesian inference. Probability distribution with a scale mixture of normal expression could be grouped into two categories: symmetric probability distributions (Andrews and Mallows (1974), and West (1987)) and asymmetric probability distributions (Reed and Yu (2009), da Silva Ferreira et al. (2011), and Kozumi and Kobayashi (2011)).

Returning to the ALD, by using the identity of Andrews and Mallows (1974),

$$\exp(-|ab|) = \int_0^\infty \frac{a}{\sqrt{2\pi v}} \exp\left\{-\frac{1}{2}(a^2 v + b^2 v^{-1})\right\} dv,$$

for any $a, b > 0$, letting $a = 1/\sqrt{2\sigma}$ & $b = \epsilon/\sqrt{2\sigma}$ and multiplying a factor of $\exp\{-(2\tau - 1)\epsilon/2\sigma\}$, to express the probability density function of the ALD errors as its scale mixture of normal representation,

$$f^{\mathrm{AL}}(\epsilon_i) = \int_0^\infty \frac{1}{\sqrt{4\pi\sigma^3 v_i}} \exp\left\{-\frac{(\epsilon_i - (1-2\tau)v_i)^2}{4\sigma v_i} - \frac{\tau(1-\tau)v_i}{\sigma}\right\} dv_i, \quad i = 1, \ldots, n,$$

as proposed by Reed and Yu (2009), and Kozumi and Kobayashi (2011). This representation can be utilised to facilitate Gibbs sampling algorithms (Kozubowski and Podgórski (2001), Geraci and Bottai (2007), Kozumi and Kobayashi (2011), and Chen et al. (2013), amongst others).

1.3 Regularised Quantile Regression

We are interested in selecting a subset of important predictors, which have adequate explanatory and predictive capabilities. One of the common procedures for simultaneously facilitating the parameter estimation and variable selection is to impose a penalty function on the likelihood to arrive at the penalised loss function. Regularisation has been shown to be effective in improving the predictive accuracy (Li and Zhu (2008), and Wu and Liu

(2009)). The Lasso, adaptive Lasso and Elastic Net estimates are all regularised estimates and the differences amongst them are only at their penalty terms. Specifically, they are all solutions to the following form of minimisation problem for regularised quantile regression

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} L_{\tau}^{\text{Quantile}}(y_i - \boldsymbol{x}_i\boldsymbol{\beta}) + \lambda_1 g_1(\boldsymbol{\beta}) + \lambda_2 g_2(\boldsymbol{\beta}), \tag{1.10}$$

for some $\lambda_1, \lambda_2 \geq 0$ and penalty functions $g_1(\cdot)$ and $g_2(\cdot)$. The Lasso corresponds to $\lambda_1 > 0$, $\lambda_2 = 0$, $g_1(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $g_2(\boldsymbol{\beta}) = 0$ (Tibshirani (1996)). The adaptive Lasso corresponds to $\lambda_1 = \lambda_{1j} > 0$, $j = 1, \ldots, k$, $\lambda_2 = 0$, $g_1(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $g_2(\boldsymbol{\beta}) = 0$ (Zou (2006)). The Elastic Net corresponds to $\lambda_1 > 0$, $\lambda_2 > 0$, $g_1(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ and $g_2(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$ (Zou and Hastie (2005)).

Letting $\tau = 0.5$, the first term of Equation (1.10) reduces to $\sum_{i=1}^{n} |y_i - \boldsymbol{x}_i\boldsymbol{\beta}|$ and the corresponding method is called the least absolute deviation (LAD) regression, which is known to be robust against outliers in response variables. However, the LAD regression might underestimate regression coefficients for non-outlying observations.

1.4 Bayesian Regularisation

Bayesian inference is one of the most popular approaches for the regression analysis. It makes inference for an entire posterior distribution of a parameter of interest, as well as incorporation of parameter uncertainty and prior information about data. This encourages the use of Bayesian analysis over standard frequentist approaches. We will present the existing Bayesian regularisation priors for regularised regression, such as regularised quantile regression.

The Bayesian Lasso, proposed by Park and Casella (2008), is defined as

$$\pi(\boldsymbol{\beta}|\lambda) = \prod_{j=1}^{k} \frac{\lambda}{2} \exp\left\{-\lambda|\boldsymbol{\beta}_j|\right\}, \tag{1.11}$$

where λ is the penalisation parameter.

The Bayesian adaptive Lasso, proposed by Sun et al. (2010), is defined as

$$\pi(\boldsymbol{\beta}|\boldsymbol{\lambda}) = \prod_{j=1}^{k} \frac{\lambda_j}{2} \exp\left\{-\lambda_j|\boldsymbol{\beta}_j|\right\}, \tag{1.12}$$

where λ_j is the penalisation parameter being assigned to each regression coefficient. What

makes the Bayesian adaptive Lasso advantageous over the Bayesian Lasso is that the adaptive Lasso penalty allows coefficient-specific penalties. Also, Figure 1 of Park and Casella (2008) showed that the coefficient paths of the Bayesian Lasso are a compromise between the coefficient paths of the Lasso and ridge regression. On the other hand, intuitively, Sun et al. (2010) suggested that the coefficient paths of the Bayesian adaptive Lasso approaches the log penalty by iteratively applying the adaptive Lasso penalty. Figure 1 of Sun et al. (2010) clearly outlined the difference between the Bayesian Lasso and the Bayesian adaptive Lasso.

Note that the Bayesian adaptive Lasso is originally proposed by Griffin and Brown (2007), and the main difference between the work of Griffin and Brown (2007) and Sun et al. (2010) is that the latter studied the fully Bayesian approach for the adaptive Lasso penalty in contrast to the former.

The Bayesian Elastic Net, proposed by Li et al. (2010), is defined as

$$\pi(\boldsymbol{\beta}|\lambda_3, \lambda_4) = \prod_{j=1}^{k} C\left(\tilde{\lambda}_3\, \lambda_4\right) \frac{\lambda_3}{2} \exp\left\{-\lambda_3|\beta_j| - \lambda_4\beta_j^2\right\}, \tag{1.13}$$

where $C\left(\tilde{\lambda}_3\, \lambda_4\right) = \Gamma^{-1}\left(1/2, \tilde{\lambda}_3\right)\left(\tilde{\lambda}_3\right)^{-1/2} \exp\left\{-\tilde{\lambda}_3\right\}$ is the normalising constant and $\tilde{\lambda}_3 = \lambda_3^2/(4\lambda_4)$. The computations of the normalising constant is provided in Appendix B of Li et al. (2010).

In contrast to the Lasso and adaptive Lasso penalties, the Elastic Net penalty is a flexible regularisation that uses a mixture of the Lasso and ridge penalties. It addresses three inherent issues of the Lasso method, as stated in Zou and Hastie (2005): (1) the Lasso method cannot select more regression coefficients than the sample size due to the nature of the convex optimisation in case of frequentist approach; (2) the Lasso method tends to select only one regression coefficient from a group, whilst disregarding others when there is some group structure amongst the regression coefficients; and (3) the Lasso method performs poorly in case of highly correlated regression coefficients.

We have presented the quantile regression analysis and Bayesian regularisation so far, and we need the computational methods to fit such models. The Gibbs sampling algorithm, as mentioned earlier, is one of the well-known Markov chain Monte Carlo (MCMC) methods in Bayesian analysis.

1.5 Markov Chain Monte Carlo

MCMC methods are powerful techniques for sampling from probability distributions, using Markov chains to approximate the posterior distribution of a parameter of interest. Typically, they are used in data modelling for different problems for Bayesian inference and numerical integration. One often deals with high-dimensional integrals or multivariate probability distributions where Bayesian analysis requires integrating over the posterior distribution of parameters of interest given the data. It is well known that Bayesian inference regarding unknown quantities is entirely based on their probabilistic description.

Let $\boldsymbol{y}$ be a vector of n independently and identically distributed observations and Θ a vector including latent variables and the parameters. We compute the posterior distribution by Bayes' rule,

$$p(\Theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \Theta)}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\Theta)p(\Theta)}{p(\boldsymbol{y})}, \tag{1.14}$$

where $p(\boldsymbol{y}, \Theta)$ is the joint posterior density of parameters and data, $p(\boldsymbol{y}|\Theta)$ is the likelihood of data given parameters, $p(\Theta)$ is the prior density of parameters, and $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\Theta)p(\Theta)d\Theta$ is the marginal likelihood or the normalising constant.

However, for the majority of regression models, including quantile regression, it is analytically infeasible to compute these integrals directly. This is when the MCMC methods are considered by either making inference about parameters of interest or making predictions from sampling. Fundamentally, they consist of Monte Carlo integration using Markov chain (Gilks et al. (1995)). Their basic concept will be described in the following subsections.

1.5.1 Monte Carlo Integration

Monte Carlo integration uses importance sampling to obtain the approximation of an integral. As in Gilks et al. (1995) and Hastings (1970), suppose that we sample $x_1, x_2, \ldots, x_N$ from the proposed distribution $f(x)$ and we desire to compute

$$\theta = \mathbb{E}[\phi(x)] = \int_{-\infty}^{\infty} \phi(x)f(x)dx, \tag{1.15}$$

where θ is a parameter of interest, $f(x)$ is a probability density function and $\phi(x)$ is neither close to a constant nor has a large variance. The integral in Equation (1.15) can be rewritten

as

$$\theta_g = \mathbb{E}[\psi(x)] = \int_{-\infty}^{\infty} \psi(x)g(x)dx, \tag{1.16}$$

where $\psi(x) = \phi(x)f(x)/g(x)$, and $g(x)$ is some probability density function. Then we can approximate the theoretical mean (Equation (1.16)) by sample mean to obtain the approximation,

$$\hat{\theta}_g = \mathbb{E}[\psi(x)] \approx \frac{1}{N}\sum_{i=1}^{N} \psi(x_i).$$

If $g(x)$ is appropriately chosen then the importance function $\psi(x)$ would be as close to a constant as possible. This results in the variance of $\hat{\theta}_g$ being significantly lower than that of $\hat{\theta}$ without importance sampling.

If $\{x_i\}_{i=1}^N$ are assumed to be independent random variables then it follows that the random variables are independently and identically distributed. Hence, by using the Strong Law of Large Numbers, the accuracy of approximation would increase, in other words,

$$\frac{1}{N}\sum_{i=1}^{N} \psi(x_i) \to \theta, \quad \text{as } N \to \infty.$$

However, in most cases, the samples being drawn from a density function are often not independent, since they may be correlated. To remedy this issue, Markov chain can be used to assume the stationary distribution (Gilks et al. (1995)).

1.5.2 Markov Chain

Suppose that we sample the sequence of random variables $x_0, x_1, \ldots$ from the distribution $\pi(x_{r+1}|x_r)$ such that next sample x_{r+1} depends only on the current state x_r, whilst it does not depend on the further history of sequence $x_0, x_1, \ldots, x_{r-1}$. Hence, this stochastic process is time-homogeneous. This type of sequence is called a Markov chain and $\pi(\cdot|\cdot)$ is called the transition kernel of the chain.

Therefore, the MCMC methods aim to construct generated chains that are progressively more likely realisations of the distribution of interest. We will briefly describe some commonly-used MCMC methods, including the Metropolis-Hastings algorithm (Hastings (1970)) and its special case, the Gibbs sampling algorithm (Geman and Geman (1984)).

1.5.3 Metropolis-Hastings Algorithm

The Metropolis algorithm was first introduced by Metropolis et al. (1953) then it was generalised to the Metropolis-Hastings algorithm by Hastings (1970). It is made into mainstream statistics and engineering, and all the other MCMC methods, including the Gibbs sampling algorithm, stem from the Metropolis-Hastings algorithm.

The Metropolis-Hastings algorithm is a random walk that uses an acceptance-rejection rule to converge to the target distribution. The acceptance-rejection rule is defined as

$$\omega\left(\Theta^{(r+1)}, \Theta^{(r)}\right) = \min\left(1, \frac{p\left(\Theta^{(r+1)}|\boldsymbol{y}\right) g\left(\Theta^{(r)}|\Theta^{(r+1)}, \boldsymbol{y}\right)}{p\left(\Theta^{(r)}|\boldsymbol{y}\right) g\left(\Theta^{(r+1)}|\Theta^{(r)}, \boldsymbol{y}\right)}\right),$$

where $g\left(\Theta^{(r)}|\Theta^{(r+1)}, \boldsymbol{y}\right)$ is the proposal distribution at state r and $p\left(\Theta^{(r+1)}|\boldsymbol{y}\right)$ is the posterior distribution at state $r+1$, as specified in Equation (1.14). Because the normalising constant does not depend on the parameters, we have $p\left(\Theta^{(r+1)}|\boldsymbol{y}\right) \propto p\left(\boldsymbol{y}|\Theta^{(r+1)}\right) p\left(\Theta\right)$.

The Metropolis-Hastings algorithm is iterative based on the acceptance-rejection rule above where generated samples converge to a target distribution. It is simple to implement because the computations only depend on the posterior and proposal distributions without requiring the normalising constant. Therefore, the marginal distribution does not need to be known, and no factorisation or integration of the posterior distribution is required.

1.5.4 Gibbs Sampling Algorithm

The Gibbs sampling algorithm, proposed by Geman and Geman (1984), is a special case of the Metropolis-Hastings algorithm wherein proposals are always accepted with a probability of 1. The idea of the Gibbs sampling algorithm is to draw a sequence of samples from a multivariate probability distribution. It transfers from multivariate sampling to one-dimensional sampling because it assumes independence amongst the samples (Geman and Geman (1984)). These samples approximate the marginal distribution of one of the parameters, or some subset of the parameters. In short, the algorithm generates a sample from the univariate distribution of each parameter in turn, conditional on the current values of the other variables.

Suppose we want to obtain N samples, $\Theta^{(0)} = \left(\theta_1^{(0)}, \ldots, \theta_N^{(0)}\right)$, from a joint posterior distribution, $p(\Theta)$, as the initial samples. At state $r+1$, we need to draw a new set of samples, $\Theta^{(r+1)} = \left(\theta_1^{(r+1)}, \ldots, \theta_N^{(r+1)}\right)$. To sample each component, $\theta_l^{(r+1)}$ $(l = 1, \ldots, N)$. we update

it according to the univariate distribution specified by $p\big(\theta_l^{(r+1)}|\theta_1^{(r+1)},\ldots,\theta_{l-1}^{(r+1)},\theta_1^{(r)},\ldots,\theta_N^{(r)}\big)$. We use the $(r+1)^{\text{th}}$ component depending on the l^{th} samples. This follows from the basic idea of the Markov chain. Finally, we iterate the procedure until the set of sample, $\Theta^{(r+1)}$, reaches the target distribution.

1.6 Variational Inference

Variational inference (VI) is useful for Bayesian inference and a method to deal with the approximation of probability densities. Generally, this technique exchanges sampling, as in MCMC procedures, for optimisation. By choosing a flexible family of approximate densities, we search for a member of this family, which minimises some optimal criterion, for example, the Kulback-Leibner (KL) divergence. The variational Bayesian (VB) method is useful for approximating intractable or difficult-to-compute posterior distribution with some optimal density. Compared to MCMC, the VB method is better in a fast computational problem, whilst achieving a comparable prediction (Blei et al. (2017)). We first start by describing the well-known mean-field VI for approximate posterior densities that can be normalised in closed form belonging to some conjugate family of densities.

1.6.1 Mean-field Variational Inference

The basic concepts behind the VB method can be easily followed in Blei et al. (2017) and Ormerod and Wand (2010). Several examples are presented in the Bishop book (Bishop and Nasrabadi (2006)). The log marginal data distribution, also known as evidence integral, is denoted by $\log p(\boldsymbol{y})$. Evidence integrals are often unavailable in closed form and requires exponential time to be evaluated. To avoid calculating the evidence integrals, one searches for a lower bound, which is known as evidence lower bound (ELBO) and will be denoted as $\text{LB}(q)$.

Jensen's inequality (Jensen (1906)) states that if a function f is concave then we have

$$f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)].$$

We use this Jensen's inequality on the log marginal data distribution,

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log \int_{\Theta} p(\boldsymbol{y}, \Theta) d\Theta \\
&= \log \int_{\Theta} p(\boldsymbol{y}, \Theta) \frac{q(\Theta)}{q(\Theta)} d\Theta \qquad (1.17) \\
&= \log \left(\mathbb{E}_{q(\Theta)} \left[\frac{p(\boldsymbol{y}, \Theta)}{q(\Theta)} \right] \right) \\
&\geq \mathbb{E}_{q(\Theta)} \left[\log p(\boldsymbol{y}, \Theta) \right] - \mathbb{E}_{q(\Theta)} \left[\log q(\Theta) \right] \\
&= \mathrm{LB}(q). \qquad (1.18)
\end{aligned}$$

This is the ELBO. Note that the term $-\mathbb{E}_{q(\Theta)} \left[\log q(\Theta) \right]$ is the entropy, which is another quantity from the theory. The ELBO in Equation (1.18) has the relation with the KL divergence. By using Bayes' rule (Equation (1.14)), we compute the KL divergence as

$$\begin{aligned}
\mathrm{KL}(q(\Theta) \| p(\boldsymbol{y}, \Theta)) &= \mathbb{E}_{q(\Theta)} \left[\log \frac{q(\Theta)}{p(\Theta|\boldsymbol{y})} \right] \\
&= \mathbb{E}_{q(\Theta)} \left[\log q(\Theta) \right] - \mathbb{E}_{q(\Theta)} \left[\log p(\Theta|\boldsymbol{y}) \right] \\
&= \mathbb{E}_{q(\Theta)} \left[\log q(\Theta) \right] - \mathbb{E}_{q(\Theta)} \left[\log p(\boldsymbol{y}, \Theta) \right] + \mathbb{E}_{q(\Theta)} \left[\log p(\boldsymbol{y}) \right] \\
&= - \left(\mathbb{E}_{q(\Theta)} \left[\log p(\boldsymbol{y}, \Theta) \right] - \mathbb{E}_{q(\Theta)} \left[\log q(\Theta) \right] \right) + \log p(\boldsymbol{y}) \\
&= -\mathrm{LB}(q) + \log p(\boldsymbol{y}),
\end{aligned}$$

since $\log p(\boldsymbol{y})$ does not depend on $q(\Theta)$. Here, $\mathrm{KL}(q\|p)$ is the KL divergence between the approximate distribution on the latent variables and the posterior distribution of the latent variables.

Thus, minimising the KL divergence is the same as maximising the lower bound. In other words,

$$\arg\min_{q} \mathrm{KL}(q\|p) \simeq \arg\max_{q} \mathrm{LB}(q).$$

Also, the KL divergence is always positive that is, $\mathrm{KL}(q\|p) \geq 0$ with equality if and only if $p(\Theta|\boldsymbol{y}) = q(\Theta)$ (Kullback and Leibler (1951)). Generally, it is difficult to obtain this posterior distribution, thus, the approach is to choose a family of tractable densities. We make the following assumption,

$$q(\Theta) = \prod_{l}^{N} q_l(\theta_l), \qquad (1.19)$$

where a partition of the Θ into N disjoint groups is denoted as θ_l.

The central idea is to maximise each factor (blocks of θ's) of $q(\Theta)$ in turn. We keep $q_{l\neq h}$ fixed and maximise $\mathrm{LB}(q)$. Note that:

$$\begin{aligned}
\mathrm{LB}(q) &= \int_\Theta \prod_l^N q_l(\theta_l) \log p(\boldsymbol{y}, \Theta) d\Theta - \int_\Theta \prod_l^N q_l(\theta_l) \log \left(\prod_l^N q_l(\theta_l) \right) d\Theta \\
&= \int_\Theta \log p(\boldsymbol{y}, \Theta) \prod_l^N q_l(\theta_l) d\Theta - \int_\Theta \prod_l^N q_l(\theta_l) \sum_l^N \log q_l(\theta_l) d\Theta \\
&= \int_{\theta_h} q_h(\theta_h) \left(\int \log p(\boldsymbol{y}, \Theta) \prod_{l\neq h} q_l(\theta_l) \right) d\theta_h - \int_{\theta_h} q_h(\theta_h) \log q_h(\theta_h) d\theta_h + \text{const} \\
&= \int_{\theta_h} q_h(\theta_h) \log \tilde{p}(\boldsymbol{y}, \theta_h) d\theta_h - \int_{\theta_h} q_h(\theta_h) \log q_h(\theta_h) d\theta_h + \text{const}, \qquad (1.20)
\end{aligned}$$

where $\tilde{p}(\boldsymbol{y}, \theta_h) = \mathbb{E}_{l\neq h}[\log p(\boldsymbol{y}, \Theta)]$ and $\mathbb{E}_{l\neq h}[\cdot]$ is the expectation evaluated for $q_{l\neq h}(\theta_{l\neq h}) = \prod_{l\neq h} q_l(\theta_l)$. Note that $\mathrm{LB}(q)$ depends on the variational parameters.

Because Equation (1.20) is equal to $-\mathrm{KL}(q\|p)$, maximising this equation is equivalent to minimising KL. Thus, the optimal solution is

$$q^*(\theta_h) = \exp\left\{ \mathbb{E}_{l\neq h}[\log p(\boldsymbol{y}, \Theta)] \right\}. \qquad (1.21)$$

Because all the latent variables are assumed to be independent in the family of tractable densities, the expectations on the right-hand side do not involve the h^{th} variational factor $q_h(\theta_h)$. Equation (1.21) implies that $\mathbb{E}_{l\neq h}[\cdot]$ is not associated with the h^{th} variational factor $q_h(\theta_h)$. The optimal variational density $q^*(\theta_h)$ can be achieved when $q^*(\theta_{l\neq h})$ on the right-hand side are the optimal choices. This results in the optimisation problem and the algorithm used to solve this problem is coordinate ascent VI (CAVI). The CAVI algorithm, proposed by Bishop and Nasrabadi (2006), iterates between updating $q^*(\theta_h)$ via Equation (1.21), whilst keeping other approximate densities fixed, and updating others, whilst keeping $q^*(\theta_h)$ fixed. After the algorithm converges, we take the optimal variational density $q^*(\theta_h)$ as an optimal estimate of parameter θ_h.

Moreover, $q^*(\theta_h)$ depends on the full conditional distributions, as usually denoted in the MCMC literature (Casella and George (1992)). Therefore, there is a natural link with the Gibbs sampling algorithm. Even so, the proposed approach leads to tractable solutions involving only local operations.

1.6.2 Laplace Variational Inference

Often, the mean-field VI approach is infeasible when a desired approximate density does not belong to the family of tractable densities, or in other words, they are not conjugate to any available closed-form distributions. Laplace VI method is alternatively proposed by Wang and Blei (2013) that embeds Laplace approximation (Tierney et al. (1989), and MacKay (1992)) within a variational optimisation algorithm.

Suppose that we have a variable amongst the parameter vector Θ, which is non-conjugate, denoted as θ_t. We arrive at the following coordinate update,

$$\begin{aligned} q_t(\theta_t) &= \exp\left\{\mathbb{E}_{l\neq t}[\log p(\boldsymbol{y}, \Theta)] + \text{const}\right\} \\ &\propto \exp\left\{\chi(\theta_t)\mathbb{E}_{l\neq t}[\xi(\Theta_{l\neq t})] + \log p(\theta_t)\right\}, \end{aligned} \tag{1.22}$$

where $\chi(\theta_t)$ is a function of θ_t that are to be updated, $\xi(\Theta_{l\neq t})$ is a function of the remaining components of Θ except the component θ_t, and $p(\theta_t)$ is the prior density of θ_t. Define the function $h(\theta_t)$ to contain the terms inside the exponent of the update (Equation (1.22)),

$$h(\theta_t) = \chi(\theta_t)\mathbb{E}_{l\neq t}[\xi(\Theta_{l\neq t})] + \log p(\theta_t).$$

The terms of $h(\theta_t)$ come from the model and involve $q_{l\neq t}(\Theta_{l\neq t})$ or θ_t. Note that we can compute $\mathbb{E}_{l\neq t}[\xi(\Theta_{l\neq t})]$ via the mean VI approach. Since $q_t(\theta_t) \propto \exp\{h(\theta_t)\}$ cannot be normalised in closed form, instead, we approximate the coordinate update by a second-order Taylor approximation of $h(\theta_t)$ around its maximum, $\hat{\theta}_t$. This follows the same logic as from the Laplace approximation. The Taylor approximation for $h(\theta_t)$ around $\hat{\theta}_t$ is

$$h(\theta_t) \approx h(\hat{\theta}_t) + \nabla h(\hat{\theta}_t)(\theta_t - \hat{\theta}_t) + \frac{1}{2}(\theta_t - \hat{\theta}_t)^2\nabla^2 h(\hat{\theta}_t), \tag{1.23}$$

where $\nabla h(\hat{\theta}_t)$ and $\nabla^2 h(\hat{\theta}_t)$ are the first and second derivatives estimated at $\hat{\theta}_t$, respectively. Notice that $\hat{\theta}_t$ is the maximum value of θ_t and its gradient $\nabla h(\hat{\eta})$ is 0. This implies that the term $\nabla h(\hat{\theta}_t)(\theta_t - \hat{\theta}_t)$ vanishes. Then Equation (1.23) is simplified to

$$q_t(\theta_t) = \exp\left\{h(\hat{\theta}_t) + \frac{1}{2}(\theta_t - \hat{\theta}_t)^2\nabla^2 h(\hat{\theta}_t)\right\}.$$

Therefore, the approximate update for $q_t(\theta_t)$ is

$$q^*(\theta_t) \approx N\left(\hat{\theta}_t, -\nabla^2 h(\hat{\theta}_t)^{-1}\right).$$

Observe that its normal form stems from Taylor approximation so, we do not assume the normality. During the CAVI algorithm, we require the computation of some expectation of θ_t. Due to several great properties of a normal distribution, we can compute them directly, for example, $\mathbb{E}[\theta_t] = \hat{\theta}_t$ or $\text{Var}(\theta_t) = -\nabla^2 h(\hat{\theta}_t)^{-1}$. Otherwise we may take a second-order Taylor approximation around some difficult-to-compute function $g(\theta_t)$ given the optimal update $q^*(\theta_t)$ for computing $\mathbb{E}[g(\theta_t)]$.

1.7 Thesis Motivation

Collectively, the introductory chapter has presented different loss functions, quantile regression, Bayesian regularisation and Bayesian methods. As Bayesian analysis is capable of full probabilistic uncertainty quantification, this thesis focuses on the development of novel Bayesian methods in parametric statistical inference tackling different issues that are as follows.

Non-linearity is prominent when quantifying the relationship between a response variable and predictors, particularly in medical applications. Quantile regression is based on the conditional quantile function that permits a more complete portrayal of the relationship as well as is being robust to outliers. This led to tackling the first issue in this thesis, namely non-linearity in relationships, whilst taking into account of robustness, by introducing the quantile-dependent regularised prior to the non-linear model in facilitating the variable selection method for a specific medical application.

However, quantile regression amongst several regression models faces challenging issues in a high-dimensional problem and at the same time, asymmetric distributions, including heavy-tailed distributions are difficult to work with and are required to take into account the robustness of methods. Whilst quantile regression enjoys the benefit of robustness, it has different modelling aims from robust regression. Thus, this thesis tackles the second issue for a high-dimensional problem by proposing the novel loss function having properties of asymmetry, heavy-tailedness and robustness, and Bayesian robust regularisation. This led to a new variant of Bayesian regularised quantile regression model using the conventional method. Yet, when the amount of data increases, the conventional method becomes computationally burdensome and the alternative approach is needed. This thesis further tackles the third issue for a computational problem by replacing the conventional method with the approximate-based technique that involves optimisation problems.

1.8 Thesis Outline

The outline of the thesis is as follows. Chapter 2 proposes a Bayesian variable selection with parametric non-linear quantile regression model by employing Bayesian variable selection with quantile-dependent prior for the fractional polynomial (FP) model. We explore the application in the analysis of blood pressure (BP) amongst United Status (US) adults. Higher BP results in hypertension, which is a highly prevalent chronic medical condition and a strong risk factor for cardiovascular disease (CVD). The FPs act as a concise and accurate formula for examining the smooth relationship between response variable and predictors. Since modelling conditional mean functions observes the partial view of a distribution of response variable, conditional quantile functions with FPs provide comprehensive relationships between the response variable and its predictors, such as median and extremely high BP measures. Then modelling extremely high BP could explore CVD insight deeply and precisely. The aim of this chapter is to examine a non-linear relationship between BP measures and their risk factors across median and upper quantile levels using data extracted from the 2007-2008 National Health and Nutrition Examination Survey (NHANES). The comparative studies are conducted with existing methods and the data analysis is provided.

Chapter 3 proposes a new Bayesian Huberised regularisation and its extension to quantile regression taking account of robustness, asymmetry and heavy tails in a high-dimensional setting. Robust regression has recently received a great amount of attention in the literature, particularly for taking asymmetricity into account simultaneously and for high-dimensional analysis. The literature review suggests that the majority of research on the topics falls in frequentist approaches, which are not capable of full probabilistic uncertainty quantification. This motivates the development of robust Bayesian regularisation. Firstly, the chapter proposes a new Huberised-type of asymmetric loss function and its corresponding probability distribution, which has the scale mixture of normal representation. Secondly, we introduce a new Bayesian Huberised regularisation for robust regression. Finally, a by-product of the research is that a new Bayesian Huberised regularised quantile regression is also derived. We further study the theoretical properties. We compare the proposed method with existing regularised methods in recent literature under this topic, via extensive simulation experiments. Three real data examples are also given.

Chapter 4 proposes a novel VB regularisation. It follows from Chapter 3 that MCMC methods are common techniques for Bayesian probabilistic models where posterior distributions cannot be computed directly. However, they are not desirable due to extremely high com-

putational cost when the amount of data information increases. The alternative approach is needed. This chapter builds upon Chapter 3 by replacing its MCMC method with the VB method based on KL divergence, to propose VB Huberised Lasso quantile regression and VB Huberised adaptive Lasso quantile regression for a fast-computational and high-dimensional problem. Both mean-field VI and Laplace VI methods are used to compute approximate densities in place of exact posterior distributions. The CAVI algorithms with their ELBO are derived. The computational performance is compared between the proposed VB algorithms and the MCMC method from Chapter 3 via a wide variety of simulation studies, and a real data example is also provided.

And finally, Chapter 5 provides a brief of the thesis and outlines some future research directions along the research topics in the thesis.

Appendix A contains the mathematical proofs of the quantile property of the probability density of asymmetric Huberised loss function, the posterior propriety and unimodality of the joint posterior density for Bayesian Huberised regularisation in Chapter 3.

Appendix B provides the derivation of the full Gibbs sampling algorithms for the given hierarchical models of Bayesian Huberised regularised quantile regression in Chapter 3.

Appendix C provides the derivation of the ELBO required for the VB algorithms in Chapter 4.

Appendix D contains the figures for the data analysis in Chapter 2 and those for the simulation studies in Chapters 3-4.

Chapter 2

Bayesian Fractional Polynomial Approach to Quantile Regression and Variable Selection

Polynomial regression quantifies the non-linear relationship between the response variable and predictors parametrically. FP regression is the extension of polynomial regression to include the fractional orders, which allows for the smoother relationship. Modelling conditional quantile function observes the full view of a distribution of a response variable, providing comprehensive relationships. This chapter presents a Bayesian variable selection with parametric non-linear quantile regression model with application to the analysis of BP amongst United States (US) adults. The comparative studies between the proposed method and existing methods show that the variable selection signified the importance of the FP model in the Bayesian quantile regression model.

2.1 Introduction

Over the past three decades, the number of adults aged 30-79 with hypertension has increased from 648 million to 1.278 billion globally (Zhou et al. (2021)). Worldwide, approximately 17.9 million deaths each year are caused by CVD (World Health Organisation (2013a)). Out of these human losses, high BP accounts for approximately 9.4 million deaths globally every year (Lim et al. (2012)). Hypertension is a highly prevalent chronic medical condition and a strong modifiable risk factor for CVD, as it attributes to more than 45% of CVD and 51% of stroke deaths (World Health Organization (2013b)). The risk of CVD

in individuals rises sharply with increasing BP (Prospective Studies Collaboration (2002), Ettehad et al. (2016), Navar et al. (2016), Bundy et al. (2017), and Clark et al. (2019)).

Continuous BP measurement has proven to be one of effective incident prevention. This implies that BP is the essential physiological indicator of the human body. When the heart beats, it pumps blood to the arteries resulting in changes in BP during the process. When the heart contracts, BP in the vessels reaches its maximum, which is known as systolic BP (SBP). When the heart rests, BP reduces to its minimum, which is known as diastolic BP (DBP).

Linear regression and polynomial regression analyses have been used in assessing the association between BP and risk factors contributing to various diseases (Koh et al. (2022), Liu et al. (2022), and Yeo et al. (2022), amongst others). It is evident that the polynomial regression models fit the data accurately in some research studies due to its adaptability of non-linearity property, yet face high order polynomial approximation. The FPs, proposed by Royston and Altman (1994), act as a concise and accurate formulae for examining smooth relationships between response and predictors, and a compromise between precision and generalisability. The FPs are parametric in nature and then intuitive for the interpretation of the analysis results. The FP approach has clearly established a role in the non-linear parametric methodology, especially with application by clinicians from various research fields, such as obstetrics and gynaecology (Tilling et al. (2014)), gene expression studies in clinical genetics (Tan et al. (2011)) and cognitive function of children (Ryoo et al. (2017)), and other medical applications (Wong et al. (2011), Ravaghi et al. (2020), and Frangou et al. (2021), amongst others).

However, modelling conditional mean functions observes the partial view of a distribution of response variables, as the distributions of many response variables such as the BP measures are typically skew. Then 'average' BP may link to CVD, yet extremely high BP could explore CVD insight deeply and precisely. So, existing mean-based FP approaches for modelling the relationship between factors and BP cannot answer key questions in need. It is attractive to model conditional quantile functions with FPs that accommodate skewness readily. Quantile regression, introduced by Koenker and Bassett (1978), provides comprehensive relationships between the response variable and its predictors, which are useful for median and extremely high BP measures in practical data analysis generally.

Zhan et al. (2021) suggested quantile regression with FP as a suitable approach for an application, such as age-specific reference values of discrete scales, in terms of model con-

sistency, computational cost and robustness. This approach is also used to derive reference curves and reference intervals in several applications (Chitty and Altman (2003), Bell et al. (2010), Bedogni et al. (2012), Kroon et al. (2017), Casati et al. (2019), Cai et al. (2020), and Loef et al. (2020), amongst others), which allow quantiles to be estimated as a function of predictors without requiring parametric distributional assumptions. This is essential for data that do not assume normality, linearity and constant variance. Recently, reasonable amount of non-linear quantile regression analyses have been conducted in medical data analysis (Maidman and Wang (2018), Huang et al. (2023), and Wu et al. (2023), amongst others).

However, Bayesian approach to quantile regression has advantages over the frequentist approach, as it can lead to exact inference in estimating the influence of risk factors on the upper quantiles of the conditional distribution of BP compared to the asymptotic inference of the frequentist approach (Yu et al. (2005)). It also provides estimation that incorporates parameter uncertainty fully (Yu and Moyeed (2001), and Yu et al. (2005)). Some comparison studies have been conducted for both Bayesian and frequentist approaches, such as the analysis of risk factors for female CVD patients in Malaysia (Juhan et al. (2020)) and the analysis of risk factors of hypertension in South Africa (Kuhudzai et al. (2022)). The former revealed that the Bayesian approach has smaller standard errors than that of the frequentist approach. The latter also revealed that credible intervals of the Bayesian approach are narrower than confidence intervals of the frequentist approach. These findings suggested that the Bayesian approach provides more precise estimates than the frequentist approach.

Variable selection in Bayesian quantile regression has been widely studied in the literature (Li et al. (2010), Alhamzawi et al. (2012), Alhamzawi and Yu (2013), Chen et al. (2013), Adlouni et al. (2018), Alhamzawi et al. (2019), and Dao et al. (2022), amongst others). It plays an important role in building a multiple regression model, provides regularisation for good estimation of effects, and identifies important variables. Sabanés Bové and Held (2011) combined variable selection and 'parsimonious parametric modelling' of Royston and Altman (1994) to formulate a Bayesian multivariate FP model with variable selection that efficiently selects best fitted FP model via stochastic search algorithm. However, in the present, no research studies have been conducted for variable selection in Bayesian parametric non-linear quantile regression for medical application, even though there is a limited amount of studies in case of non-regularised models, such as mixed effect models (Wang (2012), and Yu and Yu (2023)).

Therefore, in this chapter, we explore a new quantile regression model using FPs and employ Bayesian variable selection with quantile-dependent prior for a more accurate representation of the risk factors on BP measures. The three-stage computational scheme of Dao et al. (2022) is employed as a variable selection method due to its fast convergence rate, low approximation error and guaranteed posterior consistency under model misspecification. So, we propose a Bayesian variable selection with non-linear quantile regression model to assess how body mass index (BMI) amongst US adults influences BP measures, including SBP and DBP. The objective of this chapter is to examine non-linear relationships between BP measures and their risk factors across median and upper quantile levels. The dataset used in this chapter is the 2007-2008 NHANES, including the information on BP measurements, body measures and socio-demographic questionnaires.

Section 2.2 presents the concept of FPs (Royston and Altman (1994)) and Bayesian variable selection with quantile-dependent prior (Dao et al. (2022)). The details of the NHANES 2007-2008 dataset used for the analysis are provided in Section 2.3. Section 2.4 applies the proposed method to the analysis, performs comparative analysis with two quantile regression methods and provides all the findings. Section 2.5 concludes this chapter.

2.2 Methodology

Rather than the conventional linear model, we will be using the FP model to develop the non-linear model under Bayesian quantile regression and variable selection. FPs provide a powerful and flexible extension to conventional polynomial regression. They enable the modelling of relationships that might not be linear, thereby capturing the subtleties of medical surveys more effectively. The utility of FP quantile regression models scrutinises medical surveys and the flexibility of the FP model is exploited to refine quantile regression models for superior accuracy in regression. The inclusion of the variable selection method within Bayesian quantile regression further aids in validating the choice of the model for better data fitting and capturing comprehensive relationships, whilst adopting a non-linear FP function (Royston and Altman (1994), and Royston and Sauerbrei (2008)).

2.2.1 Fractional Polynomials

Box and Tidwell (1962) introduced the transformation now known as the Box-Tidwell transformation,

$$x^{(a)} = \begin{cases} x^a, & \text{if} \quad a \neq 0, \\ \log(x), & \text{if} \quad a = 0, \end{cases}$$

where a is a real number. Royston and Altman (1997) extended the classical polynomials to a class which they called FPs.

An FP of degree m with powers $p_1 \leq \ldots \leq p_m$ and corresponding coefficients $\alpha_1, \ldots, \alpha_m$ is

$$f^m(x; \boldsymbol{\alpha}, \boldsymbol{p}) = \sum_{j=1}^{m} \alpha_j h_j(x),$$

where $h_0(x) = 1$ and

$$h_j(x) = \begin{cases} x^{(p_j)}, & \text{if } p_j \neq p_{j-1} \ , \\ h_{j-1}(x)\log(x), & \text{if } p_j = p_{j-1}, \end{cases} \tag{2.1}$$

for $j = 1 \ldots, m$. Note that the definition $h_j(x)$ allows the repeated powers. The bracket around the exponent denotes the Box-Tidwell transformation (Equation (2.1)). For $m \leq 3$, Royston and Altman (1994) constrained the set of possible powers p_j to the set

$$\mathrm{S}^{\mathrm{FP}} = \left\{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2, 3\right\}, \tag{2.2}$$

which encompasses the classical polynomial powers $1, 2, 3$, yet also offers square roots and reciprocals. Royston and Sauerbrei (2008) argued that this set is sufficient to approximate all powers in internals $[-2, 3]$. The simple example of the FP model is as follows.

Example 2.1. *An FP with $m = 3$ powers and its power vector $\boldsymbol{p} = (p_1, p_2, p_3) = (-1/2, 2, 2)$ would be*

$$f^3(x; \boldsymbol{\alpha}, \boldsymbol{p}) = \alpha_1 x^{-1/2} + \alpha_2 x^2 + \alpha_3 x^2 \log(x),$$

where the last term reflects the repeated power 2.

Generalisation to the case of multiple predictors:

$$\eta(\boldsymbol{x}) = \sum_{l=1}^{k} f_l^{m_l}(x_l; \boldsymbol{\alpha}_l, \boldsymbol{p}_l) = \sum_{l=1}^{k} \sum_{j=1}^{m_l} \alpha_{lj} h_{lj}(x_l). \tag{2.3}$$

This is called the multiple FP model. Suppose we continue examining k continuous predictors $x_1, \ldots, x_k$ and content themselves with a maximum degree of $m_{\max} \leq 3$ for each $f_l^{m_l}$, for instance, $0 \leq m_l \leq m_{\max}$ for $l = 1, \ldots, k$, where $m_l = 0$ denotes the omission of x_l from the model. From the powers set S^{FP}, m_l powers are chosen, which need not be different due to the inclusion of logarithmic terms for repeated powers (Equation (2.1)), we now employ the τ^{th} non-linear quantile regression with the scale mixture of normal representation of the ALD errors,

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \kappa_1 \boldsymbol{v} + \sqrt{\kappa_2 \boldsymbol{v} \sigma^2} \boldsymbol{z}, \tag{2.4}$$

where the $(n \times D)$-matrix $\boldsymbol{B}$ is a function of the l^{th} predictor for the i^{th} observations, x_{il} ($i = 1, \ldots, n$, and $l = 1, \ldots, k$), the unknown parameter vector $\boldsymbol{\beta} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)^T$ with $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{lm_l})$ for $l = 1, \ldots, k$, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ is a vector of exponential random variables with a rate of $\tau(1-\tau)/\sigma$, $\boldsymbol{z} = (z_1, \ldots, z_n)^T$ is a vector of standard normal random variables and z_i is independent of v_i for $i = 1, \ldots, n$, $\kappa_1 = (1 - 2\tau)/(\tau(1 - \tau))$, and $\kappa_2 = 2/(\tau(1 - \tau))$. Each entry of matrix $\boldsymbol{B}$ is a vector, $\boldsymbol{B}_{id} = \boldsymbol{B}(x_{id}) = (h_{l1}(x_{il}), \ldots, h_{lm_l}(x_{il}))^T$, for $i = 1, \ldots, n$, $l = 1, \ldots, k$, and $d = 1, \ldots, D$.

A special way of defining the matrix $\boldsymbol{B}$ is through the use of FPs. In this case, the basis function $B(\boldsymbol{x}_l)$ is chosen as the transformation h_{lj} in Equation (2.3) ($j = 1, \ldots, m_l$). The transformation $\boldsymbol{h}_j$ is determined by the power vector $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k$ through their definition (Equation (2.1)). Note that the $\boldsymbol{p}_l$ is empty if the predictor $\boldsymbol{x}_l$ is not included in the model ($m_l = 0$).

2.2.2 Bayesian Approach and Variable Selection

Given the model in Equation (2.4), the likelihood function conditional on $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_D)^T$, σ, $\boldsymbol{v} = (v_1, \ldots, v_n)^T$ can be written as

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma, \boldsymbol{v}, \boldsymbol{B}) = \prod_{i=1}^{n} \frac{1}{\sqrt{4\pi\sigma^3 v_i}} \exp\left\{ -\frac{(y_i - \boldsymbol{B}(\boldsymbol{x}_i)^T\boldsymbol{\beta} - (1-2\tau)v_i)^2}{4\sigma v_i} - \frac{\tau(1-\tau)v_i}{\sigma} \right\}.$$

We employ the three-stage algorithm of Dao et al. (2022) for Bayesian non-linear quantile

regression with variable selection. It can be summarised, as follows.

The first-stage is the expectation-maximisation (EM) algorithm consisting of two main steps: the Expectation step (E step) and the Maximum step (M step). Dempster et al. (1977) proposed the EM algorithm, which is a statistical simulation method and it aims to solve the complex data analysis problem with missing data.

Suppose the complete data $(\boldsymbol{y}, \boldsymbol{v})$ is composed of the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and missing data $\boldsymbol{v} = (v_1, \ldots, v_n)^T$, whereas $\boldsymbol{B}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, is treated as a function of fixed predictors. Maximum likelihood estimates (MLE) can be obtained by maximising log-likelihood function $\log f(\boldsymbol{\beta}, \sigma|\boldsymbol{y}, \boldsymbol{v})$ of the complete data. The EM algorithm has the following two steps: the E step and the M step

- [E step] Given initial values of $\boldsymbol{\beta}^{(0)}$ and $\sigma^{(0)}$, we denote $\boldsymbol{\beta}^{(q-1)}$ and $\sigma^{(q-1)}$ as the $(q-1)^{\text{th}}$ iteration value of parameters $\boldsymbol{\beta}$ and σ in the EM algorithm, and we define the mathematical expectation of the complete data as a Q-function

$$Q(\boldsymbol{\beta}, \sigma|\boldsymbol{y}, \boldsymbol{\beta}^{(q-1)}, \sigma^{(q-1)}) = \mathbb{E}_{\boldsymbol{y}, \boldsymbol{\beta}^{(q-1)}, \sigma^{(q-1)}}[\log f(\boldsymbol{\beta}, \sigma|\boldsymbol{y}, \boldsymbol{v})].$$

- [M step] We obtain the updated values of $\boldsymbol{\beta}^{(q)}$ and $\sigma^{(q)}$ by maximising $Q(\boldsymbol{\beta}, \sigma|\boldsymbol{y}, \boldsymbol{\beta}^{(q-1)}, \sigma^{(q-1)})$ over parameters $\boldsymbol{\beta}$ and σ:

$$\boldsymbol{\beta}^{(q)} = (\boldsymbol{B}^T \boldsymbol{W}^{(q-1)} \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{W}^{(q-1)} (\boldsymbol{y} - \kappa_1 \boldsymbol{\Delta 3}),$$

where

$$\boldsymbol{\Delta 3} = \left(\left| y_1 - \boldsymbol{B}(x_1)^T \boldsymbol{\beta}^{(q-1)} \right|, \ldots, \left| y_n - \boldsymbol{B}(x_n)^T \boldsymbol{\beta} \right|^{(q-1)} \right)^T,$$
$$\boldsymbol{W}^{(q-1)} = \text{diag}(1/\Delta 3_1, \ldots, 1/\Delta 3_n),$$

and

$$\sigma^{(q)} = \frac{1}{2(3n+2)} \left\{ \sum_{i=1}^{n} \Delta 2_i + \sum_{i=1}^{n} \frac{(y_i - \boldsymbol{B}(\boldsymbol{x}_i)^T \boldsymbol{\beta}^{(q)})^2}{\Delta 3_i} - 2\kappa_1 \sum_{i=1}^{n} (y_i - \boldsymbol{B}(\boldsymbol{x}_i)^T \boldsymbol{\beta}^{(q)}) \right\},$$

where $\Delta 2_i = \left| y_i - \boldsymbol{B}(\boldsymbol{x}_i)^T \boldsymbol{\beta}^{(q-1)} \right| + 2\sigma^{(q-1)}$ for $i = 1, \ldots, n$.

Repeat both E-step and M-step until the EM algorithm meets the required condition, then the final iteration values are set as the posterior modes of $\boldsymbol{\beta}$ and σ, denoted by $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}$, respectively.

The second-stage algorithm is the Gibbs sampling algorithm. The quantile-specific Zellner's g-prior (Alhamzawi and Yu (2013)) is used for the prior specification and it is given by

$$\boldsymbol{\beta}|\sigma, \boldsymbol{V}, \boldsymbol{B} \sim \mathrm{N}\left(0, 2\sigma g \boldsymbol{\Sigma}_v^{-1}\right) \quad \text{and} \quad p(\sigma) \propto \frac{1}{\sigma}, \tag{2.5}$$

where $\mathrm{N}(\cdot)$ is the multivariate Normal distribution, g is a scaling factor, $\boldsymbol{V} = \mathrm{diag}(1/v_1, \ldots, 1/v_n)$, and $\boldsymbol{\Sigma}_v = \boldsymbol{B}^T \boldsymbol{V} \boldsymbol{B}$. This prior specification has an advantage, as it contains information that is dependent upon the quantile levels, which increases posterior inference accuracy.

Given the posterior modes, $\tilde{\boldsymbol{\beta}}$ and $\tilde{\sigma}$ as the starting value, we denote $\boldsymbol{\beta}^{(r-1)}$ and $\sigma^{(r-1)}$ as the $(r-1)^{\text{th}}$ iteration value of parameters $\boldsymbol{\beta}$ and σ in the Gibbs sampling algorithm.

- Sample $v_i^{(r)}$ from

$$p(v_i|\boldsymbol{y}, \boldsymbol{\beta}^{(r-1)}, \sigma^{(r-1)}) \sim \mathrm{GIG}\left(0, \frac{1}{2\sigma}, \frac{(y_i - \boldsymbol{B}(\boldsymbol{x}_i)^T\boldsymbol{\beta})^2 + \frac{1}{g}\boldsymbol{\beta}^T\boldsymbol{B}(\boldsymbol{x}_i)\boldsymbol{B}(\boldsymbol{x}_i)^T\boldsymbol{\beta}}{2\sigma}\right),$$

  for $i = 1, \ldots, n$, and $\mathrm{GIG}(x|\nu, c, d)$ is the generalised inverse Gaussian (GIG) with its density,

$$f_{\mathrm{GIG}}(x) = \frac{(c/d)^{\nu/2}}{2K_1\left(\sqrt{cd}\right)} x^{\nu-1} \exp\left(-\frac{1}{2}(cx + dx^{-1})\right), \quad v > 0, \tag{2.6}$$

  where $K_\nu(\cdot)$ is the modified Bessel function of the second kind at index ν (Barndorff-Nielsen and Shephard (2001)).

- Sample $\sigma^{(r)}$ from

$$p(\sigma|\boldsymbol{y}, \boldsymbol{v}^{(r)}) \sim \mathrm{IG}\left(\frac{3n}{2}, \frac{1}{4}(\boldsymbol{y} - \kappa_1\boldsymbol{v})^T\boldsymbol{V}\boldsymbol{H}_v(\boldsymbol{y} - \kappa_1\boldsymbol{v}) + \frac{2}{\kappa_2}\sum_{i=1}^{n} v_i\right),$$

  where $\mathrm{IG}(\cdot)$ is the inverse gamma distribution, $\boldsymbol{H}_v = \boldsymbol{I}_n - \frac{g}{g+1}\boldsymbol{B}\boldsymbol{\Sigma}_v^{-1}\boldsymbol{B}^T\boldsymbol{V}$.

- Sample $\boldsymbol{\beta}^{(r)}$ from

$$p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{v}^{(r)}, \sigma^{(r)}) \sim \mathrm{N}\left(\frac{g}{g+1}\boldsymbol{\Sigma}_v^{-1}\boldsymbol{B}^T\boldsymbol{V}(\boldsymbol{y} - \kappa_1\boldsymbol{v}), \frac{2\sigma g}{g+1}\boldsymbol{\Sigma}_v^{-1}\right).$$

- Calculate the important weights

$$w^{(r)} = \frac{p(\boldsymbol{\beta}^{(r)}, \sigma^{(r)}, \boldsymbol{v}^{(r)}|\boldsymbol{y})}{p(\boldsymbol{\beta}^{(r)}|\boldsymbol{v}^{(r)}, \sigma^{(r)}, \boldsymbol{y})p(\sigma^{(r)}|\boldsymbol{v}^{(r)}, \boldsymbol{y})p(\boldsymbol{v}^{(r)})},$$

  based on $\boldsymbol{v}^{(r)}$, $\sigma^{(r)}$ and $\boldsymbol{\beta}^{(r)}$. This is to adjust for the GIG approximation of the marginal posterior of $\boldsymbol{v}$ given $\boldsymbol{y}$, which is given by its unnormalised density function

$$\pi(\boldsymbol{v}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{v}|\tilde{\boldsymbol{\beta}}, \tilde{\sigma}, \boldsymbol{y})}{p(\tilde{\boldsymbol{\beta}}|\boldsymbol{y}, \boldsymbol{v}.\tilde{\sigma})p(\tilde{\sigma}|\boldsymbol{y}, \boldsymbol{v})},$$

  where $p(\boldsymbol{v}|\tilde{\boldsymbol{\beta}}, \tilde{\sigma}, \boldsymbol{y})$ is an importance sampling density in the importance sampling algorithm. The importance weights will be used to determine the acceptance probability of each $\{\boldsymbol{\beta}^{(r)}, \sigma^{(r)}, \boldsymbol{v}^{(r)}\}$.

The algorithm iterates until the Gibbs sampling algorithm reaches the final MCMC iteration indexed at R and discards a burn-in period.

Finally, the third-stage is the important re-weighting step. The S samples are drawn from the importance weights without replacement where $S < R$ is the number of importance weighting steps. A random indicator vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_D)^T$ is introduced to the non-linear model

$$\boldsymbol{M}_{\boldsymbol{\gamma}} : \boldsymbol{y} = \boldsymbol{B}_{\boldsymbol{\gamma}}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{B}_{\boldsymbol{\gamma}}$ is the $(n \times D_{\boldsymbol{\gamma}})$ matrix consisting of important predictors and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ of length $D_{\boldsymbol{\gamma}}$ is the non-zero parameter vector. The same prior specification in Equation (2.5) is employed along with a prior on γ_d, $d = 1, \ldots, D$, and a beta prior on π:

$$p(\boldsymbol{\gamma}|\pi) \propto \pi^{\sum_{d=1}^{D}\gamma_d}(1-\pi)^{D-\sum_{d=1}^{D}\gamma_d} \quad \text{and} \quad p(\pi) \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right),$$

where $\pi \in [0, 1]$ is the prior probability of randomly including a predictor in the model. Then π is marginalised out from $p(\boldsymbol{\gamma}|\pi)$ resulting as

$$p(\boldsymbol{\gamma}) \propto \text{Beta}\left(\sum_{d=1}^{D}\gamma_d + \frac{1}{2}, D - \sum_{d=1}^{D}\gamma_d + \frac{1}{2}\right).$$

The marginal likelihood of $\boldsymbol{y}$ under the model $\boldsymbol{M}_{\boldsymbol{\gamma}}$ is then obtained by integrating out $\boldsymbol{\beta}$

and σ resulting as

$$p(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{v}) \sim \mathrm{t}_{2n}\left((1-2\tau)\boldsymbol{v}, \frac{4\sum_{i=1}^{n} v_i}{\sigma\kappa_2}\left(\boldsymbol{V} - \frac{g}{g+1}\boldsymbol{V}\boldsymbol{B}_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_v(\boldsymbol{\gamma})^{-1}\boldsymbol{B}_{\boldsymbol{\gamma}}^T\boldsymbol{V}\right)^{-1}\right),$$

where $t_{2n}(\cdot)$ is the multivariate Student t-distribution with $2n$ degrees of freedom. The posterior probability of $\boldsymbol{M}_{\boldsymbol{\gamma}}$ is therefore given by $p(\boldsymbol{\gamma}|\boldsymbol{y},\boldsymbol{v}) \propto p(\boldsymbol{y}|\boldsymbol{\gamma},\boldsymbol{v})p(\boldsymbol{\gamma})$. Lastly, the independent samples of $\boldsymbol{v}$ from the second-stage algorithm are drawn based on the S samples and the important re-weighting step is iterated until the S samples of $\boldsymbol{\gamma}$ are obtained. Then the posterior inclusion probability is estimated, as follows

$$\hat{p}(\gamma_d = 1|\boldsymbol{y},\boldsymbol{v}) = \frac{1}{\tilde{S}}\sum_{s=1}^{\tilde{S}} \gamma_d^{(s)}, \quad d = 1,\ldots,D,$$

where $\tilde{S}$ is the number of iterations after discarding a burn-in period.

2.3 Data Preparation and Data Analysis

This study is based on the data of the NHANES during 2007-2008. The survey conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention used a complex, stratified, multistage sampling design to select a representative sample of non-institutionalised population in the US civilians to participate in a series of comprehensive health-related interviews and examinations. In total, 12,943 people participated in the NHANES 2007-2008 study.

The study variables included SBP and DBP as the response variables. The BP measurements were taken as follows. After a resting period of 5 minutes in a sitting position and determination of maximal inflation level, three consecutive BP readings were recorded. A fourth reading was recorded if a BP measurement is interrupted or incomplete. All the results were taken in the Mobile Examination Center. The BP measurements are essential for hypertension screening and disease management, since hypertension is an important risk factor for cardiovascular and renal disease. Then in this study, SBP and DBP were selected as response variables where each was averaged over the second and third readings. Predictor variables were BMI, age, ethnicity, gender and marital status.

We initially included 9,762 participants who have completed both BP and body measure examinations in the study. From 9,762 participants, we excluded those who had not underwent examinations. Then amongst the remaining 4,612 participants, we further excluded

those who refused to reveal their marital status. Finally, 4,609 participants were included for analysis in this study.

The NHANES protocols were approved by the National Center for Health Statistics research ethics review boards, and informed consent was obtained from all participants. The research adhered to the tenets of the Declaration of Helsinki.

Both 'quantreg' and 'Brq' R packages were employed to fit the frequentist and Bayesian approaches of the quantile regression model with FPs, respectively. The source R code was provided from the main author of Dao et al. (2022) to fit the Bayesian quantile regression with variable selection and FPs via the three-stage algorithm.

This study considered two quantile models at the 50^{th}, 75^{th} and 95^{th} percentiles. When modelling hypertension, it is preferable to model both median and extremely high values of SBP and DBP, which correspond to the median and upper distributions of SBP and DBP, respectively (Kuhudzai et al. (2022)). The following two quantile models were used for the analysis for the fixed quantile level τ:

$$\begin{aligned}\text{SBP}_i = {} & \text{BMI}_i\beta_1 + \text{BMI}_i^{0.5}\beta_2 + \text{Age}_i\beta_3 + \text{Age}_i^{0.5}\beta_4 + \text{Ethnicity}_i\beta_5 + \text{Gender}_i\beta_6 \\ & + \text{MaritalStatus}_i\beta_7, \\ \text{DBP}_i = {} & \text{BMI}_i\beta_1 + \text{BMI}_i^{0.5}\beta_2 + \text{Age}_i\beta_3 + \text{Age}_i^{0.5}\beta_4 + \text{Ethnicity}_i\beta_5 + \text{Gender}_i\beta_6 \\ & + \text{MaritalStatus}_i\beta_7,\end{aligned}$$

for $i = 1, \ldots, 4609$.

The power of 0.5 was chosen for continuous variables, including BMI and age. The remaining variables were linear because they are categorical. Similar FP models were employed to model BP within the linear regression framework (Dong et al. (2016), Takagi and Umemoto (2013), and Thompson et al. (2009), amongst others).

2.4 Results

In this section, both descriptive and model analyses are provided for the NHANES 2007-2008 dataset using the proposed model. To evaluate the performance of the proposed model, we included two existing methods, including quantile regression and Bayesian quantile regression, with the FP model for a fair comparative analysis. The model comparison is discussed outlining the advantages of the proposed model over these two methods. All the results are

provided in this section through tables and figures for each regression analysis.

2.4.1 Descriptive Analysis

For this analysis, continuous variables were collapsed into categorical variables, including SBP, DBP, BMI and age. According to the guidelines of Whelton et al. (2018), the BP variables were divided into three groups: normal ($<$ 120 mmHg for SBP, $<$ 80 mmHg for DBP), pre-hypertension (120$-$139 mmHg for SBP, 80$-$89 mmHg for DBP) and hypertension ($\geq$ 140 mmHg for SBP, $\geq$ 90 mmHg for DBP). The BMI variable was also divided into six groups: underweight ($<$ 18.5), healthy (18.5$-$24.9), overweight (25$-$29.9), obese (30$-$34.9), very obese (35$-$39.9) and morbidly obese ($\geq$ 40) (Centers for Disease Control and Prevention (2022)).

Tables 2.1-2.2 report SBP and DBP proportions amongst US adults by demographic and lifestyle characteristics, including BMI, age, ethnicity, gender and marital status. The Cramér's V value was used to measure the magnitude of the association between SBP, DBP, socio-demographic characteristics and BMI of the participants. Their values with p-values are also presented in Tables 2.1-2.2 and compared with with guidelines given by Rea and Parker (2014): 0.00 to under 0.10 = very weak association, 0.10 to under 0.20 = weak association, 0.20 to under 0.40 = moderate association and 0.40 and above = strong association.

It is evident from Tables 2.1-2.2 that hypertension was more prevalent in underweight, very obese and morbidly obese participants for both BP measures where the very obese and morbidly obese had the highest prevalence for DBP and SBP measures, respectively. The same trend is observed on the proportions of elevated BP for DBP measure. It is clear that healthy participants had the highest prevalence of normal BP for both BP measures.

Concerning age, the prevalence of both elevated BP and hypertension increased with age, with the 40-49 years age group having the highest proportions for DBP measure and the 50 years and above age group for SBP measure. In regards to ethnicity, the non-Hispanic Black participants had the highest prevalence of hypertension compared to other races for both BP measures.

Tables 2.1-2.2 also show that men had the highest prevalence of both elevated BP and hypertension for both BP measures. Participants who were separated or divorced and those who became widowed had the highest prevalence of hypertension for DBP and SBP measures, respectively.

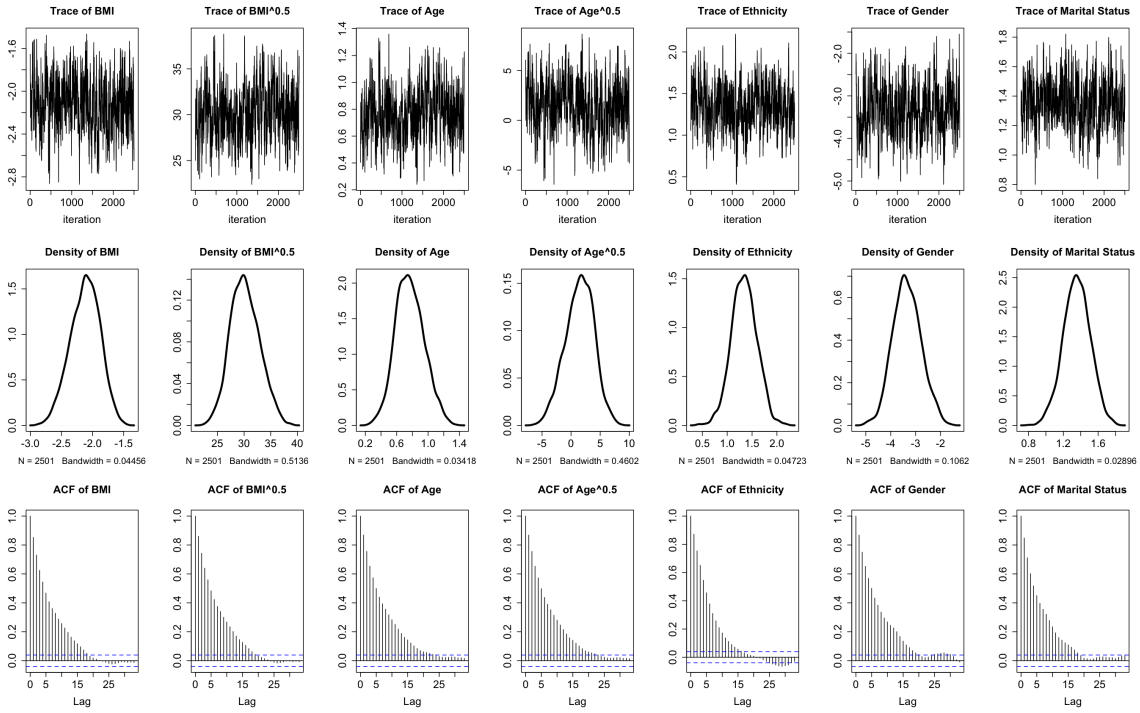
Table 2.1: SBP amongst US adults by BMI and socio-demographic characteristics.

		Normal BP (< 120 mmHg)	Pre-Hypertension (120-139 mmHg)	Hypertension (≥ 140 mmHg)
BMI	Underweight	37 (56.92%)	16 (24.62%)	12 (18.46%)
	Healthy	734 (60.31%)	343 (28.18%)	140 (11.50%)
	Overweight	781 (49.49%)	565 (35.80%)	232 (14.70%)
	Obese	415 (41.71%)	414 (41.61%)	166 (16.68%)
	Very obese	201 (42.68%)	187 (39.70%)	83 (17.62%)
	Morbidly obese	106 (37.46%)	116 (40.99%)	61 (21.55%)
P-value (Cramér's V value)		P-value < 0.01 (0.1106)		
Age	20-29 years	493 (73.36%)	164 (24.40%)	15 (2.23%)
	30-39 years	543 (65.66%)	251 (30.35%)	33 (3.99%)
	40-49 years	460 (55.89%)	285 (34.63%)	78 (9.48%)
	≥50 years	778 (34.02%)	941 (41.15%)	568 (24.84%)
		P-value < 0.01 (0.2535)		
Ethnicity	Mexican American	456 (54.29%)	279 (33.21%)	105 (12.50%)
	Other Hispanic	286 (53.16%)	186 (34.57%)	66 (12.27%)
	Non-Hispanic white	1006 (47.61%)	793 (37.53%)	314 (14.86%)
	Non-Hispanic black	425 (45.31%)	324 (34.54%)	189 (20.15%)
	Other non-Hispanic race	101 (56.11%)	59 (32.78%)	20 (11.11%)
		P-value < 0.01 (0.0665)		
Gender	Male	999 (43.28%)	957 (41.46%)	352 (15.25%)
	Female	1275 (55.41%)	684 (29.73%)	342 (14.86%)
		P-value < 0.01 (0.1310)		
Marital Status	Married	1219 (48.39%)	927 (36.80%)	373 (14.81%)
	Widowed	84 (30.11%)	103 (36.92%)	92 (32.97%)
	Divorced	226 (44.14%)	182 (35.55%)	104 (20.31%)
	Separated	89 (52.05%)	57 (33.33%)	25 (14.62%)
	Never married	468 (58.87%)	256 (32.20%)	71 (8.93%)
	Living with partner	188 (56.46%)	116 (34.83%)	29 (8.71%)
		P-value < 0.01 (0.1251)		

Table 2.2: DBP amongst US adults by BMI and socio-demographic characteristics.

		Normal BP (< 80 mmHg)	Pre-Hypertension (80-89 mmHg)	Hypertension (≥ 90 mmHg)
BMI	Underweight	49 (75.38%)	12 (18.46%)	4 (6.15%)
	Healthy	1025 (84.22%)	148 (12.16%)	44 (3.62%)
	Overweight	1265 (80.16%)	243 (15.40%)	70 (4.44%)
	Obese	772 (77.59%)	168 (16.88%)	55 (5.53%)
	Very obese	356 (75.58%)	78 (16.56%)	37 (7.86%)
	Morbidly obese	217 (76.68%)	47 (16.61%)	19 (6.71%)
P-value (Cramér's V value)		P-value < 0.01 (0.0587)		
Age	20-29 years	619 (92.11%)	47 (6.99%)	6 (0.89%)
	30-39 years	681 (82.35%)	118 (14.27%)	28 (3.39%)
	40-49 years	584 (70.96%)	173 (21.02%)	66 (8.02%)
	≥50 years	1800 (78.71%)	358 (15.65%)	129 (5.64%)
		P-value < 0.01 (0.1118)		
Ethnicity	Mexican American	699 (83.21%)	116 (13.81%)	25 (2.98%)
	Other Hispanic	444 (82.53%)	70 (13.01%)	24 (4.46%)
	Non-Hispanic white	1687 (79.84%)	327 (15.48%)	99 (4.69%)
	Non-Hispanic black	711 (75.80%)	154 (16.42%)	73 (7.78%)
	Other non-Hispanic race	143 (79.44%)	29 (16.11%)	8 (4.44%)
		P-value < 0.01 (0.0569)		
Gender	Male	1732 (75.04%)	423 (18.33%)	153 (6.63%)
	Female	1952 (84.83%)	273 (11.86%)	76 (3.30%)
		P-value < 0.01 (0.1244)		
Marital Status	Married	2017 (80.07%)	385 (15.28%)	117 (4.64%)
	Widowed	231 (82.80%)	38 (13.62%)	10 (3.58%)
	Divorced	386 (75.39%)	87 (16.99%)	39 (7.62%)
	Separated	133 (77.78%)	26 (15.20%)	12 (7.02%)
	Never married	656 (82.52%)	103 (12.96%)	36 (4.53%)
	Living with partner	261 (78.38%)	57 (17.12%)	15 (4.50%)
		P-value = 0.0516 (0.0444)		

Figure 2.1: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.95$ under the Bayesian quantile regression model with FPs.



Lastly, at the 1% significance level, Tables 2.1-2.2 exhibit very weak to weak associations between BP measures, BMI and socio-demographic characteristics amongst US adults. However, there is a moderate association between SBP measure and age. There is no statistically significant association between DBP measure and marital status at the 5% level.

2.4.2 Model Analysis

Tables 2.3-2.4 provide the coefficients for predictors relating to SBP and DBP responses for three quantile regression models with FPs at three quantile levels ($\tau = 0.5, 0.75, 0.95$), including one frequentist and two Bayesian approaches with one using variable selection. For Bayesian approaches, parameters were obtained via posterior mean. The 95% confidence intervals were also obtained for the frequentist approach, whilst the 95% credible intervals were obtained for the Bayesian approaches. A confidence interval describes a probability, for instance, if a user constructs a confidence interval with some confidence level then they are confident that an estimate would fall within the interval. On the other hand, a credible interval is an interval in the domain of a posterior probability distribution where an unobserved parameter value falls with a particular probability. We denote the frequentist approach as the QR-FP model, and two Bayesian approaches as the BQR-FP and BQRVS-FP models

Figure 2.2: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.95$ under the Bayesian quantile regression model with FPs.

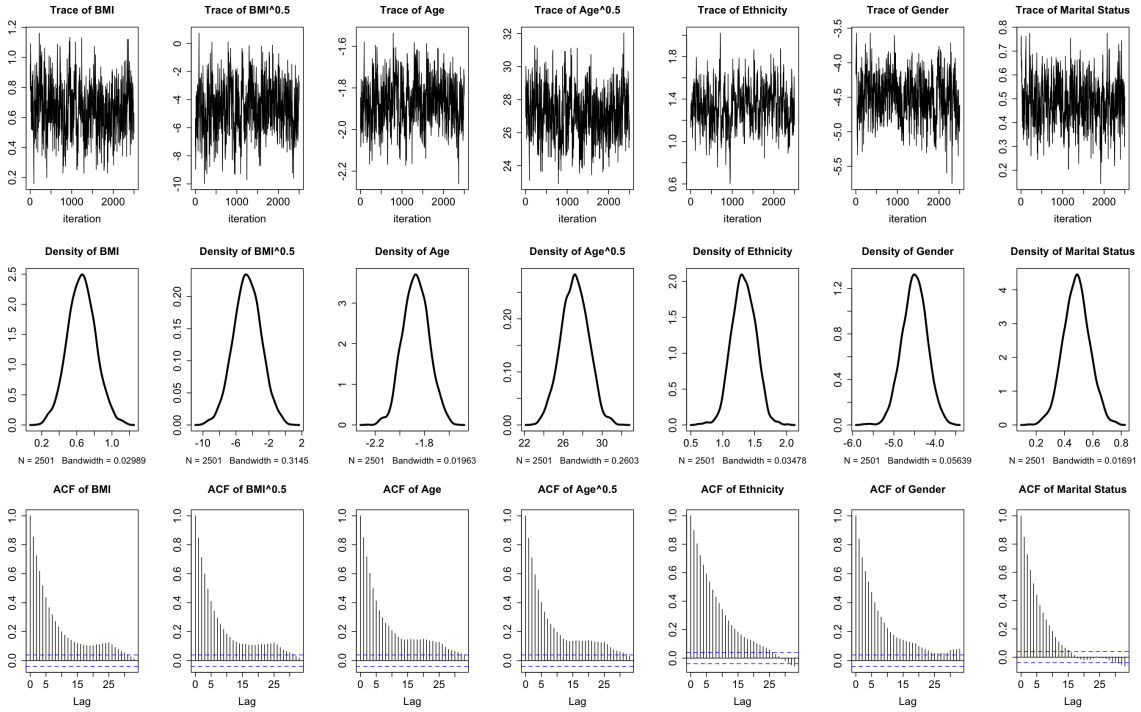


Figure 2.3: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.95$ under the Bayesian quantile regression model with FPs and variable selection.

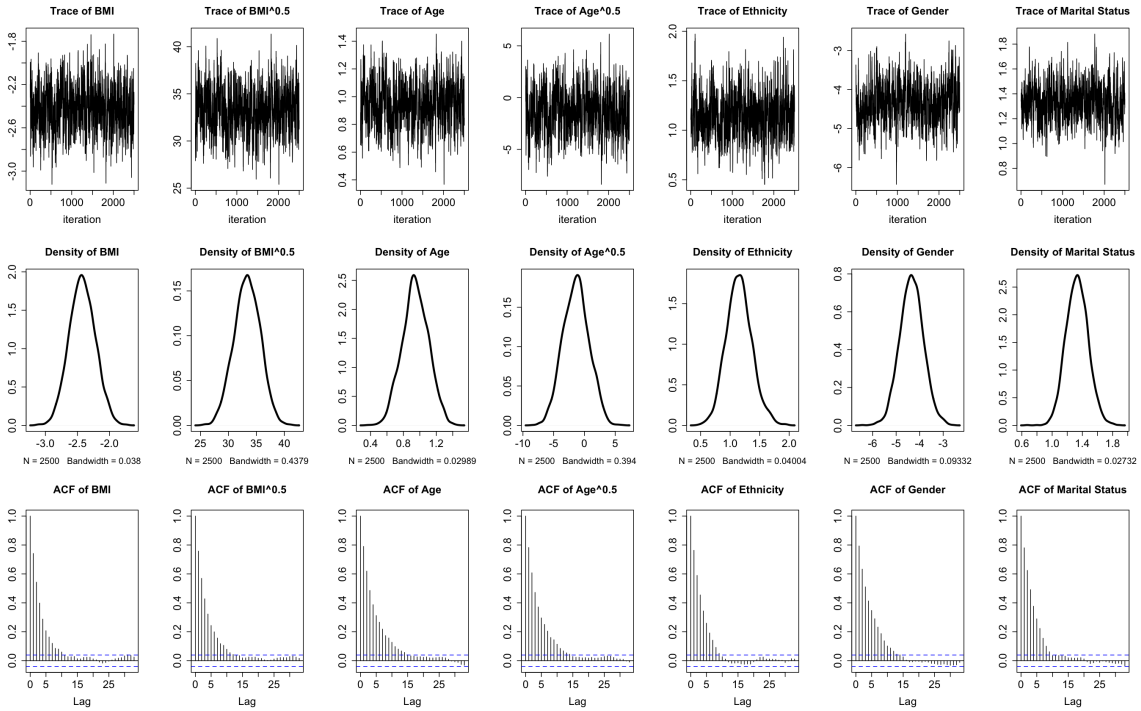


Table 2.3: One frequentist and two Bayesian quantile regression analyses for the relationship between SBP and risk factors at different quantile levels ($\tau = 0.5, 0.75, 0.95$).

Quantile Regression			
τ	0.5	0.75	0.95
BMI	-2.856 (-3.278, -2.280)	-2.198 (-3.040, -1.715)	-2.024 (-3.141, -0.798)
$BMI^{0.5}$	36.085 (29.932, 40.529)	29.210 (23.907, 38.130)	29.113 (15.239, 42.302)
Age	0.510 (0.130, 0.785)	0.317 (-0.003, 0.885)	0.710 (-0.220, 1.630)
$Age^{0.5}$	-1.758 (-5.430, 3.339)	3.297 (-4.116, 7.654)	2.300 (-9.906, 14.672)
Ethnicity	0.626 (0.154, 1.040)	0.995 (0.366, 1.495)	1.214 (0.199, 2.642)
Gender	-4.323 (-5.302, -3.512)	-3.813 (-5.231, -2.506)	-3.278 (-6.147, -0.762)
Marital Status	0.894 (0.612, 1.155)	1.327 (0.916, 1.746)	1.400 (0.650, 2.037)
Bayesian Quantile Regression			
τ	0.5	0.75	0.95
BMI	-2.818 (-3.208, -2.447)	-2.255 (-2.669, -1.889)	-2.120 (-2.603, -1.685)
$BMI^{0.5}$	35.628 (31.653, 39.794)	29.825 (25.763, 34.419)	30.191 (25.146, 35.809)
Age	0.484 (0.233, 0.734)	0.364 (0.103, 0.664)	0.768 (0.428, 1.142)
$Age^{0.5}$	-1.366 (-4.737, 2.002)	2.735 (-1.237, 6.249)	1.446 (-3.550, 6.077)
Ethnicity	0.640 (0.288, 0.979)	0.957 (0.561, 1.359)	1.341 (0.839, 1.829)
Gender	-4.376 (-5.138, -3.645)	-3.809 (-4.784, -2.823)	-3.346 (-4.397, -2.190)
Marital Status	0.888 (0.656, 1.125)	1.347 (1.055, 1.637)	1.354 (1.041, 1.649)
Bayesian Quantile Regression FP & Variable Selection			
τ	0.5	0.75	0.95
BMI	-2.812 (-3.164, -2.468)	-2.581 (-2.974, -2.168)	-2.426 (-2.813, -2.027)
$BMI^{0.5}$	35.547 (31.789, 39.269)	33.335 (28.817, 37.747)	33.335 (28.815, 37.784)
Age	0.459 (0.226, 0.680)	0.537 (0.274, 0.806)	0.945 (0.643, 1.256)
$Age^{0.5}$	-1.129 (-4.197, 2.029)	-0.051 (-3.717, 3.536)	-1.382 (-5.473, 2.680)
Ethnicity	0.571 (0.258, 0.898)	0.843 (0.484, 1.212)	1.152 (0.753, 1.616)
Gender	-4.577 (-5.300, -3.899)	-4.291 (-5.053, -3.518)	-4.343 (-5.301, -3.351)
Marital Status	0.828 (0.632, 1.033)	1.139 (0.893, 1.381)	1.331 (1.052, 1.617)

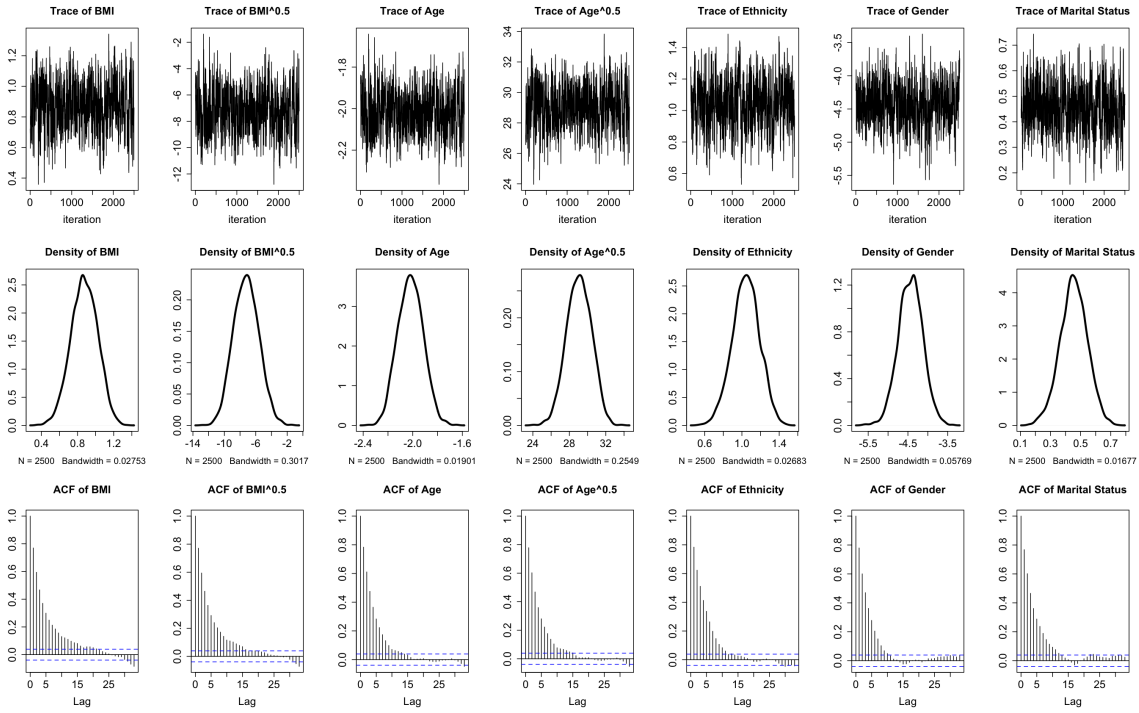
Table 2.4: One frequentist and two Bayesian quantile regression analyses for the relationship between DBP and risk factors at different quantile levels ($\tau = 0.5, 0.75, 0.95$).

Quantile Regression			
τ	0.5	0.75	0.95
BMI	1.174 (0.705, 1.496)	0.761 (0.507, 1.096)	0.582 (0.022, 1.572)
$BMI^{0.5}$	-12.200 (-15.675, -7.071)	-7.179 (-10.821, -4.242)	-3.995 (-13.869, 2.247)
Age	-2.266 (-2.477, -1.979)	-2.018 (-2.252, -1.832)	-1.852 (-2.418, -1.418)
$Age^{0.5}$	31.329 (27.308, 34.170)	28.298 (25.758, 31.451)	26.918 (21.199, 34.557)
Ethnicity	0.561 (0.203, 0.841)	0.712 (0.411, 1.030)	1.264 (0.345, 2.013)
Gender	-3.345 (-4.160, -2.651)	-3.619 (-4.337, -2.976)	-4.592 (-5.769, -3.047)
Marital Status	0.210 (-0.041, 0.448)	0.368 (0.171, 0.549)	0.466 (0.143, 0.934)
Bayesian Quantile Regression			
τ	0.5	0.75	0.95
BMI	1.153 (0.836, 1.433)	0.798 (0.539, 1.056)	0.656 (0.345, 0.974)
$BMI^{0.5}$	-11.923 (-15.007, -8.505)	-7.554 (-10.406, -4.748)	-4.624 (-7.981, -1.332)
Age	-2.253 (-2.431, -2.058)	-2.040 (-2.224, -1.863)	-1.870 (-2.064, -1.663)
$Age^{0.5}$	31.131 (28.434, 33.566)	28.594 (26.243, 31.077)	27.176 (24.467, 29.773)
Ethnicity	0.536 (0.291, 0.777)	0.706 (0.455, 0.966)	1.328 (0.981, 1.667)
Gender	-3.391 (-3.999, -2.778)	-3.635 (-4.169, -3.109)	-4.498 (-5.086, -3.924)
Marital Status	0.220 (0.030, 0.408)	0.374 (0.222, 0.533)	0.484 (0.304, 0.667)
Bayesian Quantile Regression FP & Variable Selection			
τ	0.5	0.75	0.95
BMI	1.101 (0.823, 1.381)	0.808 (0.568, 1.041)	0.874 (0.584, 1.147)
$BMI^{0.5}$	-11.299 (-14.374, -8.289)	-7.620 (-10.207, -4.940)	-7.217 (-10.158, -4.080)
Age	-2.217 (-2.397, -2.033)	-2.031 (-2.203, -1.867)	-2.018 (-2.206, -1.821)
$Age^{0.5}$	30.603 (28.089, 33.030)	28.381 (26.127, 30.639)	29.063 (26.415, 31.577)
Ethnicity	0.505 (0.278, 0.727)	0.630 (0.391, 0.868)	1.043 (0.747, 1.319)
Gender	-3.401 (-3.934, -2.888)	-3.733 (-4.219, -3.233)	-4.436 (-5.032, -3.827)
Marital Status	0.193 (0.033, 0.347)	0.371 (0.222, 0.523)	0.454 (0.270, 0.628)

Table 2.5: Selected predictors for both SBP and DBP models under the Bayesian quantile regression model with FPs and variable selection at different quantile levels ($\tau = 0.5, 0.75, 0.95$).

	Model	BMI	$BMI^{0.5}$	Age	$Age^{0.5}$	Ethnicity	Gender	MaritalStatus
$\tau = 0.5$	SBP	1.0000	1.0000	0.9920	0.3062	0.9733	1.0000	1.0000
	DBP	1.0000	1.0000	1.0000	1.0000	0.9973	1.0000	0.8628
$\tau = 0.75$	SBP	1.0000	1.0000	0.9920	0.2423	1.0000	1.0000	1.0000
	DBP	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\tau = 0.95$	SBP	1.0000	1.0000	1.0000	0.4034	1.0000	1.0000	1.0000
	DBP	1.0000	0.9986	1.0000	1.0000	1.0000	1.0000	0.9986

Figure 2.4: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.95$ under the Bayesian quantile regression model with FPs and variable selection.



where the latter uses variable selection.

For the BQR-FP model, the algorithm was implemented for 10,000 MCMC iterations and 1,000 MCMC iterations were discarded as a burn-in. For the BQRVS-FP model, the first-stage algorithm ran for 1,000 EM iterations and repeated for 2 replications. Then 5,000 MCMC iterations were drawn for the second-stage algorithm, whilst discarding 2,500 MCMC iterations as a burn-in. Finally, the last algorithm ran for 1,250 important re-weighting steps of which 500 steps were discarded as a burn-in. The value of g was selected as 1,000 for all implementations of the variable selection model.

It is evident from Table 2.3 that all the risk factors except both linear and non-linear terms of age were found to have statistically significant associations with SBP across the two upper quantile levels according to their 95% confidence intervals containing no zero value under the QR-FP model. Looking at the median level, the linear term had association with SBP under the same approach. When looking at the BQR-FP and BQRVS-FP models, only the non-linear term of age did not have a statistically significant association for all quantile levels. On the other hand, Table 2.4 observes that all the risk factors including non-linear terms had statistically significant associations with DBP across all quantile levels for all model approaches. Still, when looking at the median level under the QR-FP model,

it revealed that the marital status did not have statistically significant association.

Table 2.3 also observes that the BMI, non-linear term of age and gender have negative associations with SBP, whilst the non-linear term of BMI, age and gender have negative associations with DBP from Table 2.4 for all three model approaches. Under the SBP model, the coefficients of BMI, ethnicity, gender and marital status increased when the quantile levels increased. The same trend is observed for the coefficients of BMI's non-linear term, age, ethnicity and marital status under the DBP model. Observing the coefficient of age's non-linear term, all models saw the reverse U-shaped trend under the SBP model and on the other hand, both QR-FP and BQR-FP models had decreasing trends and the BQRVS-FP had the U-shaped trend under the DBP model. Interestingly, the coefficient of BMI's non-linear term under the SBP model followed the decreasing trend for the QR-FP model, the U-shaped trend for the BQR-FP model and the square-root trend for the BQRVS-FP model.

Convergence of both Bayesian approaches was assessed using the trace plots, the density plots and autocorrelation plots. This is essential to perform various diagnostic tools for assessing the convergence (Sinharay (2003)). The convergence diagnostics are useful to check stationarity of the Markov chain or good chain mixing and to verify the accuracy of the posterior estimates (Lesaffre and Lawson (2012)). The trace plot is in the form of a time series plot indicating whether it reaches stationarity or not. The density plot represents the stationary distribution of posterior samples approximating the posterior distribution of interest. The autocorrelation plot reports the correlation of posterior samples at each chain step with previous estimates of the same variable, lagged by number of iterations. A decreasing trend indicates that the stationary distribution is more random and less dependent on initial values in the chain (Hamra et al. (2013)).

Figures 2.1-2.2 present the trace, density and autocorrelation plots for each risk factor of SBP and DBP, respectively under the BQR-FP model at $\tau = 0.95$. When looking at the trace plots, they exhibit stationarity due to relatively constant mean and variance of each plot. Thus, they show the good Markov chain mixing rate. When looking at the density plots, they reflect a smooth distribution with one peak at the mode of the distribution indicating a good convergence. It is also shown from the figures that each risk factor of SBP and DBP has increasingly random stationary posterior distribution, yet the trend has a slower decreasing rate. Similar performances are observed at two other quantile levels ($\tau = 0.5, 0.75$) and their figures are provided in Appendix D.1 (see Figures D.1-D.2 and D.5-D.6).

Figures 2.3-2.4 also present the trace, density and autocorrelation plots for each risk factor of SBP and DBP, respectively under the BQRVS-FP model at $\tau = 0.95$. All the plots show stationarity, good Markov chain mixing rates and good convergence. Particularly, each autocorrelation plot indicates that their stationary distribution became random and less correlated with the initial values at a faster rate. Similar performances are observed at two other quantile levels ($\tau = 0.5, 0.75$) and their figures are provided in Appendix D.1 (see Figures D.3-D.4 and D.7-D.8).

Table 2.5 provides marginal inclusion probabilities (MIPs) that determine which risk factors are influential on SBP and DBP for the BQRVS-FP model at three quantile levels. The risk factors that lie above the threshold of 0.9 of MIP are selected as important predictors. Across all the quantile levels for both SBP and DBP models, all the important risk factors were consistently selected including the non-linear terms. There are two cases of non-important risk factors where, unlike the SBP model, the DBP model did not select marital status at the median level and the SBP model did not select the non-linear term of age at all the quantile levels. This mostly agreed with findings on 95% credible intervals from Tables 2.3-2.4.

2.4.3 Model Comparison

Hespanhol et al. (2019) advised that the estimation effect must always be within the confidence/credible interval, and the width of the interval represents the precision of the estimation effect. Hence, the narrower the interval the more precise is the estimate. Observing at the 95% confidence intervals of frequentist approach and the 95% credible intervals of two Bayesian approaches from Tables 2.3-2.4, the BQRVS-FP model has tighter intervals compared to the QR-FP model having wider intervals.

Another finding is from the diagnostic plots that the autocorrelation plots of BQRVS-FP model have a faster decreasing rate across all the quantile levels, whereas those of the BQR-FP model have a slower rate. This is evident that the BQRVS-FP model has more random stationary posterior distributions of interest.

When looking at Tables 2.3-2.5, the BQRVS-FP model selected the important predictors coinciding with statistically significant associations between SBP, DBP and their risk factors based on their 95% credible intervals.

These findings suggest that the Bayesian variable selection approach to quantile regression model with FPs obtained more precise estimates than the frequentist and unregularised

Bayesian approaches. The non-linear terms were selected as important variables in both SBP and DBP models indicating that FP model was necessary to examine the non-linear relationship between SBP, DBP and risk factors.

Whilst computational performance was not evaluated in this paper, it is noteworthy that all computations were executed on R version 4.2.2, utilising an Intel Core i7-4790 CPU@3.6GHz machine with 16GB DDR3 RAM memory. Both Rcpp and the Intel MKL compiler were employed to enhance the efficiency of the proposed method and reduce running time. The proposed method follows a three-stage algorithm, which, admittedly, demands more computational time compared to the unregularised Bayesian method that relies solely on a Gibbs sampling algorithm. Nevertheless, as previously mentioned, the second-stage algorithm of the proposed method, namely the Gibbs sampling algorithm, exhibits a faster convergence rate. Consequently, it necessitates fewer iterations to run compared to the unregularised Bayesian method. The first and last algorithms of the proposed method, requiring fewer iterations, contribute to a reasonable overall computational performance. It is crucial to note that, with an increasing amount of data, computational challenges may arise, potentially necessitating a big data strategy to address these issues. However, it is important to acknowledge that addressing these challenges should be left for future research work.

2.5 Chapter Summary

In this chapter, we conducted the data analysis of the impact of BMI on the BP measures, including SBP and DBP using data extracted from the 2007-2008 NHANES database. The descriptive analysis showed that the prevalence of hypertension increased by age and the hypertension was highly prevalent amongst very obese and morbidly obese participants. In particular, it was more prevalent in men than women. Moreover, there was a statistically significant moderate association between SBP and age based on the Cramér's V value, whilst the remaining associations were weaker for both BP measures. However, there was no association between DBP and marital status.

The analysis motivated a new Bayesian non-linear quantile regression under the FP model and variable selection with quantile-dependent prior. The quantile regression analysis investigates how the relationships differ across the median and upper quantile levels. The use of FPs allows for the relationships to be non-linear parametrically. The variable selection investigates for important predictors that contribute to the non-linear relationships via the Bayesian paradigm. The model analysis suggested that the proposed model provides better

estimates because the 95% credible intervals were narrower and the autocorrelation plots have faster decreasing rates of correlated posterior samples in comparison to two methods, the frequentist and Bayesian approaches of quantile regression model. The analysis of the data showed that non-linear relations do exist because the proposed model identified the non-linear terms of continuous variables, including BMI and age as important predictors in the model across all the quantile levels. On the other hand, the non-linear term of age was not selected under the SBP model. The marital status was not selected as an important risk factor for the DBP model at the median level. This agreed with findings of both descriptive and model analyses. Moreover, the data analysis suggested that the quantile-based FP approaches have the goodness of fit in comparison to mean-based FP approaches. Thus, the importance of the non-linear quantile model with FPs is significant for modelling of BP measures.

Chapter 3

Bayesian Huberised Regularisation and Beyond

Along the frequentist line, robust regression is widely used and has an ability of taking asymmetricity into account simultaneously and for high-dimensional analysis. However, the majority of research is not capable of full probabilistic uncertainty quantification. This motivates the development of a new Bayesian Huberised regularisation, including Bayesian Huberised Lasso (Kawakami and Hashimoto (2023)) and Bayesian Huberised Elastic Net. This chapter first introduces a new asymmetric Huberised loss function and its corresponding probability distribution, which has the scale mixture of normal representation. A by-product of Bayesian Huberised regularisation and asymmetric Huberised loss function results in a new Bayesian Huberised regularised quantile regression. We also studied the theoretical properties of the proposed methods, including posterior propriety and unimodality of their joint posterior density. All the proofs can be found in Appendix A. Through a wide variety of simulation studies and real data examples, the proposed methods showed promising results in terms of robustness and applications.

3.1 Introduction

Robust regression methods have a wide range of applications and have attracted a great amount of attention in the literature recently, particularly for taking asymmetricity into account simultaneously and for high-dimensional analysis, such as the adaptive Huber regression (Sun et al. (2020)) and asymmetric Huber loss and asymmetric Tukey's biweight loss functions for robust regression (Fu and Wang (2021)). The Lasso (Tibshirani (1996))

and the Elastic Net (Zou and Hastie (2005)) are some popular choices for regularising regression coefficients. The former has the ability to automatically set irrelevant coefficients to zero. The latter retains this property and the effectiveness of the ridge penalty, and it deals with highly correlated variables more effectively. Robust regularisation methods for quantile regression provide a promising technique for variable selection and model estimation in the presence of outliers or heavy-tailed errors (Li and Zhu (2008), Wu and Liu (2009), Belloni and Chernozhukov (2011), and Su and Wang (2021), amongst others). However, the majority of research on the topics falls in frequentist approaches, which are not capable of full probabilistic uncertainty quantification.

Exploring unconditional Bayesian regularisation prior, such as the Bayesian Lasso (Park and Casella (2008)) and the Bayesian Elastic Net (Li and Lin (2010)), for robust regression is not straightforward. Several issues may arise. The joint posterior may be multi-modal, which slows down the convergence of the Gibbs sampling algorithm and the point estimates may be computed through multiple modes, which lead to the inaccurate estimators (Park and Casella (2008), and Kyung et al. (2010)). The choices of hyper-parameters in gamma priors of regularisation parameters may also have strong influences on the posterior estimates. For the former, it was firstly observed by Park and Casella (2008) in the Bayesian Lasso. For the latter, it is common to employ invariant prior on scale parameter (Berger (1985)). Cai and Sun (2021) addressed these two issues by introducing the scale parameter to the Bayesian Lasso and its generalisation for quantile regression. Moreover, Kawakami and Hashimoto (2023) used the scale parameter of the hyperbolic loss function (Park and Casella (2008)) to propose the Bayesian Huberised Lasso, which is the robust version of Bayesian Lasso. Along this line, we will propose Bayesian Huberised regularisation in this chapter.

As discussed in Chapter 2, the inclusion of Bayesian modelling and variable selection in quantile regression has proven to be invaluable, owing to its exact inference, parameter uncertainty, robustness and asymmetry properties, good estimation of effects, and its ability to derive unique insights into comprehensive relationships. It is briefly mentioned in Chapter 1 that the error distribution for Bayesian quantile regression is usually assumed to follow the ALD that guaranteed posterior consistency of Bayesian estimators (Sriram et al. (2013)) and robustness (Yu and Moyeed (2001)). For regularisation, Alhamzawi et al. (2012) adopted the inverse gamma prior density to the penalty parameters and treated its hyper-parameters as unknown and estimated them along with other parameters. This allows different regression coefficients to have different penalisation parameters, which improve the predictive accuracy. Particularly, Bayesian quantile regression both on its own and in a high-dimensional setting

enjoys some robustness, such as median being more robust than mean, yet has different modelling aims from robust regression.

Delving into our proposed methodology, the approximate Gibbs sampler of Kawakami and Hashimoto (2023) can be regarded as empirical Bayesian (EB) in nature, since it estimates the tuning robustness parameter, or a hyper-parameter, directly. The idea of this sampling algorithm originated from the paper of Miller (2019) that is similar to the EB Gibbs sampling method of Casella (2001), yet is different in terms of computational algorithms for estimating hyper-parameters. Generally, Bayesian inference based on data-dependent priors is not new, and can be traced back to, at least, Berger (1984), and the history of the EB method is detailed in the paper of Casella (1985). This method has shown to be both computationally feasible and statistically valid (Morris (1983), and Casella (2001)). It is argued that if the alternative method, such as adopting a flat prior for a hyper-parameter, is employed then the usage of a flat prior may lead to models with improper posterior distributions that can be problematic for Bayesian analysis (Natarajan and McCulloch (1995), and Hobert and Casella (1996)). Precisely, Kawakami and Hashimoto (2023) stated that if the random-walk Metropolis-Hastings algorithm is utilised for estimating the tuning robustness parameter, it can result in instability due to the need of evaluation for the modified Bessel function of the second kind. Consequently, the EB approach is suitable for our proposed methodology in practice.

Therefore, this chapter first proposes a new Huberised-type of asymmetric loss function and its corresponding probability distribution, which is shown to have the scale mixture of normals. Then we introduce a new Bayesian Huberised regularisation for robust regression. Furthermore, by taking advantage of the good quantile property of this probability distribution, we develop Bayesian Huberised Lasso quantile regression and Bayesian Huberised Elastic Net quantile regression. This results in the proposed models covering both Bayesian robust regularisation and Bayesian quantile regularisation. Besides, Cai and Sun (2021) emphasised that the posterior impropriety does exist in Bayesian Lasso quantile regression and its generalisation when the prior on regression coefficients is independent of the scale parameter. Thus, we will discuss some properties of the Bayesian Huberised regularised quantile regression, including posterior propriety and posterior unimodality. The approximate Gibbs sampler of Kawakami and Hashimoto (2023) is adopted to enable the data-dependent estimation of the tuning robustness parameter in the fully Bayesian hierarchical model. Due to the fact that it is being EB in nature, the advantage of this sampling step is that it does not require cross validation evaluation of tuning parameters (see Alhamzawi (2016)

for example) nor the rejection steps, such as the inversion method and adaptive rejection sampling algorithm (see Alhamzawi et al. (2019) for example). We demonstrate the effectiveness and robustness of the Bayesian Huberised regularised quantile regression model through simulation studies followed by real data analysis.

Section 3.2 defines the asymmetric Huberised loss function with its corresponding probability density function and derives the scale mixture of normal representation for Bayesian inference. Section 3.3 presents the Bayesian Huberised regularisation including the Bayesian Huberised Lasso (Kawakami and Hashimoto (2023)) and the Bayesian Huberised Elastic Net. This results in a new robust Bayesian regularised quantile regression. In Section 3.4 and 3.5, a wide range of simulation studies and three real data examples are conducted. In Section 3.6, we draw the conclusions.

3.2 Asymmetric Huberised Loss Function

When the error distribution is asymmetric or contaminated by asymmetric outliers, the estimators obtained from Equations (1.1)-(1.4) may result in inconsistency of predictions of a conditional mean given the predictors (Fu and Wang (2021)).

Therefore, we propose the Huberised-type asymmetric loss function, that is defined as

$$L^{\text{Asy}}_{\eta,\rho^2,\tau}(x) = \sqrt{\eta\left(\eta + \frac{x}{\rho^2}\left(\tau - \mathrm{I}(x<0)\right)\right)} - \eta. \tag{3.1}$$

By letting $\eta = \sqrt{\zeta_1\zeta_2}$, $\rho^2 = \sqrt{\zeta_2/\zeta_1}$ and $\tau = 0.5$, Equation (3.1) becomes the Non-convex Huber loss function (Equation (1.3)) as a symmetric case.

The corresponding density function is

$$f(x|\mu,\eta,\rho^2,\tau) = \frac{\eta\tau(1-\tau)e^{\eta}}{2\rho^2(\eta+1)} \exp\left\{-\sqrt{\eta\left(\eta + \frac{x-\mu}{\rho^2}\left(\tau - \mathrm{I}(x<0)\right)\right)}\right\}, \tag{3.2}$$

where $\mu \in \mathbb{R}$ is a location parameter. Here, ρ^2 acts as a scale parameter and η acts as a shape parameter of this density function.

The following proposition states that the parameters μ and τ in Equation (3.2) satisfy: μ is the τth quantile of the distribution.

Proposition 3.1. *If a random variable X follows the density function in Equation (3.2) then we have $P(X \leq \mu) = \tau$ and $P(X > \mu) = 1 - \tau$.*

Proof. The proof can be found under Proposition A.1 in Appendix A. □

To observe the behaviour of the proposed loss function, we set $\eta = \sqrt{\zeta_2}\left(\sqrt{\zeta_2} + \sqrt{\zeta_2 + 1}\right)$ and $\rho^2 = \sqrt{\zeta_2}/(\sqrt{\zeta_2} + \sqrt{\zeta_2 + 1})$ then we have the following limits,

$$\lim_{\zeta_2 \to 0} L^{\text{Asy}}_{\eta,\rho^2,\tau}(x) = \sqrt{x\left(\tau - \mathrm{I}(x < 0)\right)},$$
$$\lim_{\zeta_2 \to \infty} L^{\text{Asy}}_{\eta,\rho^2,\tau}(x) = x\left(\tau - \mathrm{I}(x < 0)\right),$$

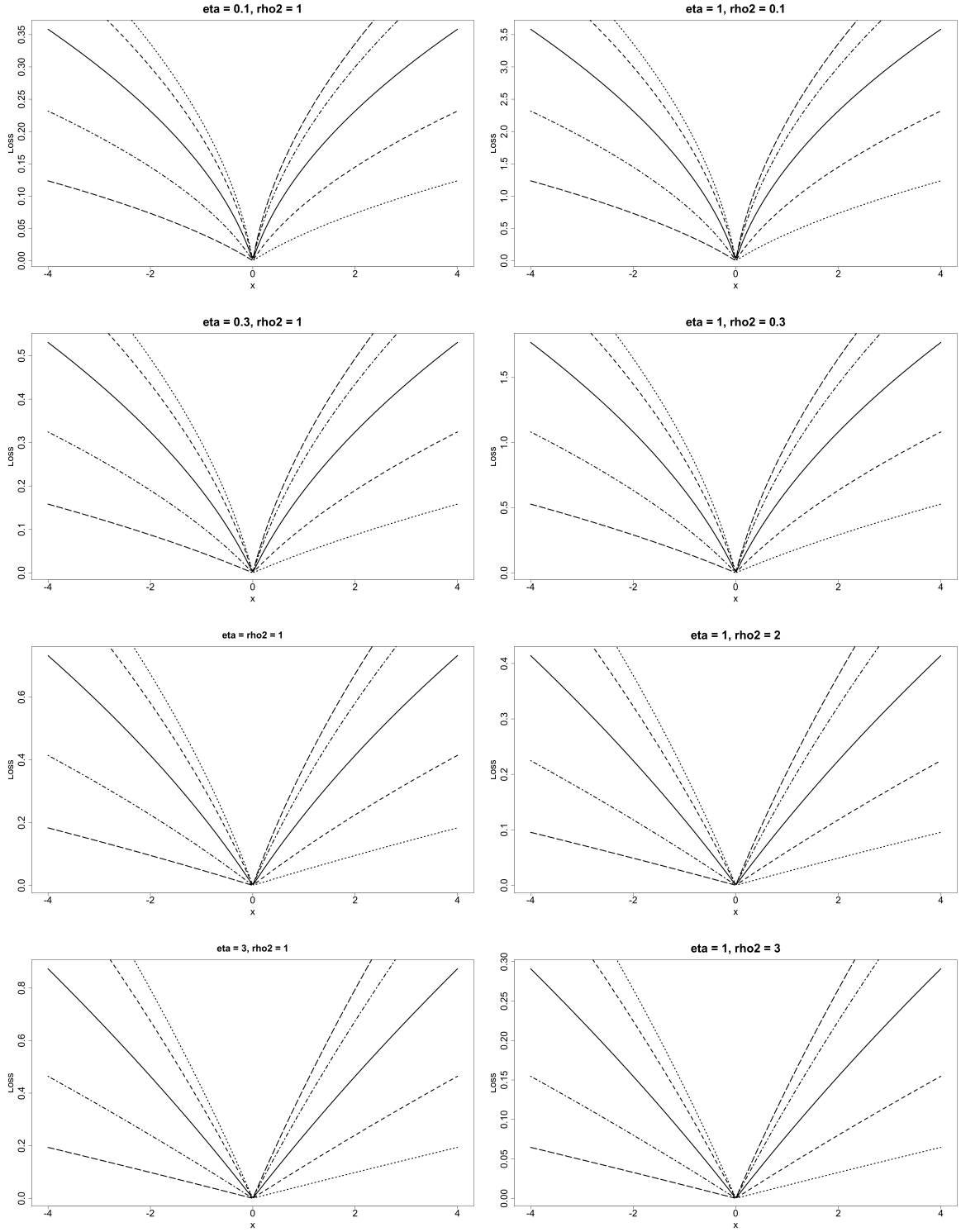
which suggest that the proposed loss bridges the quantile loss function. Daouia et al. (2018) used the quantile loss function for tail expectiles to estimate alternative measures to the value at risk and marginal expected shortfall, which are two instruments of risk protection of utmost importance in actuarial science and statistical finance. Ehm et al. (2016) showed that any scoring function that is consistent for a quantile or an expectile functional can be represented as a mixture of elementary or extremal scoring functions that form a linearly parameterised family. However, in this chapter, we show a totally new way to achieve it, and our proposed loss is a novel representative of asymmetric least squares (Daouia et al. (2019)). Figure 3.1 illustrates the asymmetric shape behaviour for five different values of τ $(0.1, 0.25, 0.5, 0.75, 0.9)$. From the figure, $L^{\text{Asy}}_{\eta,\rho^2,\tau}(x)$ approaches the square root of the quantile loss function, as $\eta \to 0$, and $L^{\text{Asy}}_{\eta,\rho^2,\tau}(x)$ approaches the quantile loss function, as $\eta \to \infty$.

Kawakami and Hashimoto (2023) discussed that it is essential to choose the right value of hyper-parameters of η and ρ^2 where ρ^2 can easily be estimated by a Gibbs sampling algorithm in a Bayesian model whereas the estimation of η is difficult. They proposed the approximate Gibbs sampler to enable the data-dependent estimation of η. This chapter will also adopt their approximate Gibbs sampler.

To fully enable the Gibbs sampling algorithm for Bayesian modelling, we show in Appendix A that the density function in Equation (3.2) has the scale mixture of normal representation with exponential and GIG densities, which is stated in the following theorem.

Theorem 3.1. *If the model error $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}$ follows the density function (Equation*

Figure 3.1: Asymmetrical behaviour of the proposed loss function for τ=0.1 (short dashed), 0.25 (normal dashed), 0.5 (solid), 0.75 (short-normal dashed), and 0.9 (long dashed) for different values of η and ρ^2.



(3.2)), then we can represent ϵ_i as the scale mixture of normals given by

$$f(\epsilon_i;\tau,\eta,\rho^2) \propto \iint N(\epsilon_i;(1-2\tau)v_i,4v_i\sigma_i)\, Exp\left(v_i;\frac{\tau(1-\tau)}{2\sigma_i}\right) GIG\left(\sigma_i;\frac{3}{2},\frac{\eta}{\rho^2},\eta\rho^2\right) dv_i d\sigma_i,$$
$$i=1,\ldots,n, \tag{3.3}$$

where GIG$(\cdot)$ denotes the GIG distribution and its density is specified by Equation (2.6), Exp$(\cdot)$ denotes the exponential distribution, and N$(\cdot)$ is the normal distribution.

Proof. The proof can be found under Theorem A.1 in Appendix A. □

3.3 Bayesian Huberised Regularised Quantile Regression Model

In this section, we consider a Bayesian modelling of Huberised regularised quantile regression.

3.3.1 Bayesian Huberised Lasso Quantile Regression

Kawakami and Hashimoto (2023) showed that the unconditional Laplace prior of $\boldsymbol{\beta}$ (Park and Casella (2008)) would lead to multi-modality of a posterior density and resolved this issue by introducing ρ^2 as a scale parameter to formulate the Bayesian Huberised Lasso, that is,

$$\pi(\boldsymbol{\beta}|\rho^2,\lambda_1)=\prod_{j=1}^{k}\frac{\lambda_1}{2\sqrt{\rho^2}}\exp\left\{-\frac{\lambda_1|\boldsymbol{\beta}_j|}{\sqrt{\rho^2}}\right\}. \tag{3.4}$$

By using the scale mixture of normal representation of Laplace distribution (Andrews and Mallows (1974)), the Bayesian Huberised Lasso can be expressed as

$$\boldsymbol{\beta}|\boldsymbol{s},\rho^2\sim \mathrm{N}(\mathbf{0},\rho^2\boldsymbol{\Lambda}),\quad s_j|\lambda_1\sim \mathrm{Exp}\left(\frac{\lambda_1^2}{2}\right),\quad j=1,\ldots,k,$$

where $\boldsymbol{s}=(s_1,\ldots,s_k)^T$ and $\boldsymbol{\Lambda}=\mathrm{diag}\,(s_1,\ldots,s_k)$.

Therefore, with the Bayesian Huberised Lasso, we present the following hierarchical model

using the scale mixture of normal representation in Theorem 3.1:

$$
\begin{aligned}
&\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{v} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+(1-2\tau)\boldsymbol{v},\boldsymbol{V}),\\
&\sigma_i|\rho^2,\eta \sim \mathrm{GIG}\left(\frac{3}{2},\frac{\eta}{\rho^2},\eta\rho^2\right),\ v_i|\sigma_i \sim \mathrm{Exp}\left(\frac{\tau(1-\tau)}{2\sigma_i}\right),\ i=1,\ldots,n,\\
&\beta_j|s_j,\rho^2 \sim \mathrm{N}(0,\rho^2 s_j),\ s_j|\lambda_1^2 \sim \mathrm{Exp}\left(\frac{\lambda_1^2}{2}\right),\ j=1,\ldots,k,\\
&\rho^2 \sim \pi(\rho^2) \propto \frac{1}{\rho^2},\ \lambda_1^2 \sim \mathrm{Gamma}(a,b),\ \eta \sim \mathrm{Gamma}(c,d),
\end{aligned}
$$

where $\boldsymbol{V}=\mathrm{diag}(4\sigma_1 v_1,\ldots,4\sigma_n v_n)$. As a prior of ρ^2, we assume the improper scale invariant prior, that is proportional to $1/\rho^2$, yet a proper inverse gamma prior can also be employed, for example. Similar to Kawakami and Hashimoto (2023) and Cai and Sun (2021), we present Propositions 3.2 and 3.3. The former shows that using the improper prior on ρ^2 will lead to a proper posterior density. This ensures stability in making inference given that the posterior density remains proper. The latter is crucial to theoretically prove that the conditional prior on $\boldsymbol{\beta}$ leads to unimodality of the joint posterior because multi-modality can affect the reliability of the sampling algorithm where it makes it difficult to find the global maximum of the posterior distribution, that is a point estimate of the parameter, and the sampling algorithm may face issues in searching for the posterior distribution and converging to the true distribution.

Based on Proposition 3.3, Section 3.4.1 will show that the unconditional prior on $\boldsymbol{\beta}$ can result in multi-modality of the joint posterior. We further impose a gamma prior on λ_1^2 and η. We set hyper-parameters $a=b=c=d=1$ for simulation studies and real data analysis although different values can be chosen and it is shown in Section 3.4.2 that different values are insensitive to the model. The sensitivity analysis of hyper-parameters is detailed in Section 3.4.2.

As for the Gibbs sampling algorithm, the full conditional distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution and those of $\boldsymbol{\sigma}$, $\boldsymbol{v}$, $\boldsymbol{s}$ and ρ^2 are GIG distributions. The full conditional distribution of λ_1^2 is a Gamma distribution. The approximate Gibbs sampler is used for η. Appendix B.1 gives the details of the full conditional posterior distributions for the Gibbs sampling algorithm.

Proposition 3.2. *Let $\rho^2 \sim \pi(\rho^2) \propto 1/\rho^2$ (improper scale invariant prior). For fixed $\lambda_1 > 0$ and $\eta > 0$, the posterior distribution is proper for all n.*

Proof. The proof can be found under Proposition A.2 in Appendix A. □

Proposition 3.3. *Under the conditional prior for $\boldsymbol{\beta}$ given ρ^2 and fixed $\lambda_1 > 0$ and $\eta > 0$, the joint posterior $(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal with respect to $(\boldsymbol{\beta}, \rho^2)$.*

Proof. The proof can be found under Proposition A.3 in Appendix A. □

3.3.2 Bayesian Huberised Elastic Net Quantile Regression

We also present the Bayesian Huberised Elastic Net, that is,

$$\pi(\boldsymbol{\beta}|\rho^2, \lambda_3, \lambda_4) = \prod_{j=1}^{k} C\left(\tilde{\lambda}_3, \lambda_4\right) \frac{\lambda_3}{2\sqrt{\rho^2}} \exp\left\{-\frac{\lambda_3|\beta_j|}{\sqrt{\rho^2}} - \frac{\lambda_4\beta_j^2}{\rho^2}\right\}, \tag{3.5}$$

where $C\left(\tilde{\lambda}_3\, \lambda_4\right) = \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right)\left(\tilde{\lambda}_3\right)^{-1/2} \exp\left\{-\tilde{\lambda}_3\right\}$ is the normalising constant and $\tilde{\lambda}_3 = \lambda_3^2/(4\lambda_4)$. Note that by letting $\rho^2 = 1$, Equation (3.5) reduces to the original Bayesian Elastic Net (Li and Lin (2010)).

By using the normal scale-mixture property (Andrews and Mallows (1974)), the Bayesian Huberised Elastic Net can be expressed as the scale mixture of normal with truncated gamma density:

$$\pi(\boldsymbol{\beta}|\rho^2, \lambda_3, \lambda_4) = \prod_{j=1}^{k} \int_0^{\infty} \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right) \sqrt{\frac{2\lambda_4 t_j}{2\pi\rho^2(t_j-1)}} \sqrt{\frac{\tilde{\lambda}_3}{t_j}} \times \mathrm{N}\left(\beta_j; 0, \frac{\rho^2(t_j-1)}{2\lambda_4 t_j}\right) \exp\left\{-\tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j > 1) d\boldsymbol{t}.$$

With the Bayesian Huberised Elastic Net, we have the following hierarchical model:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{v} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta} + (1-2\tau)\boldsymbol{v}, \boldsymbol{V}),\\
&\sigma_i|\rho^2, \eta \sim \mathrm{GIG}\left(\frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right),\ v_i|\sigma_i \sim \mathrm{Exp}\left(\frac{\tau(1-\tau)}{2\sigma_i}\right),\ i = 1, \ldots, n,\\
&\beta_j|t_j, \lambda_4, \rho^2 \sim \mathrm{N}\left(0, \frac{2\rho^2(t_j-1)}{\lambda_4 t_j}\right),\ j = 1, \ldots, k,\\
&t_j|\tilde{\lambda}_3 \sim \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right)\sqrt{\frac{\tilde{\lambda}_3}{t_j}} \exp\left\{-\tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j > 1),\ j = 1, \ldots, k,\\
&\rho^2 \sim \pi(\rho^2) \propto \frac{1}{\rho^2},\ \tilde{\lambda}_3 \sim \mathrm{Gamma}(a_1, b_1),\ \lambda_4 \sim \mathrm{Gamma}(a_2, b_2),\ \eta \sim \mathrm{Gamma}(a_3, b_3),
\end{aligned}$$

where $a_1, a_2, a_3, b_1, b_2, b_3 \geq 0$ are hyper-parameters, they are set to 1 for simulation studies and real data analysis and $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function.

Appendix B.2 gives the details of the full conditional posterior distributions for the Gibbs

sampling algorithm. The full conditional distributions are all well-known distributions except the full conditional distributions of $\tilde{\lambda}_3$ and η, and the Metropolis-Hasting algorithm is employed on $\tilde{\lambda}_3$. We also present Proposition 3.4 for the use of improper prior on ρ^2. Based on Proposition 3.5, Section 3.4.1 provides a demonstration of how the use of the unconditional prior on $\boldsymbol{\beta}$ may result in multi-modality of the joint density.

Proposition 3.4. *Let $\rho^2 \sim \pi(\rho^2) \propto 1/\rho^2$ (improper scale invariant prior). For fixed $\lambda_3 > 0$, $\lambda_4 > 0$ and $\eta > 0$, the posterior distribution is proper for all n.*

Proof. The proof can be found under Proposition A.4 in Appendix A. □

Proposition 3.5. *Under the conditional prior for $\boldsymbol{\beta}$ given ρ^2 and fixed $\lambda_3 > 0$, $\lambda_4 > 0$ and $\eta > 0$, the joint posterior $(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal with respect to $(\boldsymbol{\beta}, \rho^2)$.*

Proof. The proof can be found under Proposition A.5 in Appendix A. □

3.3.3 Approximate Gibbs Sampler for Estimation of η

In this section, we will briefly discuss the approximate Gibbs sampler for the data-dependent estimation of η that is proposed by Kawakami and Hashimoto (2023). Notice that in a Bayesian Huberised regularised quantile regression model, the full conditional distribution of η is

$$\pi(\eta|\boldsymbol{\sigma}, \rho^2) \propto \frac{1}{K_{3/2}(\eta)^n} \eta^{a-1} \exp\left\{-\eta\left(\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right) + b\right)\right\}, \tag{3.6}$$

where $a = c$ and $b = d$ in case of Bayesian Huberised Lasso quantile regression and $a = a_3$ and $b = b_3$ in case of Bayesian Huberised Elastic Net quantile regression. Since the right side of Equation (3.6) contains the modified Bessel function of the second kind, the full conditional distribution of η does not have a conjugacy property. However, it is possible to approximate Equation (3.6) by a common probability distribution.

For the selection of an initial value of the approximate Gibbs sampler, we need to approximate the modified Bessel function of the second kind. According to Abramowitz and Stegun (1964), we have $K_\nu(x) \sim (1/2)\,\Gamma(\nu)\,(x/2)^{-\nu}$ as $x \to 0$ for $\nu > 0$ and $K_\nu(x) \sim \sqrt{x/(2\pi)}e^{-x}$ as $x \to \infty$. Kawakami and Hashimoto (2023) stated that in either case, it would not make much difference in estimating η. So, we will focus on the latter case only for this chapter.

As $\eta \to \infty$, we have

$$\pi(\eta|\boldsymbol{\sigma}, \rho^2) \approx \eta^{a+n/2-1} \exp\left\{-\eta\left(\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\sigma_i}{\rho^2}+\frac{\rho^2}{\sigma_i}\right)+b-n\right)\right\},$$

which holds the approximation $\pi(\eta|\boldsymbol{\sigma}, \rho^2) \approx \text{Gamma}(\eta; a+n/2, 1/2\sum_{i=1}^{n}\left(\sigma_i/\rho^2+\rho^2/\sigma_i\right)+b-n\big)$ for large η.

The algorithm of the approximate Gibbs sampler is as follows. Given the current Markov chain states $(\boldsymbol{\sigma}, \rho^2)$, we set the initial value as $A = a+n/2$ and $B = 1/2\sum_{i=1}^{n}\left(\sigma_i/\rho^2+\rho^2/\sigma_i\right)+b-n$. For $m = 1, \ldots, M$, do the following steps

- $\eta \leftarrow A/B$;
- $A \leftarrow a + n\eta^2\partial^2/\partial\eta^2[\log K_{3/2}(\eta)]$;
- $B \leftarrow b + (A-a)/\eta + n\partial/\partial\eta[\log K_{3/2}(\eta)] + 1/2\sum_{i=1}^{n}\left(\sigma_i/\rho^2+\rho^2/\sigma_i\right)$.

until $|\eta/(A/B)-1| < \varepsilon$ or in other words, the convergence of η is met. The full derivation of the algorithm is detailed in Kawakami and Hashimoto (2023) and they also illustrated that in their simulation results, the approximation is close to the true full conditional distribution and the approximation accuracy increases as the sample size increases. For simulation studies and real data analysis, we set $M = 10$ and a tolerance $\varepsilon = 10^{-8}$.

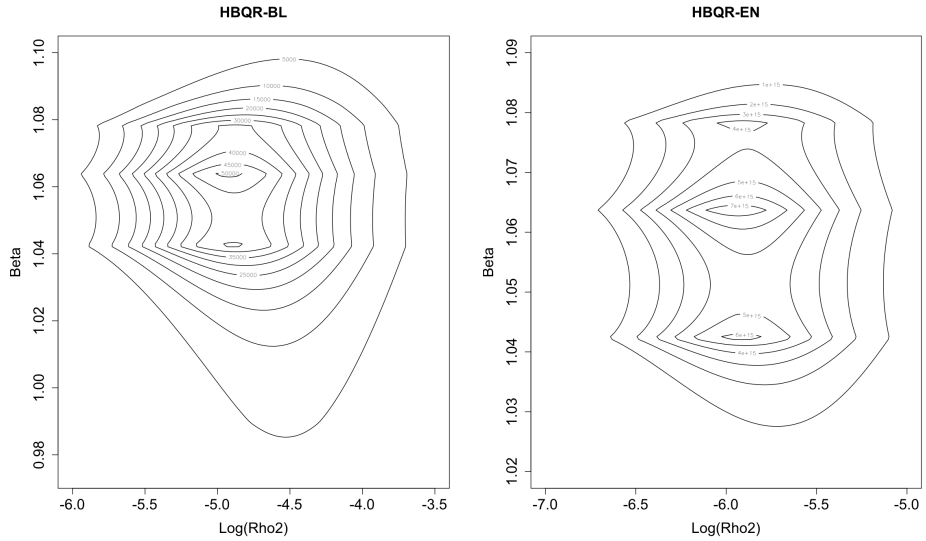
3.4 Simulations

Throughout the subsections, we conducted a wide variety of simulation studies to assess the performance of the proposed methods. Firstly, we investigated the multi-modality of the joint posterior density using unconditional prior on $\boldsymbol{\beta}$ in relation to Propositions 3.3 and 3.5. Secondly, we tested how the change in hyper-parameters of the priors for Bayesian Huberised regularised quantile regression influence the estimation. Finally, we examined the numerical performance and robustness of the proposed methods under various settings.

3.4.1 Multi-modality of Joint Posteriors

As related to Propositions 3.3 and 3.5, we presented a simple simulation to demonstrate that the unconditional prior on $\boldsymbol{\beta}$ can result in multi-modality of the joint posterior. Instead of Equations (3.4) and (3.5), we specified the unconditional Lasso prior (Equation (1.11)) and the unconditional Elastic Net prior (Equation (1.13)) with same improper prior $\pi(\rho^2) \propto$

Figure 3.2: Contour plot of an artificially generated posterior density of $(\beta, \log(\rho^2))$ of the joint posterior density (Equations (3.7) and (3.8)) for Bayesian Huberised Lasso quantile regression and Bayesian Huberised Elastic Net quantile regression, respectively. The logarithm of ρ^2 is used for better visibility.



$1/\rho^2$. Then the joint posterior distribution of $\boldsymbol{\beta}$ and ρ^2 for Bayesian Huberised Lasso quantile regression is proportional to

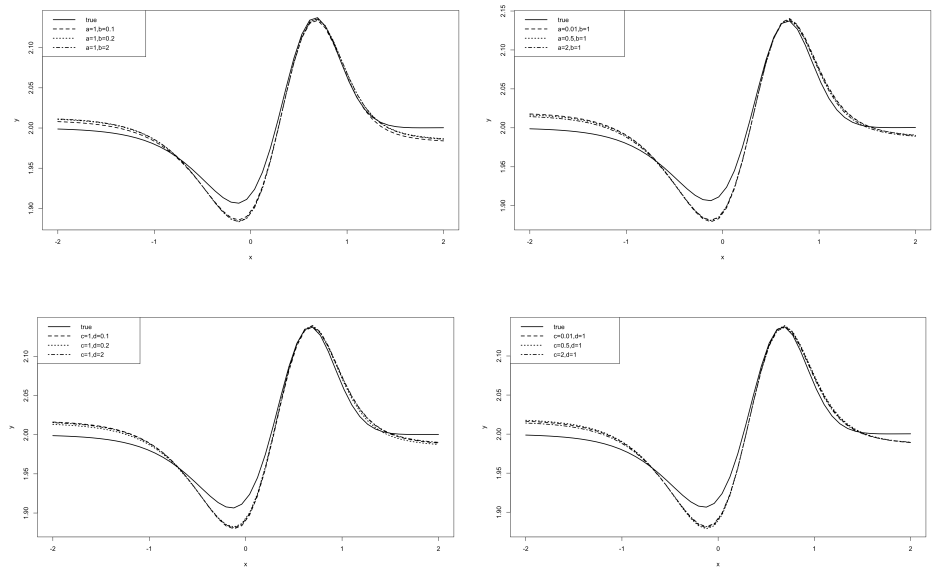
$$\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) \propto (\rho^2)^{-n/2-1} \exp\left\{-\lambda_1 \sum_{j=1}^{k} |\beta_j|\right\} \times \prod_{i=1}^{n} \exp\left\{-\sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}\right\}, \tag{3.7}$$

and that for Bayesian Huberised Elastic Net quantile regression is proportional to

$$\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) \propto (\rho^2)^{-n/2-1} \exp\left\{-\lambda_3 \sum_{j=1}^{k} |\beta_j| - \lambda_4 \sum_{j=1}^{k} \beta_j^2\right\} \times \prod_{i=1}^{n} \exp\left\{-\sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}\right\}. \tag{3.8}$$

In Appendix A, it is shown that using the conditional priors (Equations (3.4) and (3.5) leads to a unimodal posterior for any choice of $\lambda_1, \lambda_3, \lambda_4 \geq 0$ and $\eta > 0$ with an improper prior $\pi(\rho^2)$. On the other hand, the joint posteriors (Equations (3.7) and (3.8) can have more than one mode. For example, Figure 3.2 shows the contour plots of a multi-modal joint density of β and $\log(\rho^2)$, which have seen three modes in the joint density. This suggests that the sampling algorithm is unlikely to be reliable when using the unconditional priors

Figure 3.3: Sensitivity analysis of hyper-parameters for the Bayesian Huberised Lasso quantile regression.



because it does not find the global maximum of the joint posterior distribution and it would lead to issues in converging to the true distribution. Thus, this simulation study emphasises the importance of unimodality of the joint posterior density.

This particular example considered the following data generated model,

$$y_i = x_i\beta + \epsilon_i, \quad \epsilon_i \sim \text{AL}(0, \sigma = 0.03, \tau = 0.5),$$

where $\beta = 1$ and $x_i \sim \text{N}(0, 1)$ for $i = 1, \ldots, 10$, which is similar to Cai and Sun (2021). Due to multi-modality in the joint posterior with unconditional prior on $\boldsymbol{\beta}$, we use the prior for $\boldsymbol{\beta}$ conditioning on the scale parameter ρ^2.

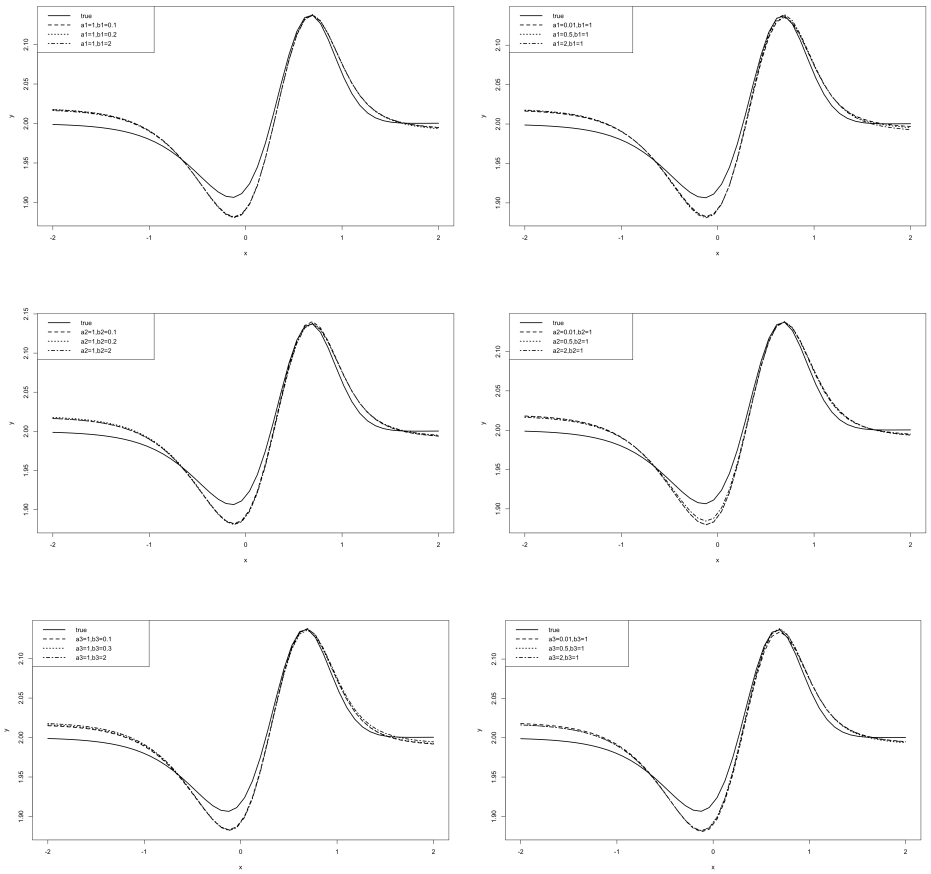
3.4.2 Sensitivity Analysis of Hyper-parameters

In this subsection, we tested the sensitivity of hyper-parameters of gamma prior of η, λ_1, λ_3 and λ_4 on the posterior estimates for the proposed methods. We equally divided $x \in [-2, 2]$ into 50 pieces and the data were generated from

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \text{AL}(0, \sigma = 0.03, \tau = 0.5), \quad i = 1, \ldots, 50,$$

where $\boldsymbol{x}_i = \left(\left(1 + e^{-4(x_i - 0.3)}\right)^{-1}, \left(1 + e^{3(x_i - 0.2)}\right)^{-1}, \left(1 + e^{-4(x_i - 0.7)}\right)^{-1}, \left(1 + e^{5(x_i - 0.8)}\right)^{-1}\right)^T$,

Figure 3.4: Sensitivity analysis of hyper-parameters for the Bayesian Huberised Elastic Net quantile regression.



and $\boldsymbol{\beta} = (1, 1, 1, 1)^T$. This indicates that the true curve is

$$f(x) = \left(1 + e^{-4(x-0.3)}\right)^{-1} + \left(1 + e^{3(x-0.2)}\right)^{-1} + \left(1 + e^{-4(x-0.7)}\right)^{-1} + \left(1 + e^{5(x-0.8)}\right)^{-1}.$$

In fact, this function was utilised in Jullion and Lambert (2007) to test the sensitivity of hyper-parameters of the gamma prior on the scale component in Bayesian P-spline.

We considered the proposed models to estimate $\boldsymbol{\beta}$. Note that there are four prior hyper-parameters a, b, c and d in the Bayesian Huberised Lasso quantile regression and six prior hyper-parameters a_1, b_1, a_2, b_2, a_3 and b_3 in the Bayesian Huberised Elastic Net quantile regression. We mainly set $a = b = c = d = a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 1$ in both simulation studies and data analysis. We generated 3000 posterior samples after discarding the first 1000 posterior samples as a burn-in. Then we plotted $y_i = \boldsymbol{x}_i\boldsymbol{\beta}$ for $i = 1, \ldots, 50$ in Figures 3.3 and 3.4 for both proposed Bayesian models, where $\boldsymbol{\beta}$ is the posterior mean for the corresponding proposed model. In Figure 3.3, we fixed $a = 1$ with b varied for the top-left plot and $b = 1$ with a varied for the top-right plot. In both cases, we kept $c = d = 1$

Table 3.1: Numerical results based on 300 replications in Simulation 1 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.3675	0.2548	0.9238	0.8711
	HBQR-EN100	0.3795	0.2585	0.9673	0.8725
	BQR-BL100	0.3956	0.2627	1.3406	0.9411
	BQR-EN100	0.3859	0.2628	0.9624	0.8821
	HBQR-BL200	0.3328	0.2108	0.6624	0.8483
	HBQR-EN200	0.3380	0.2104	0.6822	0.8576
	BQR-BL200	0.3534	0.2118	0.9076	0.9311
	BQR-EN200	0.3476	0.2123	0.6819	0.8732
$\tau = 0.5$	HBQR-BL100	0.2659	0.2059	0.9465	0.9211
	HBQR-EN100	0.2678	0.2100	0.9834	0.9289
	HBL100	0.2426	0.1891	0.9553	0.9432
	BQR-BL100	0.2468	0.1946	1.2976	0.9848
	BQR-EN100	0.2502	0.1968	0.9838	0.9413
	HBQR-BL200	0.1962	0.1550	0.6810	0.9143
	HBQR-EN200	0.1952	0.1549	0.6927	0.9192
	HBL200	0.1777	0.1404	0.6763	0.9384
	BQR-BL200	0.1825	0.1452	0.8756	0.9778
	BQR-EN200	0.1841	0.1460	0.6900	0.9329
$\tau = 0.75$	HBQR-BL100	0.3635	0.2521	0.9395	0.8756
	HBQR-EN100	0.3662	0.2503	0.9891	0.8722
	BQR-BL100	0.3943	0.2587	1.3484	0.9440
	BQR-EN100	0.3853	0.2664	0.9920	0.8859
	HBQR-BL200	0.3386	0.2141	0.6703	0.8573
	HBQR-EN200	0.3315	0.2105	0.6895	0.8562
	BQR-BL200	0.3571	0.2146	0.9053	0.9340
	BQR-EN200	0.3512	0.2174	0.6890	0.8659

fixed. Both bottom plots of Figure 3.3 followed in a similar manner. As for Figure 3.4, we also fixed $a_1 = 1$ with b_1 varied for the top-left plot, whilst keeping $a_2 = b_2 = a_3 = b_3 = 1$. The rest of Figure 3.4 also followed in a similar manner. From the figures, we observe that the estimation results do not change very much for different hyper-parameters. Thus, the choice of hyper-parameters is insensitive for the Bayesian Huberised regularised quantile regression.

3.4.3 Simulation Studies

In simulation studies, we illustrated performance of the proposed methods. We compared the point and interval estimation performance of the proposed methods with those of some existing methods. To this end, we considered the following regression model with $n \in$

Figure 3.5: Boxplots of RMSE based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.25$).

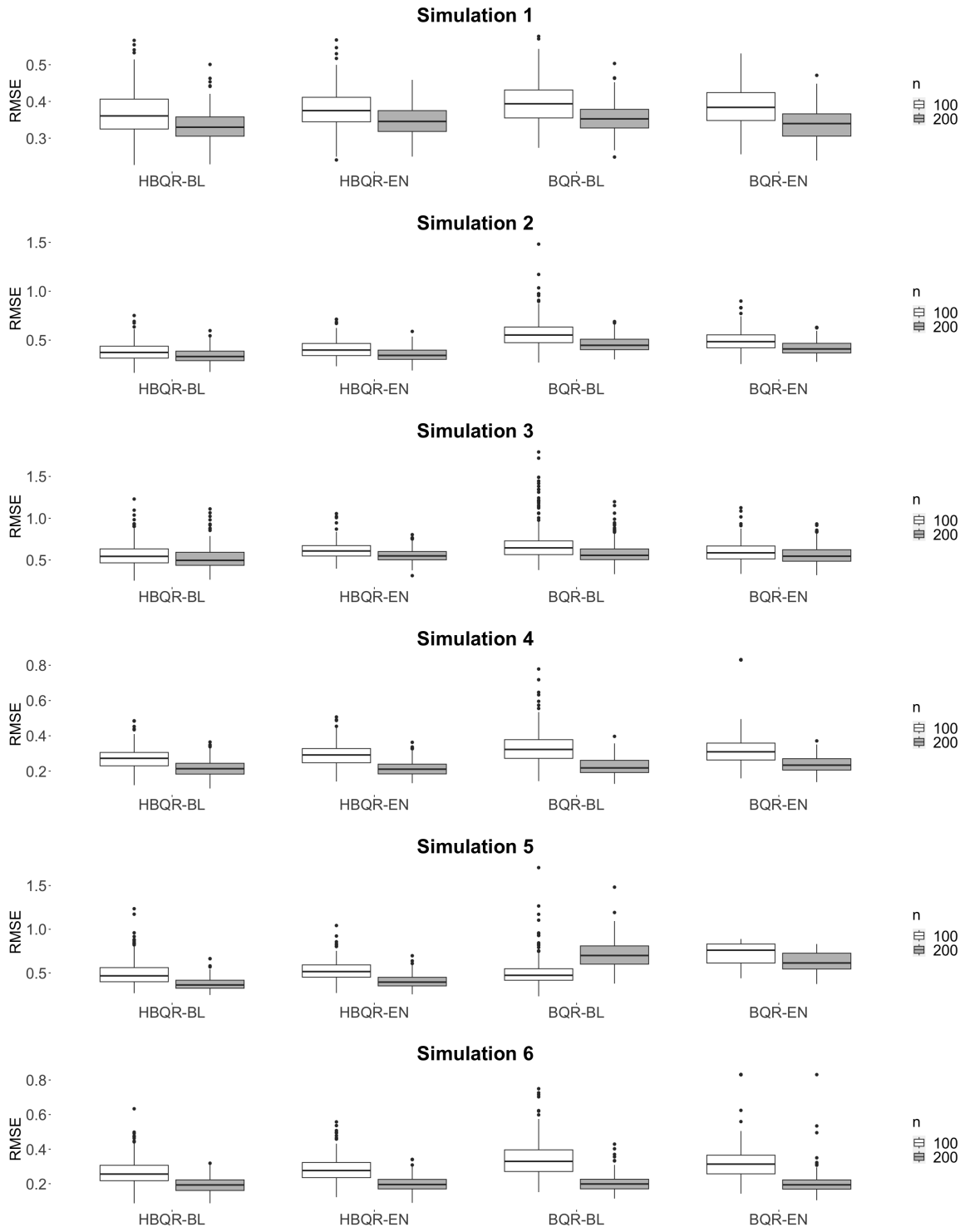


Figure 3.6: Boxplots of MMAD based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.25$).

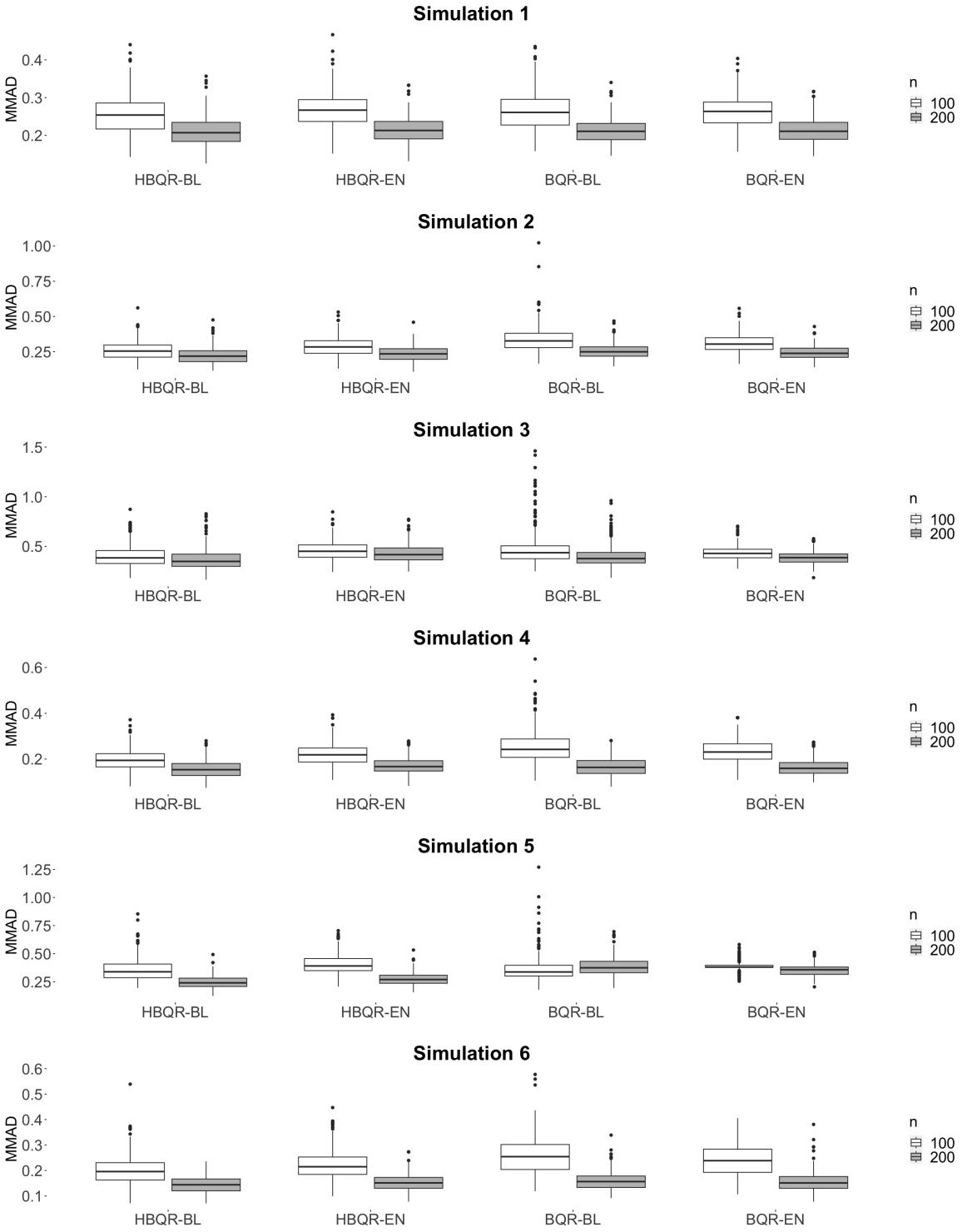


Figure 3.7: Boxplots of AL based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.25$).

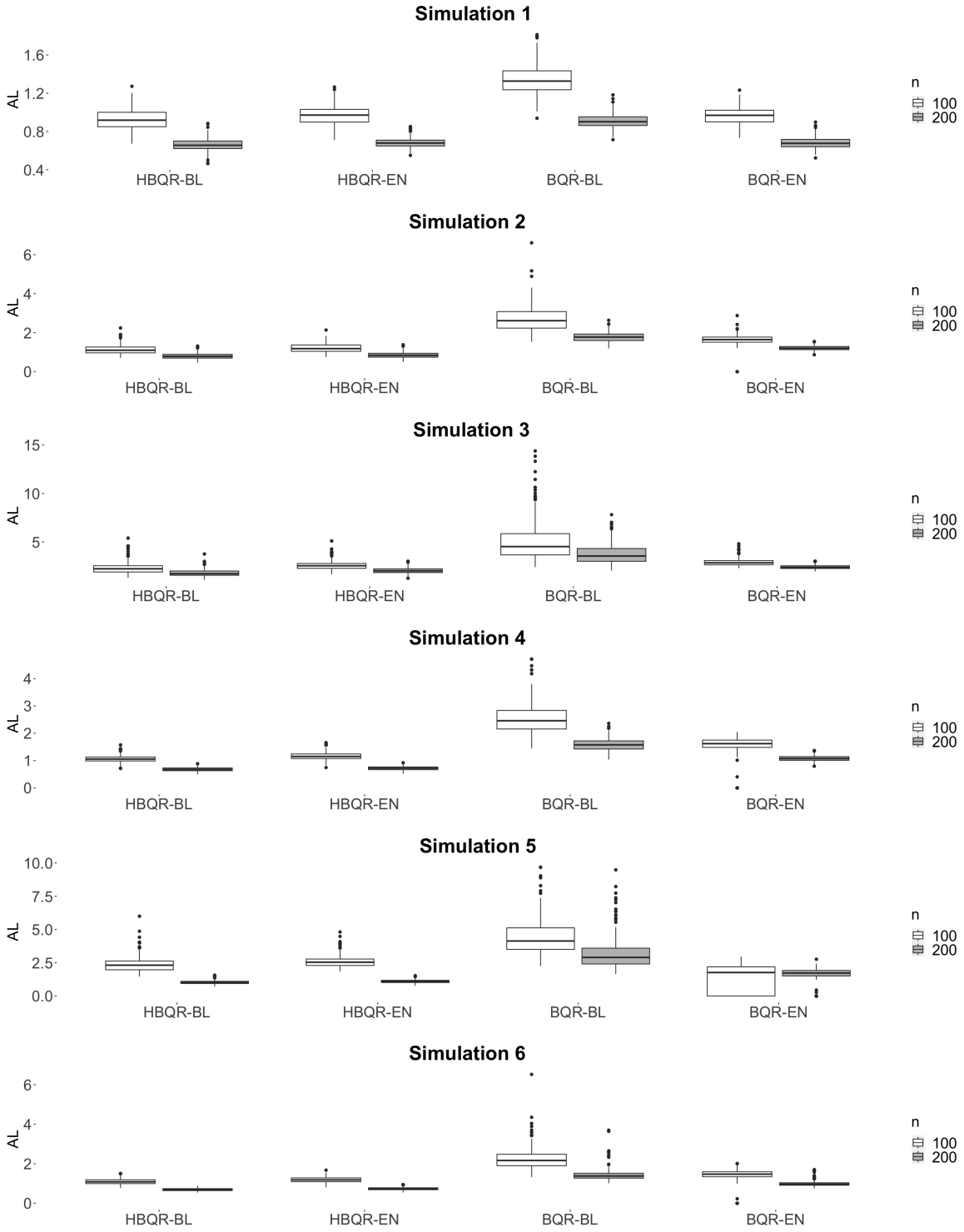


Figure 3.8: Boxplots of RMSE based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN ($\tau = 0.5$).

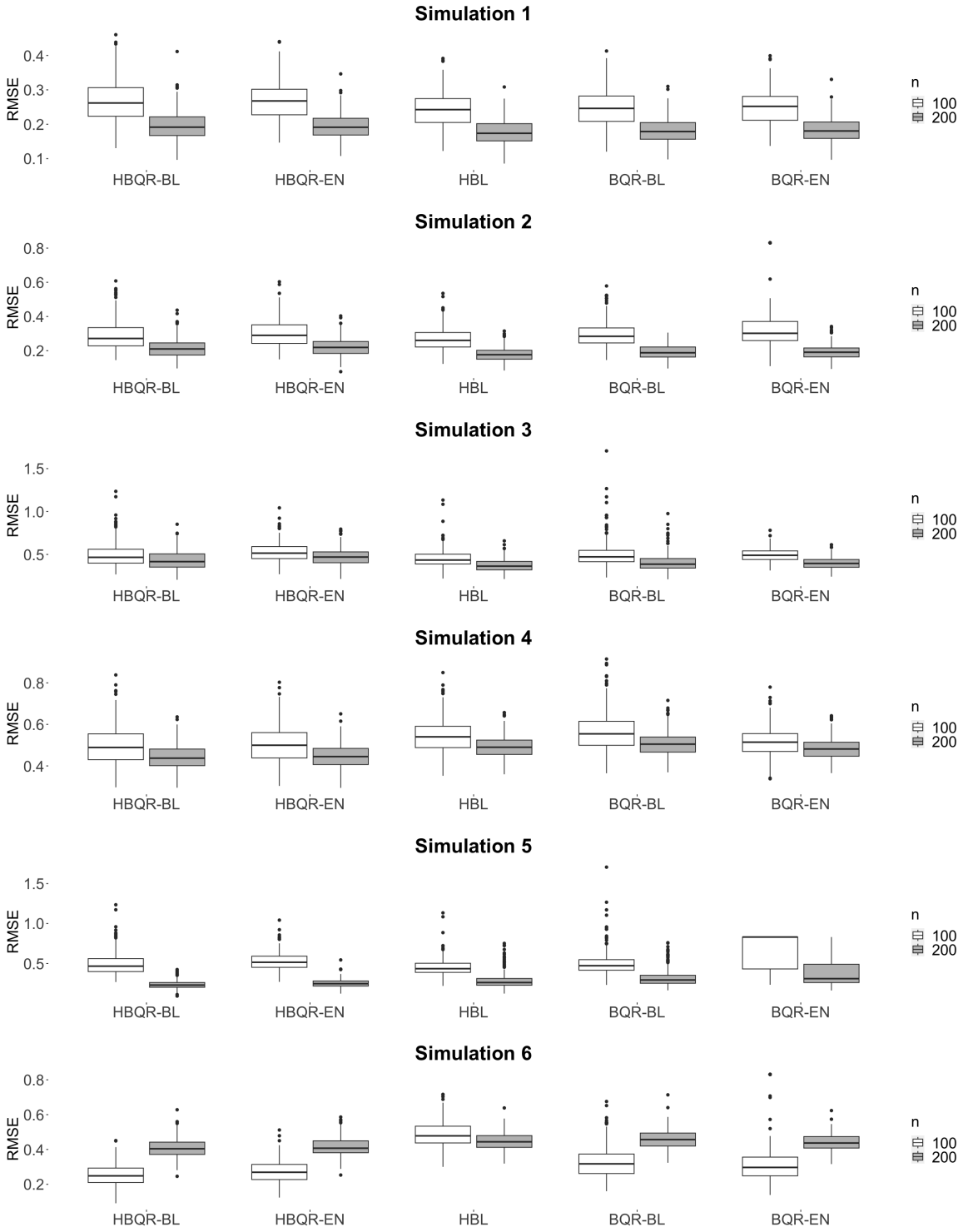


Figure 3.9: Boxplots of MMAD based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN ($\tau = 0.5$).

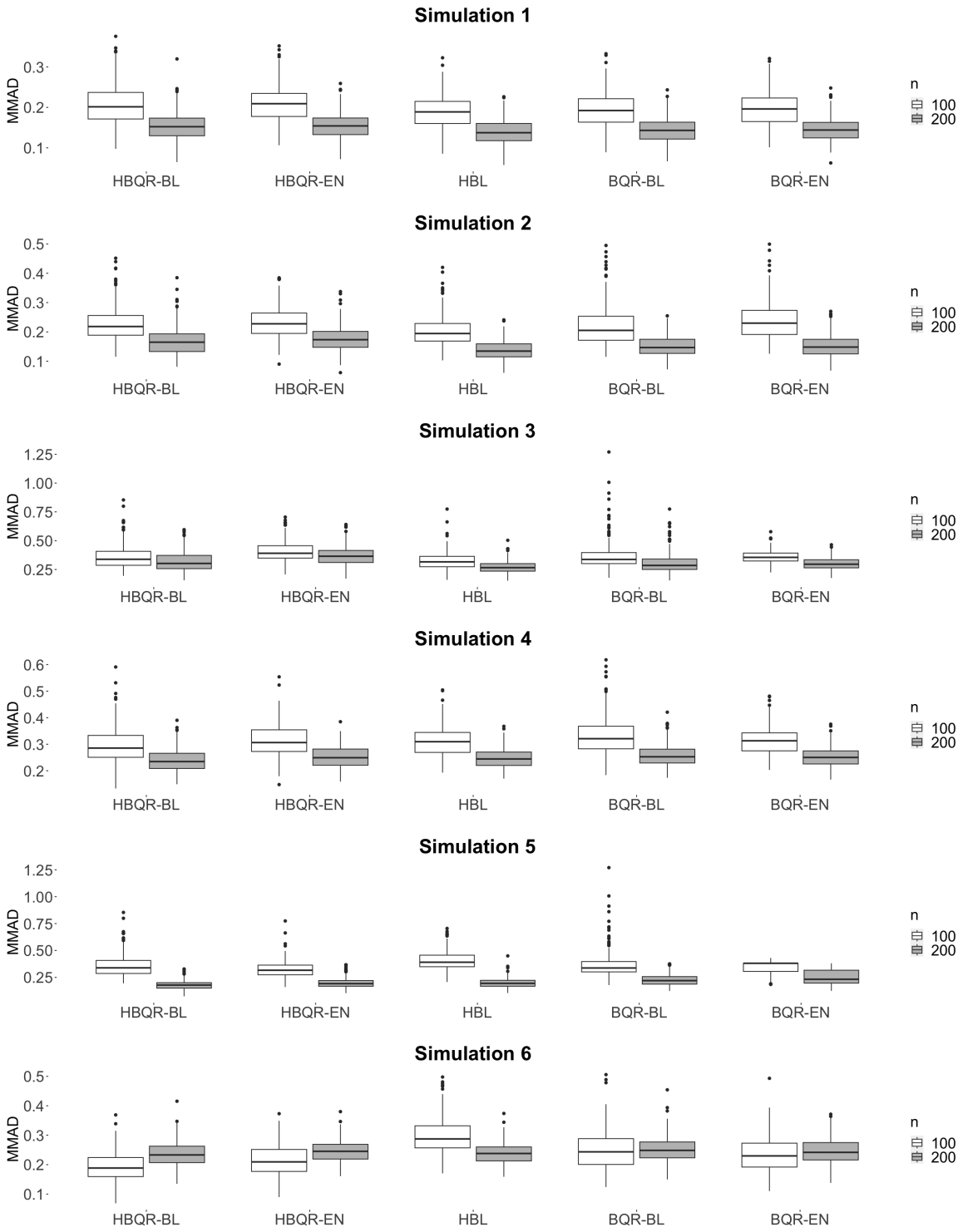


Figure 3.10: Boxplots of AL based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN ($\tau = 0.5$).

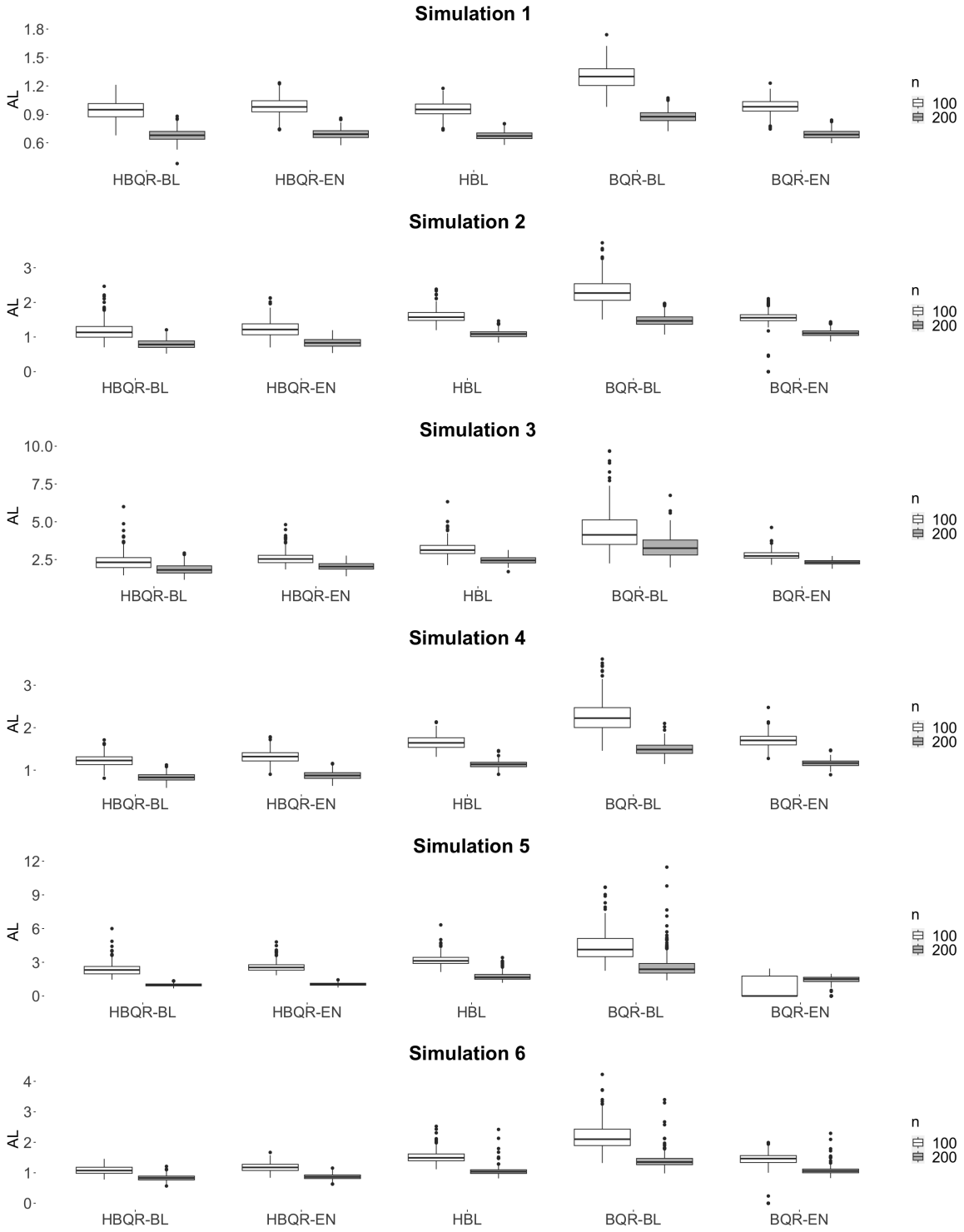


Figure 3.11: Boxplots of posterior median of η based on 300 replications in six simulation scenarios for HBQR-BL and HBQR-EN ($n = 100$).

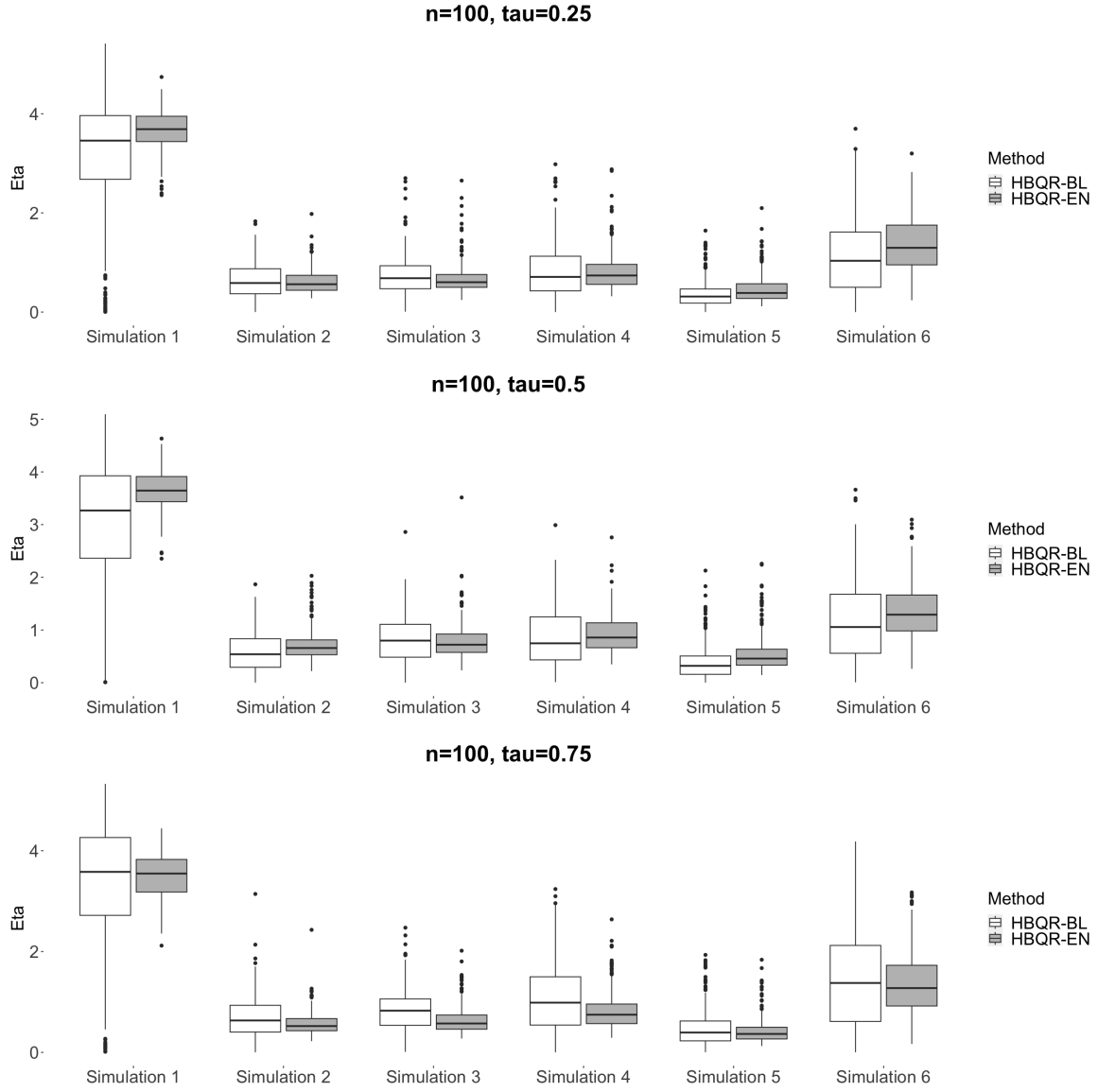


Table 3.2: Numerical results based on 300 replications in Simulation 2 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.3822	0.2579	1.1290	0.9046
	HBQR-EN100	0.4051	0.2767	1.2135	0.9002
	BQR-BL100	0.5643	0.3385	2.6992	0.9506
	BQR-EN100	0.4950	0.3100	1.6428	0.9311
	HBQR-BL200	0.3418	0.2233	0.8011	0.8762
	HBQR-EN200	0.3528	0.2368	0.8484	0.8765
	BQR-BL200	0.4588	0.2540	1.7800	0.9521
	BQR-EN200	0.4221	0.2436	1.2130	0.9394
$\tau = 0.5$	HBQR-BL100	0.2886	0.2203	1.1750	0.9533
	HBQR-EN100	0.2945	0.2336	1.2251	0.9522
	HBL100	0.2683	0.2013	1.6040	0.9946
	BQR-BL100	0.2954	0.2262	2.3330	0.9992
	BQR-EN100	0.3273	0.2340	1.5357	0.9733
	HBQR-BL200	0.2060	0.1591	0.7990	0.9279
	HBQR-EN200	0.2130	0.1679	0.8332	0.9297
	HBL200	0.1793	0.1386	1.0921	0.9943
	BQR-BL200	0.1926	0.1511	1.4813	0.9992
	BQR-EN200	0.1941	0.1522	1.1176	0.9929
$\tau = 0.75$	HBQR-BL100	0.3791	0.2624	1.1734	0.9066
	HBQR-EN100	0.3889	0.2822	1.2615	0.9000
	BQR-BL100	0.5761	0.3468	2.7564	0.9525
	BQR-EN100	0.4697	0.3086	1.8026	0.9433
	HBQR-BL200	0.3324	0.2188	0.8013	0.8792
	HBQR-EN200	0.3352	0.2198	0.8442	0.8705
	BQR-BL200	0.4530	0.2498	1.7595	0.9521
	BQR-EN200	0.4087	0.2416	1.2458	0.9437

$\{100, 200\}$, $k = 20$ and $\tau \in \{0.25, 0.5, 0.75\}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where $\beta_0 = 1$, $\beta_1 = 3$, $\beta_2 = 0.5$,$\beta_4 = \beta_{11} = 1$, $\beta_7 = 1.5$ and the other β_j's were set to 0. We assumed that $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the response vector. The predictors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^T$ were generated from a multivariate normal distribution $\mathrm{N}_k(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (r^{|i-j|})_{1 \le i,j \le k}$ for $|r| < 1$. Similar to Kawakami and Hashimoto (2023), and Lambert-Lacroix and Zwald (2011), we considered the following six scenarios.

- Simulation 1: Low correlation and Gaussian noise. $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{I}_n)$, $\sigma = 2$ and $r = 0.5$.
- Simulation 2: Low correlation and large outliers. $\epsilon = W/\sqrt{\mathrm{Var}(W)}$, $\sigma = 9.67$ and

Table 3.3: Numerical results based on 300 replications in Simulation 3 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.5676	0.4038	2.3005	0.9259
	HBQR-EN100	0.5805	0.4285	2.5853	0.9268
	BQR-BL100	0.6854	0.4754	5.0697	0.9524
	BQR-EN100	0.6147	0.4293	2.9063	0.9430
	HBQR-BL200	0.5173	0.3732	1.8027	0.9094
	HBQR-EN200	0.5319	0.3825	2.0427	0.9060
	BQR-BL200	0.5836	0.4033	3.7568	0.9519
	BQR-EN200	0.5542	0.3872	2.4114	0.9422
$\tau = 0.5$	HBQR-BL100	0.4949	0.3576	2.3700	0.9719
	HBQR-EN100	0.4958	0.3856	2.5830	0.9700
	HBL100	0.4525	0.3248	3.1990	0.9980
	BQR-BL100	0.5023	0.3671	4.4369	0.9992
	BQR-EN100	0.4910	0.3566	2.7858	0.9871
	HBQR-BL200	0.4317	0.3197	1.8692	0.9611
	HBQR-EN200	0.4317	0.3178	2.0474	0.9624
	HBL200	0.3707	0.2719	2.4534	0.9965
	BQR-BL200	0.4053	0.3041	3.3309	0.9992
	BQR-EN200	0.3975	0.2995	2.3221	0.9921
$\tau = 0.75$	HBQR-BL100	0.5563	0.3993	2.3546	0.9300
	HBQR-EN100	0.5911	0.4240	2.7935	0.9321
	BQR-BL100	0.6650	0.4591	4.9652	0.9531
	BQR-EN100	0.5984	0.4318	3.3417	0.9482
	HBQR-BL200	0.5241	0.3856	1.8819	0.9114
	HBQR-EN200	0.5390	0.3926	2.1538	0.9033
	BQR-BL200	0.5786	0.4008	3.8405	0.9522
	BQR-EN200	0.5455	0.3940	2.6987	0.9479

$r = 0.5$. W is a random variable according to the contaminated density defined by $0.9 \times \mathrm{N}(0,1) + 0.1 \times \mathrm{N}(0, 15^2)$, where $\sqrt{\mathrm{Var}(W)} = 4.83$.

- Simulation 3: High correlation and large outliers. $\epsilon = W/\sqrt{\mathrm{Var}(W)}$, $\sigma = 9.67$ and $r = 0.95$.
- Simulation 4: Large outliers and skew Student-t noise. $\epsilon_i \sim 0.9 \times \text{Skew-t}_3(\gamma = 3) + 0.1 \times \mathrm{N}(0, 20^2)$, $\sigma = 1$ and $r = 0.5$.
- Simulation 5: Heavy-tailed noise. $\epsilon_i \sim \mathrm{Cauchy}(0, 1)$, $\sigma = 2$ and $r = 0.5$.
- Simulation 6: Multiple outliers. $\epsilon_i \sim 0.8 \times \text{Skew-t}_3(\gamma = 3) + 0.1 \times \mathrm{N}(0, 10^2) + 0.1 \times \mathrm{Cauchy}(0, 1)$, $\sigma = 1$ and $r = 0.5$.

For the generated dataset, we applied the proposed robust methods denoted by HBQR-BL

Table 3.4: Numerical results based on 300 replications in Simulation 4 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.2705	0.1957	1.0598	0.9416
	HBQR-EN100	0.2803	0.2093	1.1574	0.9408
	BQR-BL100	0.3335	0.2542	2.5406	0.9990
	BQR-EN100	0.3300	0.2362	1.5557	0.9769
	HBQR-BL200	0.2265	0.1572	0.6799	0.9038
	HBQR-EN200	0.2287	0.1584	0.7199	0.9004
	BQR-BL200	0.2168	0.1679	1.5867	0.9979
	BQR-EN200	0.2143	0.1636	1.0815	0.9721
$\tau = 0.5$	HBQR-BL100	0.4965	0.2950	1.2265	0.9193
	HBQR-EN100	0.5048	0.3049	1.2383	0.9213
	HBL100	0.5434	0.3111	1.6510	0.9450
	BQR-BL100	0.5609	0.3312	2.2801	0.9503
	BQR-EN100	0.5184	0.3134	1.6965	0.9430
	HBQR-BL200	0.4407	0.2385	0.8315	0.9029
	HBQR-EN200	0.4466	0.2314	0.8457	0.9006
	HBL200	0.4927	0.2461	1.1366	0.9446
	BQR-BL200	0.5061	0.2571	1.4968	0.9506
	BQR-EN200	0.4835	0.2515	1.1630	0.9444
$\tau = 0.75$	HBQR-BL100	0.8388	0.4236	1.4484	0.9070
	HBQR-EN100	0.8417	0.4460	1.5382	0.9005
	BQR-BL100	1.1330	0.5333	2.7729	0.9492
	BQR-EN100	1.0732	0.5514	2.1302	0.9281
	HBQR-BL200	0.7868	0.3716	1.0323	0.8829
	HBQR-EN200	0.7940	0.3874	1.0884	0.8676
	BQR-BL200	1.0651	0.4624	1.9991	0.9462
	BQR-EN200	1.0206	0.4729	1.5216	0.9149

and HBQR-EN where they were employed with Bayesian Huberised Lasso and Bayesian Huberised Elastic Net, respectively. We also applied the existing robust methods, including Bayesian linear regression with Bayesian Huberised Lasso (Kawakami and Hashimoto (2023)), and Bayesian quantile regression with original Bayesian Lasso and Bayesian Elastic Net (Li et al. (2010)) denoted by HBL, BQR-BL and BQR-EN, respectively. For HBL and BQR-BL methods, we assumed $\lambda_1 \sim \text{Gamma}(a = 1, b = 1)$ and for the BQR-EN method, we assumed $\lambda_1 \sim \text{Gamma}(a_1 = 1, b_1 = 1)$ and $\lambda_2 \sim \text{Gamma}(a_2 = 1, b_2 = 1)$. For HBQR-BL and HBQR-EN methods, we implemented both Gibbs and Metropolis-within-Gibbs sampling algorithms, respectively and set all the hyper-parameters to 1. It is noteworthy that different hyper-parameters could be chosen, since it is shown in Section 3.4.2 that the choice of hyper-parameters does not matter for our proposed methods.

Table 3.5: Numerical results based on 300 replications in Simulation 5 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.4465	0.3012	1.4890	0.9169
	HBQR-EN100	0.4460	0.3193	1.5992	0.9189
	BQR-BL100	0.8163	0.4688	4.3100	0.9522
	BQR-EN100	0.7197	0.3866	1.2336	0.8401
	HBQR-BL200	0.3741	0.2456	1.0247	0.9108
	HBQR-EN200	0.3926	0.2647	1.1039	0.9035
	BQR-BL200	0.7137	0.3841	3.1585	0.9521
	BQR-EN200	0.6376	0.3508	1.5062	0.8983
$\tau = 0.5$	HBQR-BL100	0.3292	0.2257	1.4712	0.9668
	HBQR-EN100	0.2469	0.2002	1.0604	0.9616
	HBL100	0.4151	0.2771	2.3541	0.9909
	BQR-BL100	0.4577	0.3172	3.6592	0.9963
	BQR-EN100	0.6727	0.3435	0.7418	0.8184
	HBQR-BL200	0.2320	0.1780	0.9866	0.9594
	HBQR-EN200	0.2495	0.1861	1.0505	0.9570
	HBL200	0.2861	0.1996	1.7512	0.9951
	BQR-BL200	0.3155	0.2263	2.6611	0.9990
	BQR-EN200	0.4242	0.2551	1.2599	0.9279
$\tau = 0.75$	HBQR-BL100	0.4315	0.2998	1.5445	0.9321
	HBQR-EN100	0.4356	0.3011	1.7046	0.9303
	BQR-BL100	0.7818	0.4538	4.2976	0.9649
	BQR-EN100	0.6703	0.3859	1.7851	0.8868
	HBQR-BL200	0.3732	0.2442	1.0471	0.9089
	HBQR-EN200	0.3913	0.2607	1.1202	0.9008
	BQR-BL200	0.7062	0.3789	3.2171	0.9554
	BQR-EN200	0.5920	0.3386	1.8842	0.9287

Whilst applying the above methods, we generated 2000 posterior samples after discarding the first 500 samples as a burn-in. We computed posterior median of each element of β_j's for point estimates of β_j's, and the performance was evaluated via root of mean squared error (RMSE) defined as $\left[(k+1)^{-1}\sum_{j=0}^{k}(\hat{\beta}_j - \beta_j^{\text{true}})^2\right]^{1/2}$, and median of mean absolute deviation (MMAD) defined as median $\left[(k+1)^{-1}\sum_{j=0}^{k}\left|\hat{\beta}_j - \beta_j^{\text{true}}\right|\right]$. We also computed 95% credible intervals of β_j's, and calculated average lengths (ALs) and coverage probabilities (CPs) defined as $(k+1)^{-1}\sum_{j=0}^{k}|\text{CI}_j|$ and $(k+1)^{-1}\sum_{j=0}^{k}\text{I}(\beta_j \in \text{CI}_j)$, respectively. These values were averaged over 300 replications of generating datasets.

We reported numerical results in Tables 3.1-3.6. In Simulation 1, there was no outlier in generated datasets. At the median level ($\tau = 0.5$), both HBL and BQR-BL methods outperformed other methods in terms of RMSE and MMAD. Yet, at the lower and upper

Table 3.6: Numerical results based on 300 replications in Simulation 6 at different quantile levels ($\tau = 0.25, 0.5, 0.75$) for HBQR-BL, HBQR-EN, HBL BQR-BL and BQR-EN.

	Methods	RMSE	MMAD	AL	CP
$\tau = 0.25$	HBQR-BL100	0.2669	0.2034	1.0971	0.9503
	HBQR-EN100	0.2734	0.2141	1.1773	0.9546
	BQR-BL100	0.3397	0.2591	2.2468	0.9960
	BQR-EN100	0.3285	0.2438	1.4439	0.9639
	HBQR-BL200	0.1931	0.1454	0.6920	0.9203
	HBQR-EN200	0.1978	0.1466	0.7353	0.9235
	BQR-BL200	0.2025	0.1595	1.4260	0.9984
	BQR-EN200	0.2002	0.1557	0.9916	0.9814
$\tau = 0.5$	HBQR-BL100	0.2550	0.1937	1.0816	0.9516
	HBQR-EN100	0.2634	0.2050	1.1773	0.9546
	HBL100	0.4862	0.2947	1.5152	0.9384
	BQR-BL100	0.3228	0.2480	2.1960	0.9970
	BQR-EN100	0.3200	0.2370	1.4295	0.9698
	HBQR-BL200	0.4094	0.2328	0.8321	0.8971
	HBQR-EN200	0.4147	0.2355	0.8701	0.8959
	HBL200	0.4450	0.2384	1.0562	0.9379
	BQR-BL200	0.4584	0.2502	1.3950	0.9498
	BQR-EN200	0.4392	0.2452	1.0774	0.9363
$\tau = 0.75$	HBQR-BL100	0.8115	0.4248	1.4426	0.8998
	HBQR-EN100	0.8229	0.4399	1.5531	0.8946
	BQR-BL100	1.0427	0.5067	2.5333	0.9468
	BQR-EN100	0.9903	0.5243	1.8811	0.9092
	HBQR-BL200	0.7661	0.3735	1.0210	0.8689
	HBQR-EN200	0.7734	0.3836	1.0743	0.8630
	BQR-BL200	0.9763	0.4381	1.8219	0.9437
	BQR-EN200	0.9407	0.4492	1.3898	0.9019

quantile levels ($\tau = 0.25, 0.75$), the proposed methods outperformed the existing methods. In terms of AL, the HBQR-BL method has the shortest AL amongst methods at all quantile levels, whilst HBQR-EN and BQR-EN methods have similar ALs. The BQR-BL method produced wider ALs. Upon looking at Simulations 2 and 3 in the presence of large outliers, the HBL method outperformed the quantile-based methods at the median level, whilst the quantile-based methods have comparable results. When the value of τ deviated from the median level, the proposed methods outperformed the existing methods significantly. In terms of AL, the proposed methods produced shortest ALs, whilst the BQR-BL produced ALs two times wider than the other methods. Upon looking at Simulations 4-6 where there were skewed & heavy-tailed noise with large outliers, heavy-tailed noise (Cauchy distribution) and multiple outliers, respectively, the proposed methods outperformed the existing methods significantly at varying quantile levels with the exception of the latter

having lower RMSEs in Simulation 4 at $\tau = 0.25$. The behaviour of AL followed the same pattern of those in Simulations 2 and 3. Finally, in terms of CP, the proposed methods have CPs in range of 86% and 97% for different scenarios at varying quantile levels, which are lower than those of the existing methods. By increasing the sample size, the CP would see an increase, since the sample size is one of the primary factors on the CP (Wilcox (2003)).

We reported boxplot performances for visualisation in Figures 3.5-3.7 at $\tau = 0.25$. Upon looking at Figure 3.5 for RMSE performances, all four quantile-based methods produced comparable boxplots for Simulation 1, yet the proposed methods have tighter boxplots and the existing methods produced more outliers in their boxplots for Simulation 2-6. Particularly, in Simulation 5, the BQR-BL method saw an increase in RMSE, as the sample size increased from $n = 100$ to $n = 200$, which is unusual. For MMAD performances, Figure 3.6 showed the similar behaviour, and whilst observing results of Simulation 5, the existing methods produced several outliers in their boxplots and the BQR-EN method had the unusual boxplot. This is also observed in Figure 3.7 for AL performances, and the proposed methods consistently produced lower median values in their boxplots. Similar performances are observed for $\tau = 0.75$ and the figures are provided in Appendix D.2 (see Figures D.9-D.11).

We also reported boxplot performances in Figures 3.8-3.10 at $\tau = 0.5$. Upon looking at Figure 3.8 for RMSE performances, all five methods produced similar boxplots for Simulations 1-3. For Simulation 4, it is noticeable that the existing methods produced tighter boxplots. The BQR-EN method saw an unusual boxplot for Simulation 5 and the BQR-BL method produced most outliers in its boxplots for all simulations. Unlike the HBL method, Simulation 6 saw an increase in RMSE for each quantile-based method, as the sample size increased from $n = 100$ to $n = 200$. Figure 3.9 observed the similar behaviour for MMAD performances. Upon looking at Figure 3.10 for AL performances, Simulation 1 observed similar boxplots for all five methods, yet the proposed methods consistently produced tighter boxplots and their corresponding median values are lower than those of the existing methods for Simulations 2-6. Whilst looking at Simulation 5, the BQR-EN method produced the unusual boxplot with large box size, and median value being close to its lower quartile.

In terms of sample sizes, when there is an increase in the sample size, all tables and figures mentioned above generally saw a decrease in error measures, including RMSE and MMAD, and AL of credible intervals. This is due to the fact that the more data is available, the better the model fitting is.

We also presented boxplots of posterior median of η in Figure 3.11 for the sample size of $n = 100$ at different quantile levels. For Simulation 1, both proposed methods produced the relatively large value of η, whilst the HBQR-BL method produced large box size and long whiskers in its boxplot in contrast to the HBQR-EN method having its tighter boxplot with shorter whiskers. On the other hand, for Simulations 2-6, the relatively small value of η was chosen by the proposed methods, whilst both saw tighter boxplots with shorter whiskers, and those of the HBQR-EN method produced more outliers. Upon looking at Simulation 6, the proposed methods produced boxplots with slightly large box size and longer whiskers like those in Simulation 1. Nevertheless, there is no noticeable difference between η values of both proposed methods, and the behaviour of η did not differ at varying quantile levels. Thus, η was adaptively chosen for different scenarios, which can also be observed in the HBL method (Kawakami and Hashimoto (2023)). Similar performances are observed for the sample size of $n = 200$ and the figure is provided in Appendix D.2 (see Figure D.12).

To summarise the results, it is acknowledged that the proposed methods have strengths and limitations. At the median level, the existing methods are fairly competitive with the proposed methods for some scenarios. When the error assumption is Gaussian either with or without outliers, the HBL method is preferable because the HBL method dampens the effect of outlying observations well under the Gaussian error assumption. However, when the noise is non-Gaussian, the proposed methods performed significantly better because they accommodate skewness and heavy tails well under the non-Gaussian noise assumption, whilst they are adaptive for various outliers with the support of tuning robustness parameter η. When the value of τ deviated from the median level, the proposed methods generally outperformed the existing quantile-based methods. Additionally, the AL of the proposed methods' credible intervals is narrower than those of the existing methods at varying quantile levels for each scenario. It is noteworthy that the BQR-BL method may be unstable in producing estimates due to having wider ALs, and the BQR-EN method performed poorly under the assumption of Cauchy noise, whilst the proposed methods are stable in all six scenarios. Therefore, the proposed methods consistently performed well, particularly for noises with contamination, skewness and/or heavy tails.

3.5 Real Data Analysis

The robustness and efficiency of the Bayesian Huberised regularised quantile regression models were demonstrated via the analysis of three benchmarking datasets: Prostate Cancer

Figure 3.12: Posterior medians and 95% credible intervals of the regression coefficients at different quantile levels ($\tau = 0.1, 0.5, 0.9$) for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN, applied to the Crime data.

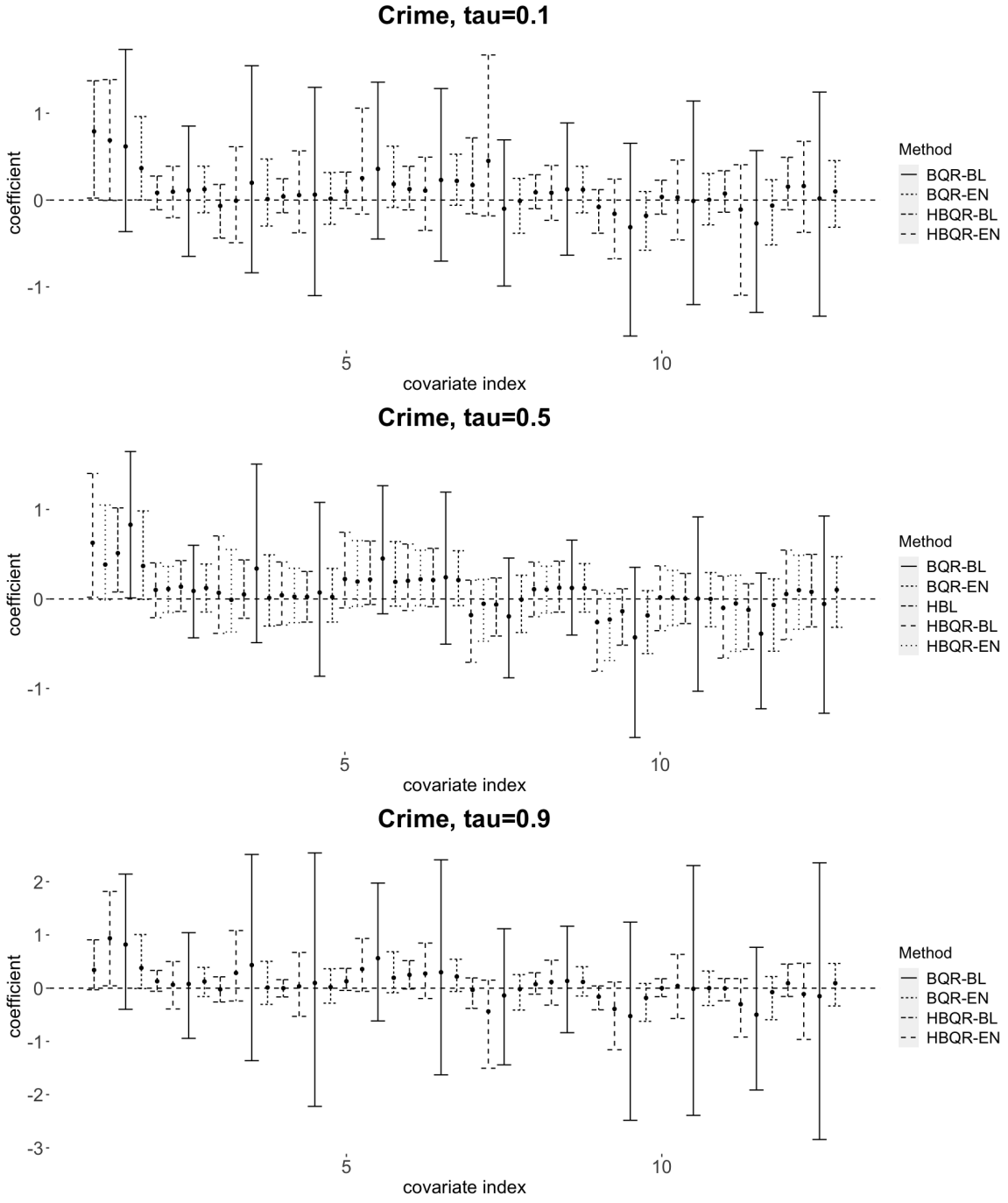


Figure 3.13: Posterior medians and 95% credible intervals of the regression coefficients at different quantile levels ($\tau = 0.5, 0.7, 0.9$) for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN, applied to the Prostate Cancer data.

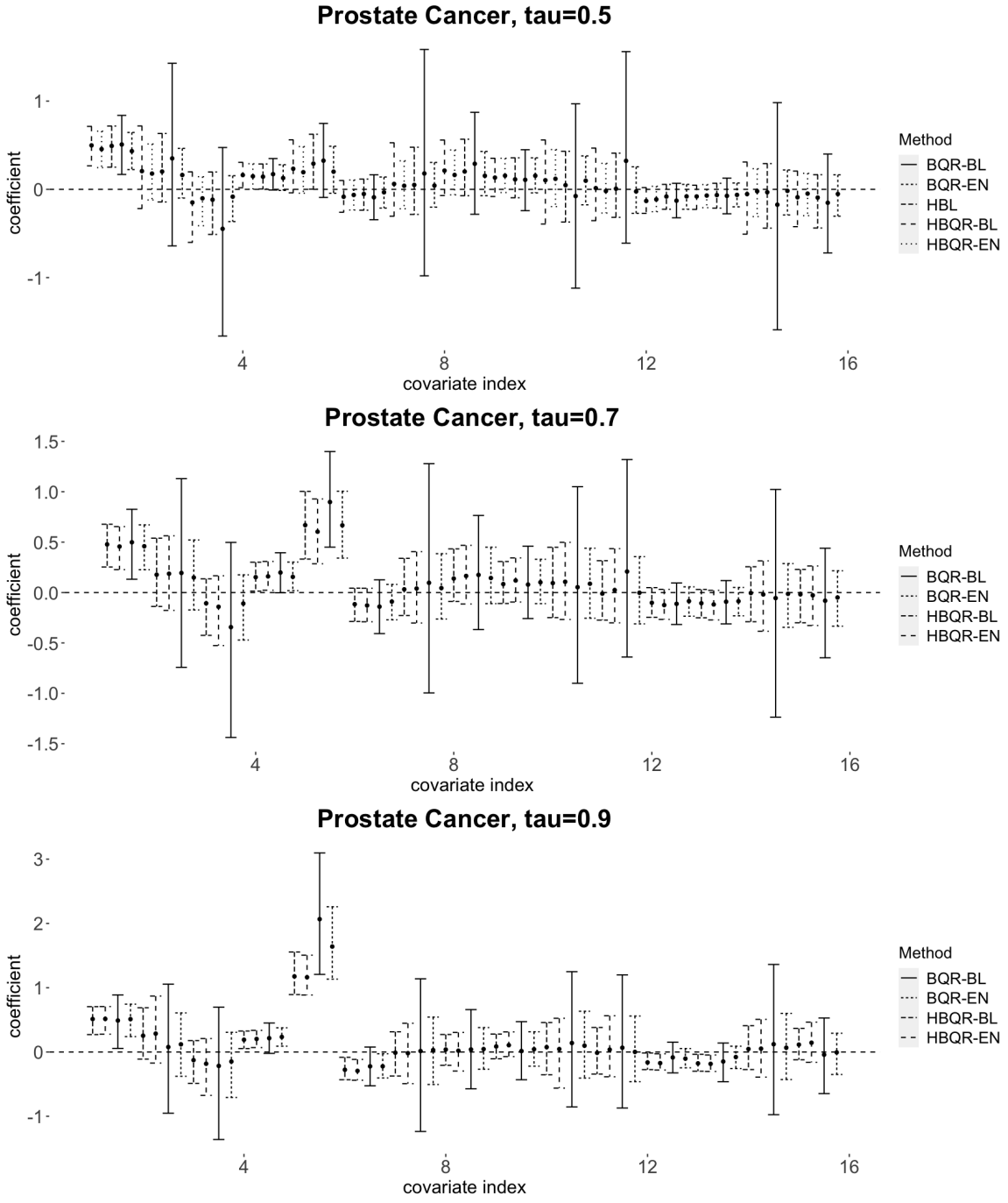


Figure 3.14: Posterior medians and 95% credible intervals of the regression coefficients at different quantile levels ($\tau = 0.1, 0.5, 0.9$) for HBQR-BL, HBQR-EN, HBL, BQR-BL and BQR-EN, applied to the Top Gear data.

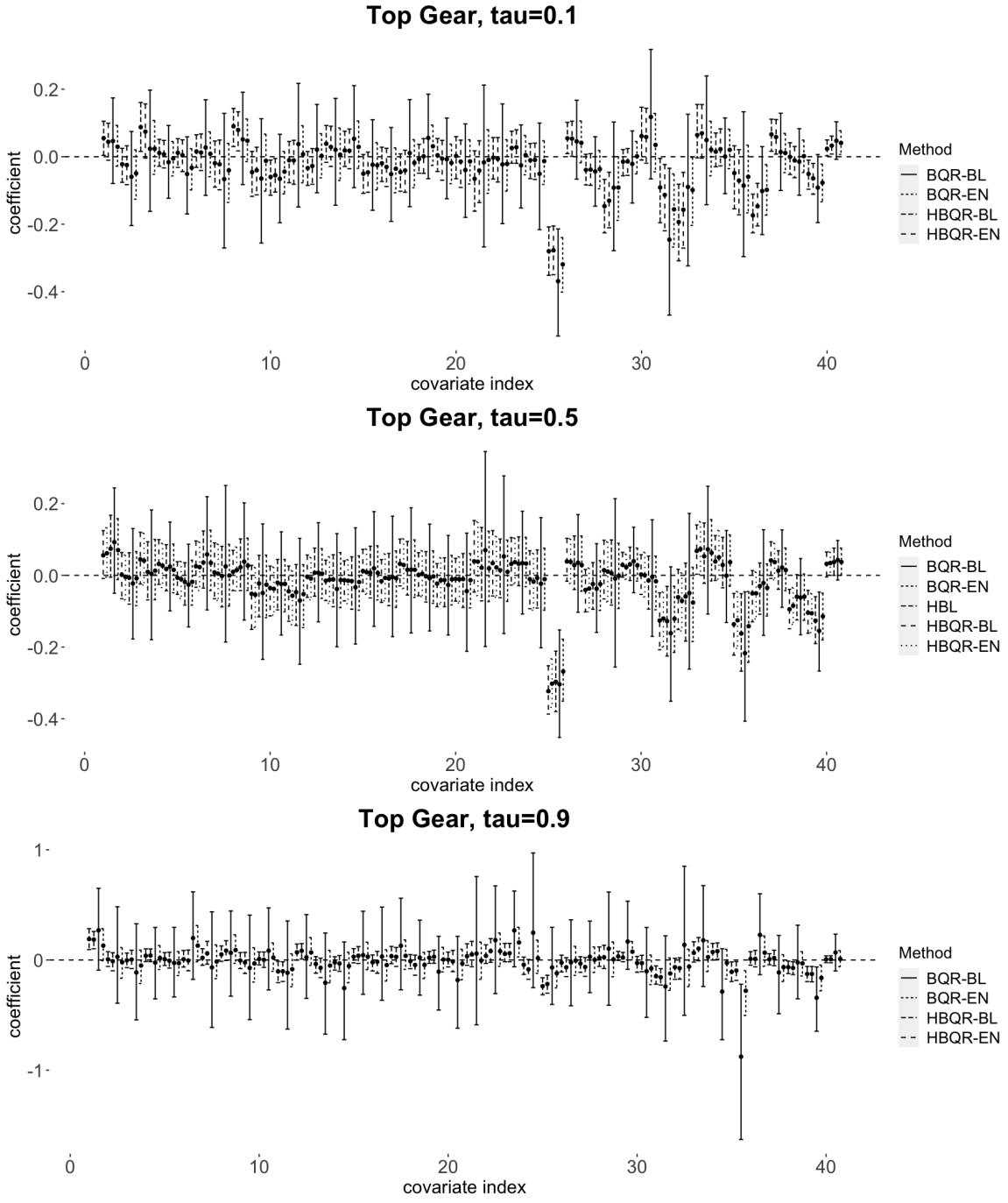


Table 3.7: MSPE, MAPE, MHPE with $\delta = 1.345$ and MedSPE for the Crime data, computed from 10-fold cross-validation.

	Methods	MSPE	MAPE	MedSPE	MHPE
$\tau = 0.1$	HBQR-BL	0.0226	0.1223	0.0044	0.0113
	HBQR-EN	0.0156	0.1034	0.0071	0.0078
	BQR-BL	0.0157	0.1054	0.0137	0.0078
	BQR-EN	0.0169	0.1285	0.0150	0.0085
$\tau = 0.5$	HBQR-BL	0.0123	0.0946	0.0067	0.0061
	HBQR-EN	0.0121	0.0938	0.0073	0.0061
	HBL	0.0304	0.1534	0.0173	0.0152
	BQR-BL	0.0192	0.1008	0.0081	0.0096
	BQR-EN	0.0250	0.1474	0.0185	0.0125
$\tau = 0.9$	HBQR-BL	0.0400	0.1439	0.0077	0.0200
	HBQR-EN	0.0278	0.1346	0.0079	0.0139
	BQR-BL	0.0453	0.1582	0.0106	0.0226
	BQR-EN	0.0401	0.1629	0.0146	0.0200

data, Crime data and Top Gear data. The Crime and Top Gear datasets have large outliers. For a better interpretation of the parameters and to put the predictors on the common scale, we standardised all the numerical predictors and response variables to have mean 0 and variance 1. Like in simulation studies, we also considered all the five methods of which we generated 10,000 posterior samples after discarding the first 5,000 posterior samples as a burn-in. Then we reported posterior medians of regression coefficients and their 95% credible intervals. For brevity, we dropped the names of predictors of the datasets and kept the corresponding number to indicate each predictor. For BQR-BL, BQR-EN, HBQR-BL and HBQR-EN, we set the quantile levels as $\tau \in \{0.1, 0.5, 0.9\}$ for the Crime and Top Gear datasets. We also chose $\tau \in \{0.5, 0.7, 0.9\}$ for the Prostate Cancer dataset like Li and Lin (2010), Alhamzawi et al. (2012) and Alhamzawi et al. (2019).

Since datasets may contain outliers, we adopted the following four criteria as measures of predictive accuracy: mean squared prediction error (MSPE), mean absolute prediction error (MAPE), mean Huber prediction error (MHPE) for $\delta = 1.345$ and median of squared prediction error (MedSPE) via the 10-fold cross validation. They are defined by MSPE $= 10^{-1}\sum_{j=1}^{10}(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})^2$, MAPE $= 10^{-1}\sum_{j=1}^{10}|\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)}|$, MHPE $= 10^{-1}\sum_{j=1}^{10} L_\delta^{\text{Huber}}$
$(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})$ and MedSPE $= \text{median}_{1\leq j\leq 10}(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})^2$, where $L_\delta^{\text{Huber}}(\cdot)$ is defined by Equation (1.1), $\hat{\boldsymbol{\beta}}^{(-j)}$ is the posterior median based on a dataset except for j^{th} validation set, and $\boldsymbol{y}_j$ and $\boldsymbol{X}_j$ are the response variables and design matrix based on the j^{th} validation set, respectively.

Table 3.8: MSPE, MAPE, MHPE with $\delta = 1.345$ and MedSPE for the Prostate Cancer data, computed from 10-fold cross-validation.

	Methods	MSPE	MAPE	MedSPE	MHPE
$\tau = 0.5$	HBQR-BL	1.1988	1.0378	1.2586	0.5977
	HBQR-EN	1.2481	1.0607	1.2175	0.6216
	HBL	1.3158	1.0906	1.3379	0.6541
	BQR-BL	1.2932	1.0814	1.3393	0.6434
	BQR-EN	1.3115	1.0882	1.3001	0.6519
$\tau = 0.7$	HBQR-BL	1.2473	1.0606	1.4409	0.6222
	HBQR-EN	1.2290	1.0513	1.4216	0.6131
	BQR-BL	1.2916	1.0809	1.5416	0.6423
	BQR-EN	1.2615	1.0644	1.4466	0.6283
$\tau = 0.9$	HBQR-BL	1.2471	1.0564	1.5681	0.6227
	HBQR-EN	1.1894	1.0245	1.4484	0.5943
	BQR-BL	1.3358	1.0868	1.2840	0.6631
	BQR-EN	1.2240	1.0338	1.3014	0.6102

3.5.1 Crime Dataset

The dataset was collected from the Statistical Abstract of the US for the 50 states and the District of Columbia (U.S. Census Bureau (2006)). This dataset was analysed in the book of Statistical Methods for the Social Sciences (Agresti and Finlay (1997)). The predictors are the number of murders per 100,000 people in the population, the percentage of the population living in Metropolitan areas, the percentage of the population who were white, the percentage of the population who were high school graduates or higher, the percentage of families living below the poverty level, and the percentage of families headed by a single parent (male householders with no wife present and with own children, or female householders with no husband present and with own children). The response of interest is the number of murders, forcible rapes, robberies, and aggravated assaults per 100,000 people in the population. In total, we have 51 observations and included squared variables, which resulted in 12 predictors in our models.

The posterior medians and 95% credible intervals of the regression coefficients based on the five methods are reported in Figure 3.12. From the figure, all the methods behaved similarly and the estimates were very close. The BQR-BL method produced wider credible intervals, which suggested that this method may be unstable in producing estimates. The similar performances are observed in $\tau = 0.1$ and $\tau = 0.9$. Table 3.7 also presents the predictive performance of the five methods for $\tau = 0.5$ and the four Bayesian quantile regression based methods for $\tau \in \{0.1, 0.9\}$. The proposed methods performed better than the existing robust

Table 3.9: MSPE, MAPE, MHPE with $\delta = 1.345$ and MedSPE for the Top Gear data, computed from 10-fold cross-validation.

	Methods	MSPE	MAPE	MedSPE	MHPE
$\tau = 0.1$	HBQR-BL	0.0288	0.1500	0.0196	0.0144
	HBQR-EN	0.0296	0.1505	0.0229	0.0148
	BQR-BL	0.0360	0.1736	0.0331	0.0180
	BQR-EN	0.0331	0.1605	0.0267	0.0166
$\tau = 0.5$	HBQR-BL	0.0127	0.0942	0.0064	0.0064
	HBQR-EN	0.0120	0.0905	0.0070	0.0060
	HBL	0.0110	0.0863	0.0055	0.0055
	BQR-BL	0.0183	0.1102	0.0101	0.0092
	BQR-EN	0.0110	0.0864	0.0063	0.0055
$\tau = 0.9$	HBQR-BL	0.0643	0.2309	0.0410	0.0322
	HBQR-EN	0.0843	0.2662	0.0628	0.0421
	BQR-BL	0.6942	0.7652	0.7337	0.3471
	BQR-EN	0.2290	0.4461	0.1790	0.1145

methods in both median and upper quantile levels. The HBL method produced relatively large error measures in the median case amongst the rest of methods. Looking at the lower quantile level ($\tau = 0.1$), MSPE, MAPE and MHPE suggested that HBQR-EN and BQR-BL performed better, whilst MedSPE suggested that both proposed methods performed better. In this case, they were very comparable.

3.5.2 Prostate Cancer Dataset

The dataset was from a prostate cancer study (Stamey et al. (1989)) and analysed by Zou and Hastie (2005). It is available in the R package 'bayesQR' (Benoit and Van den Poel (2017)). The predictors are eight clinical measures in men who were about to receive a radical prostatectomy: the logarithm of cancer volume, the logarithm of prostate weight, age, the logarithm of the amount of benign prostatic hyperplasia, seminal vesicle invasion, the logarithm of capsular penetration, the Gleason score and the percentage Gleason score 4 or 5. The response of interest is the logarithm of prostate-specific antigens. In total, we have 97 observations for the data and included squared variables, which resulted in 16 predictors in our models.

The posterior medians and 95% credible intervals of the regression coefficients based on the five methods are reported in Figure 3.13. From the figure, all the methods were comparable except the BQR-BL method that seemed to perform unstably due to its regression coefficients' wide credible intervals. The similar performances are also observed in $\tau = 0.7$ and

$\tau = 0.9$. Table 3.8 also presents the predictive performance of the five methods for $\tau = 0.5$ and the four Bayesian quantile regression based methods for $\tau \in \{0.7, 0.9\}$. The proposed methods performed better than the existing robust methods for $\tau = 0.5$ and $\tau = 0.7$. For $\tau = 0.9$, it revealed that the HBQR-EN and BQR-EN methods performed better except the existing methods performed better according to MedSPE.

3.5.3 Top Gear Dataset

The dataset used information on cars featuring on the website of the popular BBC television show Top Gear. It is available in the R package 'robustHD' (Alfons (2021)) and contained 242 observations on 29 numerical and categorical variables after removing the missing values. A description of the variables is provided in Table 3 of the paper (Alfons et al. (2016)). The response of interest is MPG (fuel consumption) and the remaining variables are predictors. For categorical variables, there were 4 binary variables and 12 variables with three levels. These 12 variables were assigned two dummy variables each. The resulting design matrix consisted of 12 numerical variables, 4 individual dummy variables, and 12 groups of two dummy variables each, giving a total of 40 predictors.

The posterior medians and 95% credible intervals of the regression coefficients based on the five methods are reported in Figure 3.14. From the figure, all the methods were comparable. Like for the Crime and Prostate Cancer datasets, the BQR-BL method produced wider credible intervals. The similar performances are also observed in $\tau = 0.1$ and $\tau = 0.9$. Table 3.9 also presents the predictive performance of the five methods for $\tau = 0.5$ and the four Bayesian quantile regression based methods for $\tau \in \{0.1, 0.9\}$. All the methods were comparable in the median case ($\tau = 0.5$) where both HBL and BQR-EN have the lowest error measures. Looking at the extreme quantile levels ($\tau = 0.1, 0.9$), the proposed methods significantly outperformed the BQR-BL and BQR-EN methods especially at the upper quantile level where the existing robust methods performed worse than the proposed methods. Furthermore, BQR-BL had the highest error measures in all cases.

3.6 Chapter Summary

In this chapter, we have presented the Bayesian Huberised regularisation. We proposed the asymmetric Huberised loss function and its corresponding probability density function that led to the scale mixture of normal distribution with exponential and GIG mixing densities.

This resulted in fully Bayesian hierarchical models for quantile regression and its Gibbs sampling algorithm with the approximate Gibbs sampler for the data-dependent estimation of the robustness parameter. We have proven theoretically that the proposed Bayesian models yielded a good posterior propriety and unimodality in their joint posterior density with conditional prior for the regression coefficients. Simulation studies and real data examples showed that the proposed methods are effective in predictive accuracy and their robustness is evident under a wide range of scenarios. In many situations, the proposed methods outperformed the existing Bayesian regularised quantile regression methods especially at the extreme quantile levels and under the non-Gaussian error assumption, since they accommodate skewness and heavy tails, whilst they are adaptive for various outliers with the support of tuning robustness parameter.. Our proposed methods have proven to be robust in obtaining valuable results.

Chapter 4

Variational Bayesian Huberised Adaptive Lasso

Chapter 3 has utilised the MCMC method that has proven to be successful in formulating full Bayesian probabilistic models when posterior distributions are analytically intractable to be computed directly. As sample size increases, the computational cost of the MCMC method becomes more expensive and burdensome. This motivates the VB method as an alternative. This chapter proposes a novel VB regularisation and its extension to quantile regression, including VB Huberised Lasso quantile regression and VB Huberised adaptive Lasso quantile regression. The full CAVI algorithms are presented. Via various simulation studies and a real data example, the comparative studies with the MCMC method have shown that the proposed methods performed much faster than the MCMC method, particularly the algorithm of VB Huberised adaptive Lasso quantile regression, whilst obtaining similar statistical results.

4.1 Introduction

MCMC methods are a common approach for Bayesian probabilistic models where posterior distributions cannot be computed directly. As discussed in the previous chapters, some Bayesian regularisation methods have been developed to simultaneously estimate model parameters and consistently select variables in a high-dimensional setting by imposing various priors on parameters. However, they are not the best choices because they take a great deal of computational time and are not scalable. The VB method is an alternative approach, which has attracted much attention in recent years, such as Chen et al. (2016), Alves et al.

(2021), Yi and Tang (2022), and Wang et al. (2023), amongst others.

The VB method transforms the problem of probabilistic inference into the optimisation problem. It approximates the exact posterior distribution based on a family of tractable densities. The KL divergence is used as a measure from the approximate posterior density to the posterior distribution. VI originated from the free energy of statistical physics, such as the mean-field free energy and Bethe/Kikuchi free energies (Saul et al. (1996), Yedidia et al. (2000), and Yedidia et al. (2001)). It then extended to Bayesian statistics and machine learning (Jordan et al. (1999), Beal (2003), Beal and Ghahramani (2003), and Beal and Ghahramani (2004), amongst others). The comprehensive reviews of the VB method are provided in Fox and Roberts (2012), Ostwald et al. (2014), Blei et al. (2017), Tran et al. (2021), and Wu and Tang (2021). Blei et al. (2017) noted that the accuracy of VI has not yet been thoroughly studied and many open questions are still there to be answered. Nonetheless, the VB method is an attractive option for approximate inference due to its sound theoretical foundation and high convergence rate (Jordan et al. (1999), and Minka (2005)).

Therefore, in this chapter, we incorporate the VB method and Bayesian Huberised Lasso (Kawakami and Hashimoto (2023)) to propose a novel VB Huberised regularisation using the Lasso (Tibshirani (1996)) and adaptive Lasso (Zou (2006)) for robust regression for a fast-computational and high-dimensional problem. Along with Bayesian Huberised regularisation, the asymmetric Huberised loss function, proposed in Chapter 3, is utilised to present both Bayesian Huberised Lasso quantile regression and Bayesian Huberised adaptive Lasso quantile regression, due to the quantile and normal scale-mixture properties of its probability distribution. The MCMC method is replaced with the VB method where the mean-field VB approach is used to formulate approximate densities in place of exact posterior distributions. Because the tuning robustness parameter does not belong to the conjugate prior family, the approximate Gibbs sampler is replaced with the Laplace VI method (Wang and Blei (2013)) that uses the Laplace approximation. This retains the advantage of the data-dependent estimation of the tuning robustness parameter. Without compromising the accuracy of parameter estimation, the VB method retains the advantages of Bayesian inference, including parameter uncertainty, priori knowledge and families of tractable densities, whilst enjoys low computational cost and fast convergence rate. The parameter estimation, efficiency and robustness of the proposed algorithms and its computational performance are demonstrated in simulation studies followed by real data analysis, whilst conducting comparative analysis with the MCMC method.

Section 4.2 presents the fully hierarchical models for Bayesian Huberised Lasso quantile regression and Bayesian Huberised adaptive Lasso quantile regression. Then we derive the VB algorithms using mean-field and Laplace methods in Section 4.3. The simulation studies are conducted to investigate the numerical performance of the proposed algorithms in Section 4.4 followed by the Boston Housing data example in Section 4.5. Section 4.6 provides the discussion of this chapter.

4.2 Bayesian Huberised Lasso Quantile Regression and its Extension

In this section, we revisit the Bayesian Huberised Lasso quantile regression from Chapter 3 and reformulate the regression problem in order to keep the notation consistent for developing the VB algorithm where the intercept term is to be estimated independently from the regression coefficient vector. We will also introduce the new Bayesian Huberised adaptive Lasso quantile regression.

4.2.1 Bayesian Huberised Lasso

We consider the following Huberised regularised quantile regression model,

$$y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n,$$

where β_0 is the intercept term, $\boldsymbol{\beta}$ is the vector of unknown coefficients, y_i is the response variable, $\boldsymbol{x}_i$ is the regression coefficient vector and ϵ follows the density function from Equation (3.2). We consider the Bayesian Huberised Lasso, that is given in Equation (3.4).

By using the scale mixture of normal representation of Laplace distribution (Andrews and Mallows (1974)) for the Bayesian Huberised Lasso and that of Theorem 3.1, we present the following hierarchical model:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{X}, \beta_0, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{v} \sim \mathrm{N}(\beta_0, \boldsymbol{X}\boldsymbol{\beta} + (1-2\tau)\boldsymbol{v}, \boldsymbol{V}),\\
&\sigma_i|\rho^2, \eta \sim \mathrm{GIG}\left(\frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right),\ v_i|\sigma_i \sim \mathrm{Exp}\left(\frac{\tau(1-\tau)}{2\sigma_i}\right),\ i = 1, \ldots, n,\\
&\beta_j|s_j, \rho^2 \sim \mathrm{N}(0, \rho^2 s_j),\ s_j|\lambda^2 \sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right),\ j = 1, \ldots, k,\\
&\rho^2 \sim \pi(\rho^2) \propto \frac{1}{\rho^2},\ \lambda^2 \sim \mathrm{Gamma}(a, b),\ \eta \sim \mathrm{Gamma}(c, d),\ \beta_0 \propto 1,
\end{aligned}$$

The prior specification follows in the same way from Chapter 3 and we assume a flat prior for β_0. Whilst the Bayesian Huberised Lasso enjoys robustness and variable shrinkage properties, it may experience overfitting like the Bayesian Lasso (Park and Casella (2008)). Zou (2006) proposed adaptive Lasso, which extended the Lasso approach (Tibshirani (1996)) allowing different penalisation parameters for different regression coefficients.

4.2.2 Bayesian Huberised Adaptive Lasso

We extend Equation (3.4) to include the adaptive Lasso penalty resulting in the Bayesian Huberised adaptive Lasso:

$$\pi(\boldsymbol{\beta}|\rho^2, \boldsymbol{\lambda}) = \prod_{j=1}^{k} \frac{\lambda_j}{2\sqrt{\rho^2}} \exp\left\{-\frac{\lambda_j|\boldsymbol{\beta}_j|}{\sqrt{\rho^2}}\right\}, \tag{4.1}$$

where λ_j is the penalisation parameter being assigned to each regression coefficient.

By using the scale mixture of normal representation of Laplace distribution (Andrews and Mallows (1974)), the Bayesian Huberised adaptive Lasso can be expressed as

$$\boldsymbol{\beta}|\boldsymbol{s}, \rho^2 \sim \mathrm{N}(\mathbf{0}, \rho^2\boldsymbol{\Lambda}), \quad s_j|\lambda_j^2 \sim \mathrm{Exp}\left(\frac{\lambda_j^2}{2}\right), \quad j = 1, \ldots, k,$$

where $\boldsymbol{s} = (s_1, \ldots, s_k)^T$ and $\boldsymbol{\Lambda} = \mathrm{diag}\,(s_1, \ldots, s_k)$.

Thus, we present the following hierarchical model with the Bayesian Huberised adaptive Lasso:

$$\begin{aligned}
&\boldsymbol{y}|\boldsymbol{X}, \beta_0, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{v} \sim \mathrm{N}(\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + (1-2\tau)\boldsymbol{v}, \boldsymbol{V}),\\
&\sigma_i|\rho^2, \eta \sim \mathrm{GIG}\left(\frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right),\ v_i|\sigma_i \sim \mathrm{Exp}\left(\frac{\tau(1-\tau)}{2\sigma_i}\right),\ i = 1, \ldots, n,\\
&\beta_j|s_j, \rho^2 \sim \mathrm{N}(0, \rho^2 s_j),\ s_j|\lambda_j^2 \sim \mathrm{Exp}\left(\frac{\lambda_j^2}{2}\right),\ \lambda_j^2 \sim \mathrm{Gamma}(a_j, b_j),\ j = 1, \ldots, k,\\
&\rho^2 \sim \pi(\rho^2) \propto \frac{1}{\rho^2},\ \eta \sim \mathrm{Gamma}(c, d),\ \beta_0 \propto 1,
\end{aligned}$$

This two-level prior distribution of regression coefficients provides flexible shrinkage weights for regression coefficients. Instead of employing the usual MCMC method like in Chapter 3, the VB approach is used, since it has a fast computational speed and retains the efficiency of parameter estimation without compromising the accuracy.

4.3 Variational Inference

This section follows the concept of VI that is fully explained in Chapter 1, and the notation is modified such that it is appropriate for a given model. For each specified model, the CAVI algorithm is developed and the evidence lower bound (ELBO), required for the convergence criterion, is also formulated. The detailed derivation of ELBO, including all the variational densities, is provided in Appendix C.

4.3.1 Bayesian Huberised Lasso Quantile Regression

Following from Chapter 1, the vector including latent variables and parameters denoted by Θ is represented in this subsection by the vector $\Theta = (\beta_0, \boldsymbol{\beta}, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{\sigma}, \rho^2, \eta, \lambda^2)$.

Following the optimisation problem (Equation (1.21), we express the approximate posterior distribution by factorisation,

$$q(\Theta) = \prod_{l=1}^{N} q(\theta_l) = \left(\prod_{j=1}^{k} q(\beta_j) q(s_j) \right) \left(\prod_{i=1}^{n} q(v_i) q(\sigma_i) \right) q(\rho^2) q(\eta) q(\lambda^2) q(\beta_0) \approx p(\Theta | \boldsymbol{y}).$$

For the sake of brevity, we denote $\mathbb{E}[\theta_l]$ as the expectation of the variable θ_l about the optimal variational posterior distribution $q^*(\theta_l)$. According to Equation (1.21), the optimal variational density $q^*(\boldsymbol{\beta})$ is

$$q^*(\boldsymbol{\beta}) \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}), \tag{4.2}$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^T \mathbb{E}[\boldsymbol{V}^{-1}] \boldsymbol{X} + \mathbb{E}\left[\frac{1}{\rho^2}\right] \mathbb{E}[\boldsymbol{\Lambda}^{-1}] \right), \\
\boldsymbol{\mu}_{\boldsymbol{\beta}} &= \boldsymbol{\Sigma} \boldsymbol{X}^T \left(\mathbb{E}[\boldsymbol{V}^{-1}](\boldsymbol{y} - \mathbb{E}[\beta_0]) - \frac{1-2\tau}{4} \mathbb{E}\left[\frac{1}{\sigma_i}\right] \right), \\
\mathbb{E}[\boldsymbol{V}^{-1}] &= \mathrm{diag}\left(\frac{1}{4} \mathbb{E}\left[\frac{1}{\boldsymbol{v}}\right] \mathbb{E}\left[\frac{1}{\boldsymbol{\sigma}}\right] \right), \\
\mathbb{E}[\boldsymbol{\Lambda}^{-1}] &= \mathrm{diag}\left(\mathbb{E}\left[\frac{1}{\boldsymbol{s}}\right] \right).
\end{aligned}$$

The mean estimator $\boldsymbol{\beta}^{\mathrm{mean}}$ of $\boldsymbol{\beta}$ for the variational density $q^*(\boldsymbol{\beta})$ is given by $\boldsymbol{\beta}^{\mathrm{mean}} = \boldsymbol{\mu}_{\boldsymbol{\beta}}$.

In the similar way, the optimal variational density $q^*(\boldsymbol{\sigma})$ is

$$q^*(\sigma_i) \sim \text{GIG}\left(0, \hat{a}_{\sigma_i}, \hat{b}_{\sigma_i}\right), \quad i = 1, \ldots, n, \tag{4.3}$$

where

$$\begin{aligned}
\hat{a}_{\sigma_i} &= \mathbb{E}[\eta]\mathbb{E}\left[\frac{1}{\rho^2}\right], \\
\hat{b}_{\sigma_i} &= \frac{1}{4}\mathbb{E}\left[\frac{1}{v_i}\right]\left((y_i - \mathbb{E}[\beta_0] - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}])^2 + \text{Var}(\beta_0) + \text{Tr}(\boldsymbol{x}_i\boldsymbol{x}_i^T\text{Var}(\boldsymbol{\beta}))\right) \\
&\quad + \left(\tau(1-\tau) + \frac{(1-2\tau)^2}{4}\right)\mathbb{E}[v_i] - \frac{1-2\tau}{2}\left(y_i - \mathbb{E}[\beta_0] - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}]\right) \\
&\quad + \tau(1-\tau)\mathbb{E}[v_i] + \mathbb{E}[\eta]\mathbb{E}[\rho^2].
\end{aligned}$$

The mean estimator σ_i^{mean} of σ_i for the variational density $q^*(\sigma_i)$ is given by $\sigma_i^{\text{mean}} = \sqrt{\hat{b}_{\sigma_i}/\hat{a}_{\sigma_i}}K_1\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)/K_0\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)$ for $i = 1, \ldots, n$, where $K_\nu(\cdot)$ is the modified Bessel function of the second kind at index ν.

The optimal variational density $q^*(\boldsymbol{v})$ can be obtained similarly as follows:

$$q^*(v_i) \sim \text{GIG}\left(\frac{1}{2}, \hat{a}_{v_i}, \hat{b}_{v_i}\right), \quad i = 1, \ldots, n, \tag{4.4}$$

where

$$\begin{aligned}
\hat{a}_{v_i} &= \mathbb{E}\left[\frac{1}{\sigma_i}\right]\left(\frac{(1-2\tau)^2}{4} + \tau(1-\tau)\right), \\
\hat{b}_{v_i} &= \frac{1}{4}\mathbb{E}\left[\frac{1}{\sigma_i}\right]\left((y_i - \mathbb{E}[\beta_0] - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}])^2 + \text{Var}(\beta_0) + \text{Tr}(\boldsymbol{x}_i\boldsymbol{x}_i^T\text{Var}(\boldsymbol{\beta}))\right).
\end{aligned}$$

The mean estimator v_i^{mean} of v_i for the variational density $q^*(v_i)$ is given by $v_i^{\text{mean}} = \sqrt{\hat{b}_{v_i}/\hat{a}_{v_i}}K_{3/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)/K_{1/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)$ for $i = 1, \ldots, n$.

Similarly, the optimal variational density $q^*(\rho^2)$ can be expressed as

$$q^*(\rho^2) \sim \text{GIG}\left(-n - \frac{k}{2}, \hat{a}_{\rho^2}, \hat{b}_{\rho^2}\right), \tag{4.5}$$

where

$$\hat{a}_{\rho^2} = \mathbb{E}[\eta] \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{\sigma_i}\right],$$
$$\hat{b}_{\rho^2} = \mathbb{E}[\eta] \sum_{i=1}^{n} \mathbb{E}[\sigma_i] + \sum_{j=1}^{k} \mathbb{E}[s_j]\mathbb{E}[\beta_j^2].$$

The mean estimator $(\rho^2)^{\text{mean}}$ of ρ^2 for the variational density $q^*(\rho^2)$ is given by $(\rho^2)^{\text{mean}} = \sqrt{\hat{b}_{\rho^2}/\hat{a}_{\rho^2}} K_{-n-k/2+1}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right) / K_{-n-k/2}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)$.

The optimal variational density $q^*(\boldsymbol{s})$ can also be expressed as

$$q^*(s_j) \sim \text{GIG}\left(\frac{1}{2}, \hat{a}_{s_j}, \hat{b}_{s_j}\right), \quad j = 1, \ldots, k, \tag{4.6}$$

where $\hat{a}_{s_j} = \mathbb{E}[\lambda^2]$ and $\hat{b}_{s_j} = \mathbb{E}[1/\rho^2]\mathbb{E}[\beta_j^2]$ for $j = 1, \ldots, k$. The mean estimator s_j^{mean} of s_j for the variational density $q^*(s_j)$ is given by
$s_j^{\text{mean}} = \sqrt{\hat{b}_{s_j}/\hat{a}_{s_j}} K_{3/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right) / K_{1/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)$ for $j = 1, \ldots, k$.
The optimal variational density $q^*(\lambda^2)$ is

$$q^*(\lambda^2) \sim \text{Gamma}\left(\hat{a}_{\lambda^2}, \hat{b}_{\lambda^2}\right), \tag{4.7}$$

where $\hat{a}_{\lambda^2} = a + k$ and $\hat{b}_{\lambda^2} = b + 1/2 \sum_{j=1}^{k} \mathbb{E}[s_j]$. The mean estimator $(\lambda^2)^{\text{mean}}$ of λ^2 for the variational density $q^*(\lambda^2)$ is given by $(\lambda^2)^{\text{mean}} = \hat{a}_{\lambda^2}/\hat{b}_{\lambda^2}$.

Finally, the optimal variational density $q^*(\beta_0)$ is

$$q^*(\beta_0) \sim \text{N}\left(\mu_{\beta_0}, \sigma_{\beta_0}^2\right), \tag{4.8}$$

where

$$\sigma_{\beta_0}^2 = \left(\frac{1}{4}\sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{\sigma_i}\right]\mathbb{E}\left[\frac{1}{v_i}\right]\right)^{-1},$$
$$\mu_{\beta_0} = \left(\sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{\sigma_i}\right]\left(\mathbb{E}\left[\frac{1}{v_i}\right](y_i - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}]) - (1-2\tau)\right)\right)\left(\sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{\sigma_i}\right]\mathbb{E}\left[\frac{1}{v_i}\right]\right)^{-1}.$$

The mean estimator β_0^{mean} of β_0 for the variational density $q^*(\beta_0)$ is given by $\beta_0^{\text{mean}} = \mu_{\beta_0}$.

The expected values involved in the definition of the above optimal variational densities are

computed as follows.

$$\mathbb{E}[\boldsymbol{\beta}] = \boldsymbol{\mu}_{\boldsymbol{\beta}}, \quad \mathrm{Var}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \quad \mathbb{E}[\beta_j^2] = [\mu_{\boldsymbol{\beta}}]_j + [\boldsymbol{\Sigma}_{\boldsymbol{\beta}}]_{jj},$$

$$\mathbb{E}[\sigma_i] = \sqrt{\frac{\hat{b}_{\sigma_i}}{\hat{a}_{\sigma_i}}} \frac{K_1\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)}{K_0\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)}, \mathbb{E}\left[\frac{1}{\sigma_i}\right] = \sqrt{\frac{\hat{a}_{\sigma_i}}{\hat{b}_{\sigma_i}}} \frac{K_{-1}\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)}{K_0\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)},$$

$$\mathbb{E}[v_i] = \sqrt{\frac{\hat{b}_{v_i}}{\hat{a}_{v_i}}} \frac{K_{3/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)}{K_{1/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)}, \quad \mathbb{E}\left[\frac{1}{v_i}\right] = \sqrt{\frac{\hat{a}_{v_i}}{\hat{b}_{v_i}}} \equiv \sqrt{\frac{\hat{a}_{v_i}}{\hat{b}_{v_i}}} \frac{K_{3/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)}{K_{1/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)} - \frac{1}{\hat{b}_{v_i}},$$

$$\mathbb{E}[\rho^2] = \sqrt{\frac{\hat{b}_{\rho^2}}{\hat{a}_{\rho^2}}} \frac{K_{-n-k/2+1}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)}{K_{-n-k/2}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)},$$

$$\mathbb{E}\left[\frac{1}{\rho^2}\right] = \sqrt{\frac{\hat{a}_{\rho^2}}{\hat{b}_{\rho^2}}} \frac{K_{-n-k/2-1}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)}{K_{-n-k/2}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)} \equiv \sqrt{\frac{\hat{a}_{\rho^2}}{\hat{b}_{\rho^2}}} \frac{K_{-n-k/2+1}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)}{K_{-n-k/2}\left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}}\right)} + \frac{2n+k}{\hat{b}_{\rho^2}},$$

$$\mathbb{E}[s_j] = \sqrt{\frac{\hat{b}_{s_j}}{\hat{a}_{s_j}}} \frac{K_{3/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)}{K_{1/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)}, \quad \mathbb{E}\left[\frac{1}{s_j}\right] = \sqrt{\frac{\hat{a}_{s_j}}{\hat{b}_{s_j}}} \equiv \sqrt{\frac{\hat{a}_{s_j}}{\hat{b}_{s_j}}} \frac{K_{3/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)}{K_{1/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)} - \frac{1}{\hat{b}_{s_j}},$$

$$\mathbb{E}[\lambda^2] = \frac{\hat{a}_{\lambda^2}}{\hat{b}_{\lambda^2}}, \quad \mathbb{E}[\beta_0] = \mu_{\beta_0}, \quad \mathrm{Var}(\beta_0) = \sigma^2_{\beta_0}, \quad \mathbb{E}[\beta_0^2] = \mu^2_{\beta_0} + \sigma^2_{\beta_0},$$

for $j = 1, \ldots, k$ and $i = 1, \ldots, n$. The symbol '$\equiv$' indicates that one formula is equivalent to another.

The derivation of the optimal variational density for η (Equation (4.13)) will be detailed in Section 4.3.3. Based on the variational densities defined in Equations (4.2)-(4.8) and (4.13), we update the ELBO through CAVI algorithm. The CAVI algorithm is summarised in Algorithm 4.3.1.

4.3.2 Bayesian Huberised Adaptive Lasso Quantile Regression

Let $\Theta = (\beta_0, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma}, \rho^2, \eta, \boldsymbol{s}, \boldsymbol{\lambda}^2)$. Based on the hierarchical model, we express the approximate posterior distribution by factorisation,

$$q(\Theta) = \prod_{i=1}^{N} q(\theta_i) = \left(\prod_{j=1}^{k} q(\beta_j) q(s_j) q(\lambda_j^2)\right) \left(\prod_{i=1}^{n} q(v_i) q(\sigma_i)\right) q(\rho^2) q(\eta) q(\beta_0) \approx p(\Theta|\boldsymbol{y}).$$

Algorithm 4.3.1 Variational inference for Bayesian Huberised Lasso quantile regression

Step 1. Initialise the variational densities $q^*(\boldsymbol{\beta})$, $q^*(\sigma_i)$, $q^*(v_i)$ $(i = 1, \ldots, n)$, $q^*(\rho^2)$, $q^*(\eta)$, $q^*(s_j)$ $(j = 1, \ldots, k)$, $q^*(\lambda^2)$, $q^*(\beta_0)$.
while ELBO does not reach convergence criterion **do**
  Step 2.
  for $i = 1, \ldots, n$ **do**
    update $q^*(\sigma_i)$ according to the density (Equation (4.3));
    update $q^*(v_i)$ according to the density (Equation (4.4));
  end for
  Step 3. update $q^*(\boldsymbol{\beta})$ according to the density (Equation (4.2));
  Step 4. update $q^*(\beta_0)$ according to the density (Equation (4.8));
  Step 5. update $q^*(\rho^2)$ according to the density (Equation (4.5));
  Step 6. update $q^*(\eta)$ according to the density (Equation (4.13));
  Step 7. update $q^*(\lambda^2)$ according to the density (Equation (4.7));
  Step 8.
  for $j = 1, \ldots, k$ **do**
    update $q^*(s_j)$ according to the density (Equation (4.6));
  end for
  Step 9. Compute ELBO;
end while
Step 10. Return $q^*(\boldsymbol{\beta})$, $q^*(\boldsymbol{\sigma})$, $q^*(\boldsymbol{v})$, $q^*(\rho^2)$, $q^*(\eta)$, $q^*(\boldsymbol{s})$, $q^*(\lambda^2)$, $q^*(\beta_0)$, $\boldsymbol{\beta}^*$, $\boldsymbol{\sigma}^*$, $\boldsymbol{v}^*$, ρ^{2*}, η^*, $\boldsymbol{s}^*$, λ^{2*}, β_0^*.

Algorithm 4.3.2 Variational inference for Bayesian Huberised adaptive Lasso quantile regression

Step 1. Initialise the variational densities $q^*(\boldsymbol{\beta})$, $q^*(\sigma_i)$, $q^*(v_i)$ $(i = 1, \ldots, n)$, $q^*(\rho^2)$, $q^*(\eta)$, $q^*(s_j)$, $q^*(\lambda_j^2)$ $(j = 1, \ldots, k)$, $q^*(\beta_0)$.
while ELBO does not reach convergence criterion **do**
  Step 2. Do Step 2-6 of Algorithm 4.3.1;
  Step 3.
  for $j = 1, \ldots, k$ **do**
    update $q^*(s_j)$ according to the density (Equation (4.9));
    update $q^*(\lambda_j^2)$ according to the density (Equation (4.10));
  end for
  Step 4. Compute ELBO;
end while
Step 5. Return $q^*(\boldsymbol{\beta})$, $q^*(\boldsymbol{\sigma})$, $q^*(\boldsymbol{v})$, $q^*(\rho^2)$, $q^*(\eta)$, $q^*(\boldsymbol{s})$, $q^*(\boldsymbol{\lambda}^2)$, $q^*(\beta_0)$, $\boldsymbol{\beta}^*$, $\boldsymbol{\sigma}^*$, $\boldsymbol{v}^*$, ρ^{2*}, η^*, $\boldsymbol{s}^*$, $\boldsymbol{\lambda}^{2*}$, β_0^*.

Then we can find the solution according to Equation (1.21) and most of the variational parameters follow in the same way as in Section 4.3.1. However, we need to compute the optimal variational densities for $\boldsymbol{s}$ and $\boldsymbol{\lambda}^2$, which are as follows.

The optimal variational density $q^*(\boldsymbol{s})$ is

$$q^*(s_j) \sim \text{GIG}\left(\frac{1}{2}, \hat{a}_{s_j}, \hat{b}_{s_j}\right), \quad j = 1, \ldots, k, \tag{4.9}$$

where $\hat{a}_{s_j} = \mathbb{E}[\lambda_j^2]$ and $\hat{b}_{s_j} = \mathbb{E}[1/\rho^2]\mathbb{E}[\beta_j^2]$ for $j = 1, \ldots, k$. The mean estimator s_j^{mean} of s_j for the variational density $q^*(s_j)$ is given by
$s_j^{\text{mean}} = \sqrt{\hat{b}_{s_j}/\hat{a}_{s_j}} K_{3/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right) / K_{1/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)$ for $j = 1, \ldots, k$.
The optimal variational density $q^*(\boldsymbol{\lambda}^2)$ is

$$q^*(\lambda_j^2) \sim \text{Gamma}\left(\hat{a}_{\lambda_j^2}, \hat{b}_{\lambda_j^2}\right), \quad j = 1, \ldots, k, \tag{4.10}$$

where $\hat{a}_{\lambda_j^2} = a + 1$ and $\hat{b}_{\lambda_j^2} = b + 1/2\mathbb{E}[s_j]$. The mean estimator $(\lambda_j^2)^{\text{mean}}$ of λ_j^2 for the variational density $q^*(\lambda_j^2)$ is given by $(\lambda_j^2)^{\text{mean}} = \hat{a}_{\lambda_j^2}/\hat{b}_{\lambda_j^2}$.

The CAVI algorithm is summarised in Algorithm 4.3.2.

4.3.3 Laplace Approximation for η

We have the following coordinate update for η,

$$\begin{aligned} q(\eta) &= \frac{1}{K_{3/2}(\eta)^n} \exp\left\{-\frac{\eta}{2}\sum_{i=1}^{n}\left(\mathbb{E}\left[\frac{1}{\rho^2}\right]\mathbb{E}[\sigma_i] + \mathbb{E}[\rho^2]\mathbb{E}\left[\frac{1}{\sigma_i}\right]\right)\right\} \eta^{c-1}\exp\{-\eta d\} \\ &= \exp\{h(\eta)\}, \end{aligned} \tag{4.11}$$

where $h(\eta) = -n\log K_{3/2}(\eta) + (c-1)\log\eta - \eta\left(1/2\sum_{i=1}^{n}\left(\mathbb{E}[1/\rho^2]\mathbb{E}[\sigma_i] + \mathbb{E}[\rho^2]\mathbb{E}[1/\sigma_i]\right) + d\right)$. Since Equation (4.11) does not belong to a family of tractable densities, we cannot update $q(\eta)$ using Equation (4.11). The remedy for this issue is to use the Laplace VI method (Wang and Blei (2013)), which is alternative to the mean-field VI method. It is useful for non-conjugate models that cannot be normalised in closed form, such as Equation (4.11).

Since $h(\eta)$ is non-conjugate and twice-differentiable, we approximate the update by taking a second-order Taylor approximation of $h(\eta)$ around its maximum. The Taylor approximation

for $h(\eta)$ around $\hat{\eta}$ is

$$h(\eta) \approx h(\hat{\eta}) + \nabla h(\hat{\eta})(\eta - \hat{\eta}) + \frac{1}{2}\nabla^2 h(\hat{\eta})(\eta - \hat{\eta})^2, \tag{4.12}$$

where $\nabla h(\hat{\eta}) = dh(\eta)/d\eta$ evaluated at $\hat{\eta}$, $\nabla^2 h(\hat{\eta}) = d^2h(\eta)/d\eta^2$ evaluated at $\hat{\eta}$ and $\hat{\eta}$ is the maximum a posterior (MAP) of $q(\eta)$, found by maximising $h(\eta)$. The derivations of $\nabla h(\hat{\eta})$ and $\nabla^2 h(\hat{\eta})$ are as follows.

$$\nabla h(\hat{\eta}) = -n\frac{d}{d\eta}\log K_{3/2}(\eta) + \frac{c-1}{\eta} - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbb{E}\left[\frac{1}{\rho^2}\right]\mathbb{E}[\sigma_i] + \mathbb{E}[\rho^2]\mathbb{E}\left[\frac{1}{\sigma_i}\right] + d\right),$$
$$\nabla^2 h(\hat{\eta}) = -n\frac{d^2}{d\eta^2}\log K_{3/2}(\eta) - \frac{c-1}{\eta^2}.$$

In the Taylor expansion of Equation (4.12), the first-order term $\nabla h(\hat{\eta})(\eta - \hat{\eta})$ is equal to 0 because $\hat{\eta}$ is the maximum of $h(\eta)$.

Then we approximate Equation (4.11) as

$$q(\eta) \propto \exp\{h(\eta)\} \approx \exp\left\{h(\hat{\eta}) + \frac{1}{2}\nabla^2 h(\hat{\eta})(\eta - \hat{\eta})^2\right\}.$$

This results in the optimal variational density for $q^*(\eta)$,

$$q^*(\eta) \approx \mathrm{N}(\mu_\eta, \sigma_\eta^2), \tag{4.13}$$

where $\mu_\eta = \hat{\eta}$ and $\sigma_\eta^2 = -\left(\nabla^2 h(\hat{\eta})\right)^{-1}$.

Within the algorithm, we iterate between holding the other coordinate updates fixed, whilst updating $q^*(\eta)$ from Equation (4.13) and holding $q^*(\eta)$ fixed, whilst updating the other coordinate updates. Each time we update $q^*(\eta)$, we require the use of numerical optimisation to obtain $\hat{\eta}$, the optimal value of $h(\eta)$, such as Brent's method (Brent (1973)). Whilst updating the other coordinate updates, the only expectation required is $\mathbb{E}[\eta]$ and this can easily be computed as $\mathbb{E}[\eta] = \mu_\eta$ by using a property of the normal distribution.

4.4 Simulations

Throughout the subsections, we conducted a wide variety of simulation studies to assess the performance of the proposed VB algorithms within the comparative studies with the MCMC method from Chapter 3. Firstly, we presented the parameter estimation under

Table 4.1: Parameter estimation and calculation speed comparison of each method in Simulation 1.

	Methods	RMSE	MMAD	β_0	β_1	β_2	β_3	Time	Iterations
$\tau = 0.25$	MCMC500	0.1004	0.1730	0.0012	0.6238	0.0092	0.0006	89.60	15000
	VBL500	0.0931	0.1678	-0.0028	0.6007	0.0150	0.0013	16.15	301.84
	VBAL500	0.0961	0.1698	-0.0031	0.6108	0.0138	0.0011	15.14	284.66
	MCMC1000	0.0934	0.1639	0.0062	0.6055	0.0018	0.0019	202.39	15000
	VBL1000	0.0824	0.1503	0.0047	0.5681	0.0030	0.0018	42.68	382.32
	VBAL1000	0.0840	0.1522	0.0047	0.5734	0.0026	0.0017	41.44	370.52
$\tau = 0.5$	MCMC500	0.0027	0.0333	-0.0010	-0.0085	0.0020	0.0001	90.38	15000
	VBL500	0.0030	0.0366	-0.0030	-0.0100	0.0049	0.0004	15.62	291.82
	VBAL500	0.0030	0.0361	-0.0033	-0.0071	0.0043	0.0004	15.35	285.28
	MCMC1000	0.0009	0.0182	-0.0008	0.0019	0.0024	-0.0001	202.03	15000
	VBL1000	0.0010	0.0176	-0.0009	0.0033	0.0024	0.0002	38.95	349.40
	VBAL1000	0.0010	0.0175	-0.0011	0.0048	0.0021	0.0002	38.29	343.50
$\tau = 0.75$	MCMC500	0.0939	0.1680	-0.0030	-0.6032	-0.0173	-0.0012	89.86	15000
	VBL500	0.0877	0.1611	-0.0052	-0.5816	-0.0167	-0.0010	16.36	305.32
	VBAL500	0.0871	0.1602	-0.0052	-0.5794	-0.0168	-0.0009	16.55	306.40
	MCMC1000	0.0997	0.1675	-0.0069	-0.6277	0.0043	0.0015	202.01	15000
	VBL1000	0.0891	0.1584	-0.0062	-0.5922	0.0019	0.0010	40.23	362.32
	VBAL1000	0.0886	0.1580	-0.0062	-0.5906	0.0019	0.0009	40.11	361.34

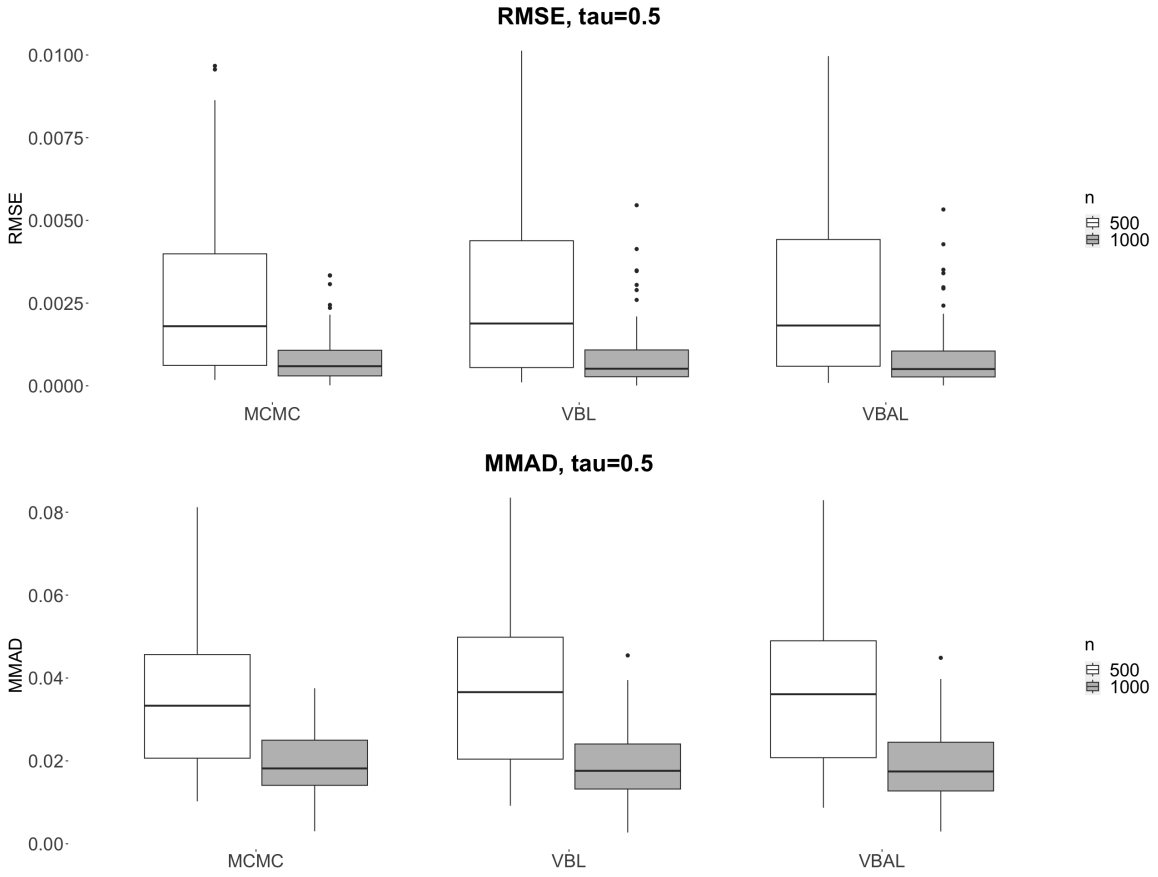
a simple setting and its computational performance. Secondly, the computational performance was further investigated under a high-dimensional settings with an increasing number of predictors for the fixed sample size. Finally, we assessed the accuracy of the proposed algorithms and their computational performance under high-dimensional setting with different error distributional assumptions. All computations were carried out on the R version 4.2.2 on an Intel (Core i7-4790) CPU@3.6GHz machine with 16GB DDR3 RAM memory.

4.4.1 Parameter Estimation

In this simulation study, the parameter estimation and computational performance of the proposed variational Bayes algorithms of Bayesian Huberised Lasso and adaptive Lasso quantile regression models were compared with its approximate Gibbs sampling algorithm. For simplicity, we denote the proposed algorithms and the approximate Gibbs sampling algorithm as VBL, VBAL and MCMC, respectively where VBL and VBAL were employed with Bayesian Huberised Lasso and Bayesian Huberised adaptive Lasso, respectively. We generated data from the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + x_{i1}\epsilon_i, \quad i = 1, \ldots, n,$$

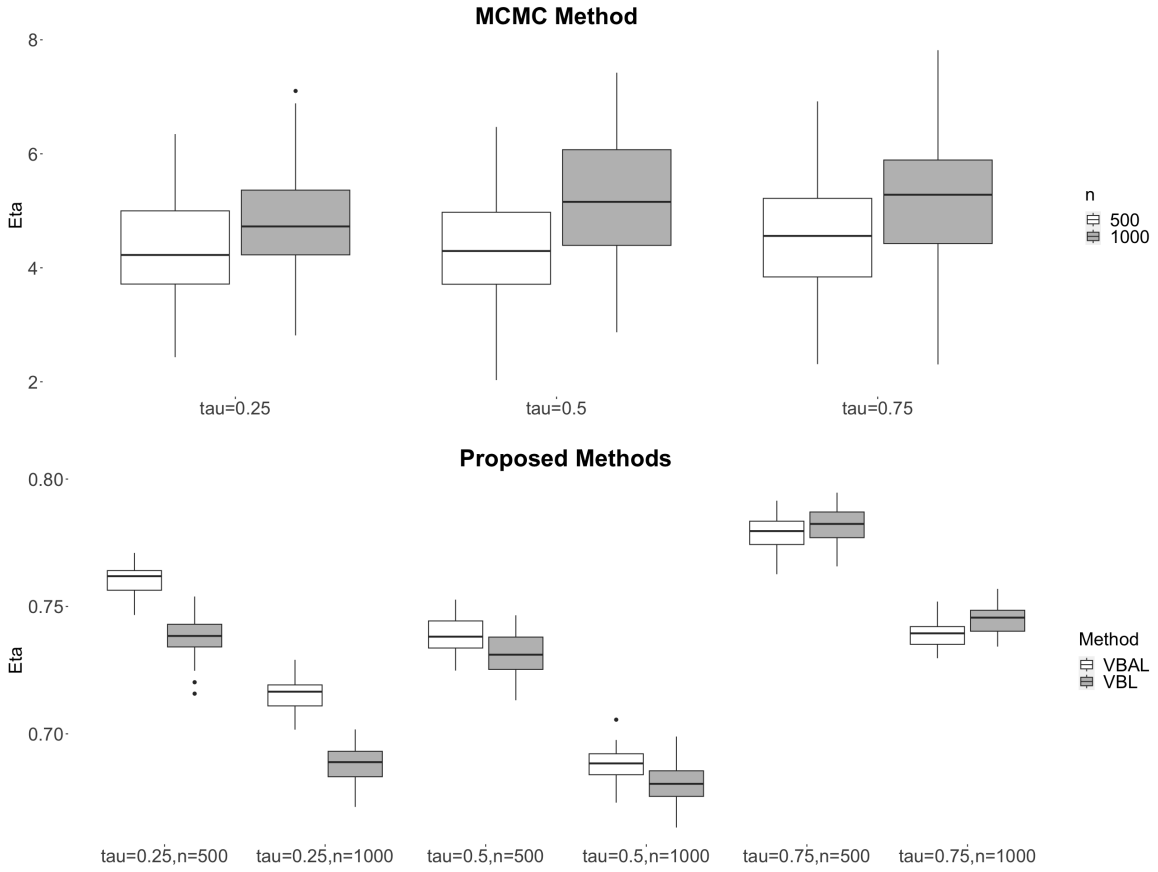
Figure 4.1: Boxplots of RMSE & MMAD based on 50 replications in parameter estimation simulation for the proposed VB algorithms and MCMC method ($\tau = 0.5$).



where y_i is the response variable. x_{i1} and x_{i2} were produced from a uniform distribution $\mathrm{U}(0, 1)$, x_{i3} was generated from the standard normal distribution $\mathrm{N}(0, 1)$, $\beta_0 = -0.5$, $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 0$ and ϵ_i was drawn from the standard normal distribution $\mathrm{N}(0, 1)$. That is, a heteroscedastic model was considered. A total of 50 replications were generated, different sample sizes $n \in \{500, 1000\}$ were chosen, and the model was fitted at three different quantile levels: $\tau \in \{0.25, 0.5, 0.75\}$. All the hyper-parameters of both proposed algorithms were set to 1.

To sample from the posterior distributions as much as possible, the MCMC method was iterated for 15,000 posterior samples, whilst discarding the first 5,000 samples and retaining the results in the last 10,000 iterations. In both VBL and VBAL algorithms, 10^{-5} was taken as the convergence criterion, and the algorithm terminated after the ELBO reached the convergence criterion. To compare the overall estimation effect of unknown parameters, we computed posterior median and optimal estimate of each element of β_j's from the MCMC method and VBL & VBAL algorithms, respectively, for point estimates of

Figure 4.2: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for parameter estimation simulation.



β_j's, and the performances were evaluated via RMSE defined as $\left[4^{-1}\sum_{j=0}^{3}(\hat{\beta}_j - \beta_j^{\text{true}})^2\right]^{1/2}$, MMAD defined as median $\left[4^{-1}\sum_{j=0}^{3}\left|\hat{\beta}_j - \beta_j^{\text{true}}\right|\right]$, and bias defined as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\text{true}})$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$. To compare the computational efficiency, we calculated the computational time (Time in seconds) and the number of iterations (Iteration) that were averaged over 50 replications.

We reported parameter estimation results in Table 4.1. From the table, the results show that the performance of the proposed algorithms was close to that of the MCMC method in terms of bias, RMSE and MMAD. All the values of bias, RMSE and MMAD were less than 0.1 except the bias of β_1 and those of RMSE and MMAD for lower and upper quantiles ($\tau = 0.25, 0.75$) because of quantile levels differing from median. This indicated that the parameter estimation effect was generally good. As the quantiles were being farther away from the median, the accuracy of parameter estimation decreased gradually at the fixed sample size. However, an increase in the sample size resulted in decreases in RMSE, MMAD and bias. This is generally the case for the median as well, which indicated that

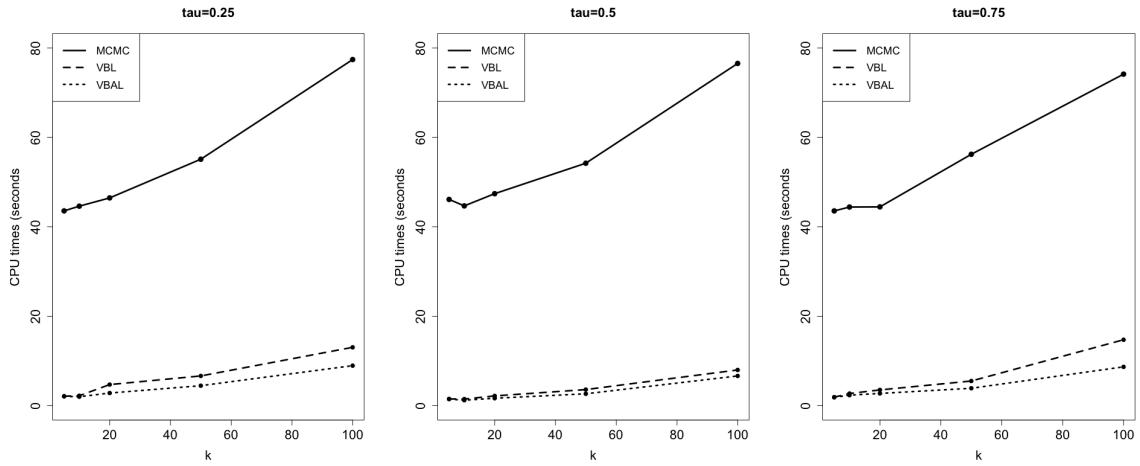
an increase in the sample size would improve the accuracy of parameter estimation. In comparison of sample size and computational time, the computational time of all three methods increased when the sample size increased. Yet, the proposed algorithms were approximately 5-6 times faster than the MCMC method. Upon looking at the number of iterations required for convergence in the proposed algorithms, they increased slightly when the quantile levels deviated from the median, They also increased with an increase in the sample size. Specifically, the VBAL algorithm required less iterations to converge than the VBL algorithm. So, the VBAL algorithm was fastest of all three methods.

We reported boxplot performances for visualisation in Figure 4.1 at $\tau = 0.5$. Upon looking at RMSE performances, each method produced the similar boxplot for the sample size of either $n = 500$ or $n = 1000$. For $n = 500$, each boxplot had a larger box with longer whiskers. As the sample size increased to $n = 1000$, each boxplot became tighter with shorter whiskers and more outliers, particularly, those of the proposed algorithms produced more outliers. Nevertheless, an increase in the sample size saw a decrease in the median value of RMSE. Upon looking at MMAD performances, the boxplots exhibited the similar behaviour as those for RMSE performances. Similar performances are observed for $\tau = 0.25, 0.75$ and the figures are provided in Appendix D.3 (see Figures D.13-D.14).

We also presented the boxplots of η that were obtained as posterior median from the MCMC method and as optimal estimate from the proposed algorithms in Figure 4.2 for all the quantile levels. In terms of method comparison, the proposed algorithms produced different estimates of η from those of the MCMC method, which may be due to variational approximation. Upon looking at the MCMC method, it is consistent in producing the same posterior estimate at varying quantile levels and varying sample sizes. On the other hand, the proposed algorithms saw subtle differences in boxplots at varying quantile levels, such as an increase in the sample size led to a small decrease in the optimal estimates. Yet, the VBL algorithm produced similar estimates as those of the VBAL algorithm. It is also evident that the proposed algorithms have narrower boxplots than those of the MCMC method. This suggested that the proposed algorithms may produce more precise estimates of η than the MCMC method.

Overall, the results observed the similar performances of parameter estimation for all three methods at varying quantile levels. Yet, the computational time of the proposed algorithms was significantly faster than that of the MCMC method.

Figure 4.3: Computational performance for the proposed VB algorithms and MCMC method for different quantile levels ($\tau = 0.25, 0.5, 0.75$).



4.4.2 Computational Details and CPU Times

We calculated the computational times for the MCMC method and the proposed VB algorithms. Let $\beta_0 = 3$, $\boldsymbol{\beta} = (0.5, 1, 1.5, 1, 0, \ldots, 0)^T$ and the data was generated from the following heteroscedastic model, $y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta} + x_{ik}\epsilon_i$, $i = 1, \ldots, n$, which is the same as that in Section 4.4.1, and $\boldsymbol{x}_i$ was generated from a multivariate normal distribution $\mathrm{N}_k(0, \Sigma)$ with $\Sigma = (r^{|i-j|})_{1 \leq i,j \leq k}$ for $|r| < 1$ and $r = 0.5$ was selected. Figure 4.3 shows the result of CPU times in seconds for $n = 200$ and varying dimensions $k \in \{5, 10, 20, 50, 100\}$, averaged over 10 replications. Like in Section 4.4.1, we retained 10,000 posterior samples after discarding the first 5,000 samples as a burn-in for the MCMC method, and we obtained the optimal estimates after the ELBO meets the convergence criterion for the VBL and VBAL algorithms.

From the figure, the proposed algorithms performed significantly faster than the MCMC method at varying quantile levels. As the dimension became higher, the computational time of all three methods increased, however, that of the proposed algorithms increased slower than that of the MCMC methods. By comparing the proposed algorithms, both VBL and VBAL algorithms have similar computational time at the median, however, the VBAL algorithm had the lowest time across all the quantiles for varying dimensions.

Remark 4.1. *During the implementation of the proposed algorithms, the approximate density of ρ^2 may have some limitations, as the number of predictors and the sample size increase. This is because the modified Bessel function of the second kind involved in optimising the approximate density can grow exponentially as a function of predictor and sample size, and becomes infinite in a software implementation.*

4.4.3 Simulation Studies

In simulation studies, we illustrated performance of the proposed algorithms. We compared both estimation accuracy and computational performance of the proposed algorithms with the MCMC methods. To this end, we considered the following heteroscedastic model:

$$y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta} + x_{ik}\sigma\epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})$, $(x_{i1}, \ldots, x_{ik-1})$ was generated from a multivariate normal distribution $\mathrm{N}_{k-1}(0, \Sigma)$ with $\Sigma = (r^{|i-j|})_{1\leq i,j\leq k}$ for $|r| < 1$, x_{ik} was generated from a uniform distribution $\mathrm{U}(0, 1)$, the intercept term $\beta_0 = -1$ and all elements in the parameter $\boldsymbol{\beta}$ were 0 except $\beta_1 = 2, \beta_2 = 1, \beta_3 = -1, \beta_4 = -2, \beta_5 = \beta_{10} = \beta_{20} = \beta_{30} = \beta_{40} = \beta_{50} = 1$. We considered the three scenarios.

- Simulation 1: Gaussian noise. $\boldsymbol{\epsilon} \sim \mathrm{N}_n(0, I_n)$ and $\sigma = 1$.
- Simulation 2: large outliers. $\epsilon = W/\sqrt{\mathrm{Var}(W)}$ and $\sigma = 9.67$. W is a random variable according to the contaminated density defined by $0.9 \times \mathrm{N}(0, 1) + 0.1 \times \mathrm{N}(0, 15^2)$, where $\sqrt{\mathrm{Var}(W)} = 4.83$.
- Simulation 3: Multiple outliers and skew Student-t noise. $\epsilon_i \sim 0.8 \times \text{Skew-t}_3(\gamma = 3) + 0.1 \times \mathrm{N}(0, 10^2) + 0.1 \times \text{Cauchy}(0, 1)$ and $\sigma = 1$.

We considered six combinations of two sample sizes $n \in \{300, 600\}$ and three correlations $r \in \{0.2, 0.5, 0.8\}$ for three different quantile levels, $\tau \in \{0.25, 0.5, 0.75\}$. The computational time (Time in seconds) and the number of iterations (Iteration) were also calculated. All the hyper-parameters were set to 1 for each method. The dataset was replicated 50 times. Both MCMC method and proposed algorithms followed the same procedure as those in Section 4.4.1. To evaluate the numerical performances, RMSE and MMAD were computed.

We reported simulation results in Tables 4.2-4.4 and Figures 4.4-4.7. From the tables, generally, the results show that the proposed algorithms converged well in all cases of sample sizes, correlation coefficients and quantile levels. At the fixed value of sample size and correlation, the algorithms have fewer iterations and faster convergence at the median than those at the lower and upper quantile levels. This is because when $\tau \neq 0.5$, the proposed algorithms require more computations involving τ terms resulting in additional computational time. The proposed algorithms have lower RMSE and MMAD than those of the MCMC method in most cases, yet they were very similar. At the fixed value of sample size and

Figure 4.4: Boxplots of RMSE based on 50 replications in Simulation 1 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

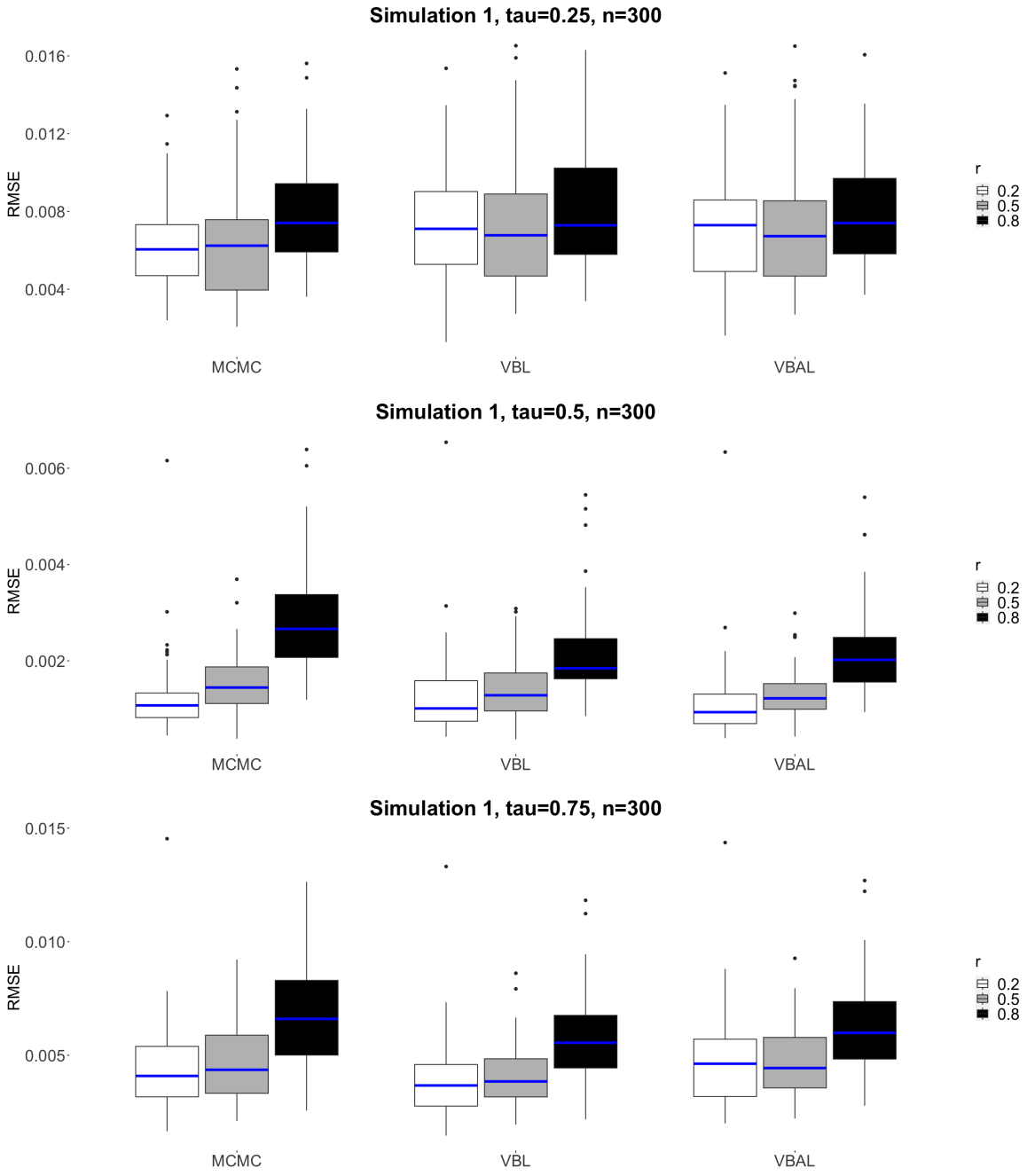


Figure 4.5: Boxplots of RMSE based on 50 replications in Simulation 2 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

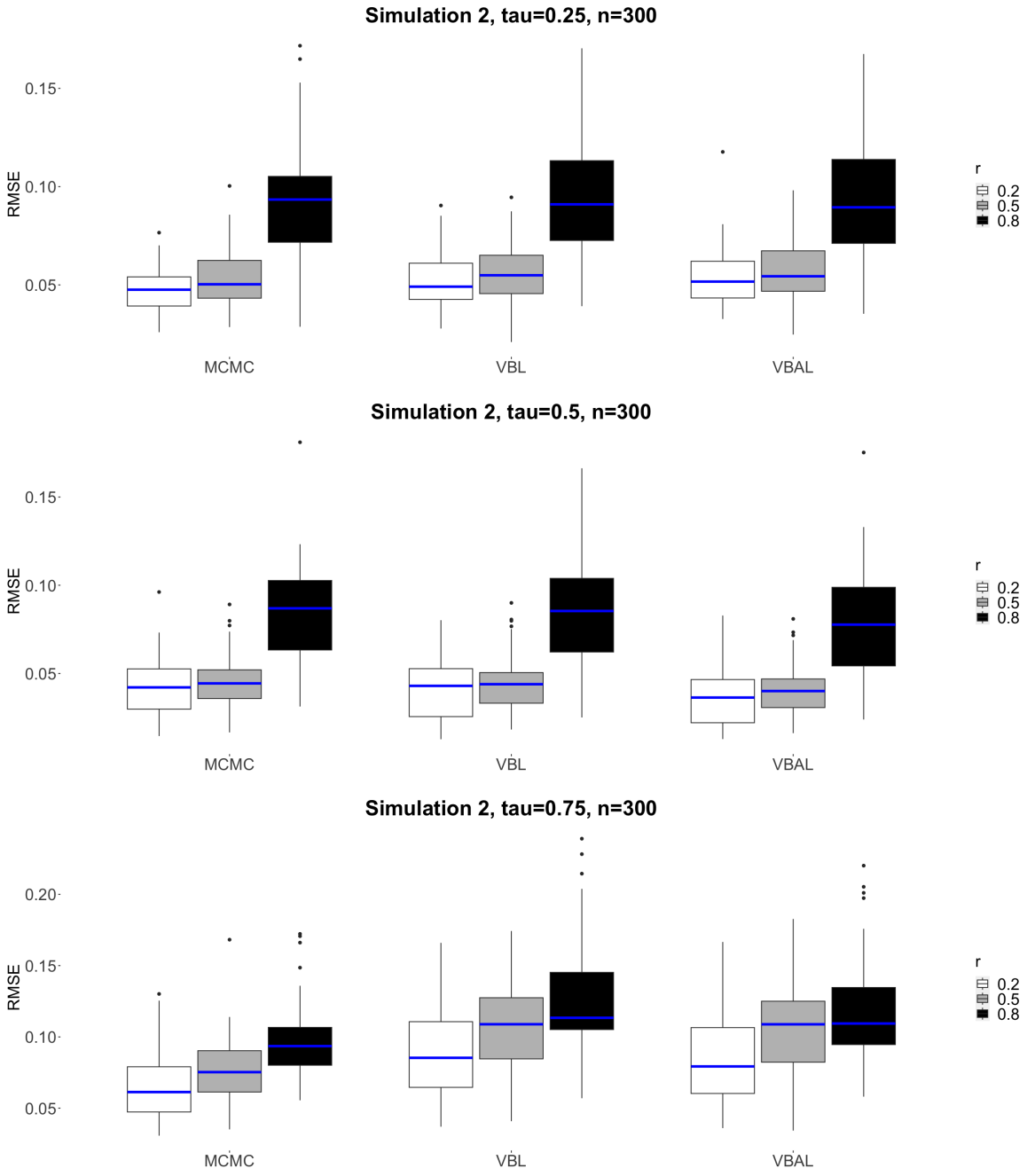


Figure 4.6: Boxplots of RMSE based on 50 replications in Simulation 3 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

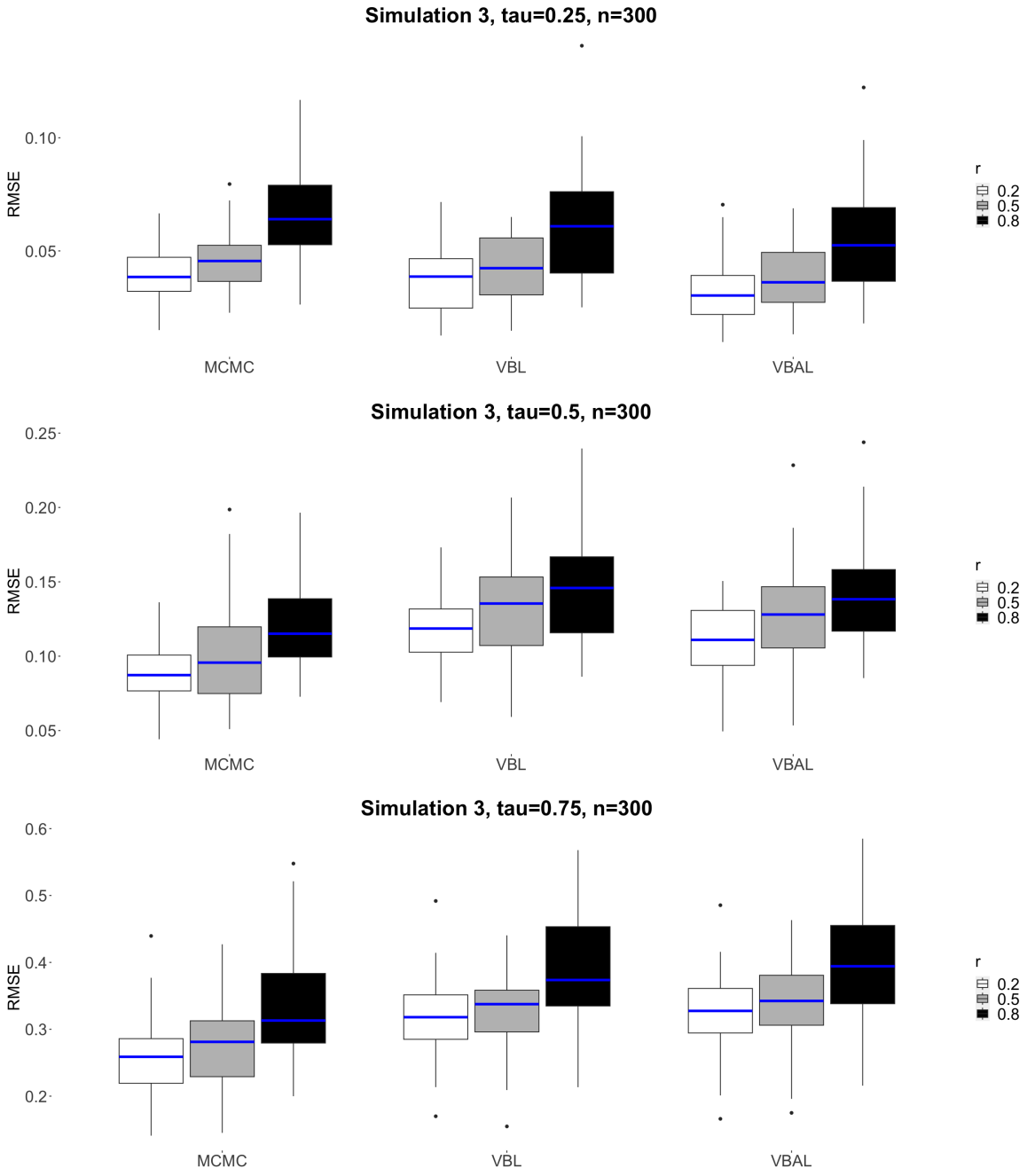


Table 4.2: Numerical results based on 50 replications in Simulation 1 at different quantile levels ($\tau = 0.25, 0.5, 0.75$), different correlation coefficients ($r = 0.2, 0.5, 0.8$) and sample sizes ($n = 300, 600$) for the proposed VB algorithm and the MCMC method.

	r	Methods	RMSE	MMAD	Time	Iterations
$\tau = 0.25$	0.2	MCMC300	0.0062	0.0338	72.15	15000
		VBL300	0.0072	0.0324	12.41	185.48
		VBAL300	0.0072	0.0335	10.42	154.22
		MCMC600	0.0061	0.0249	138.37	15000
		VBL600	0.0062	0.0240	31.29	233.06
		VBAL600	0.0062	0.0243	24.17	181.24
	0.5	MCMC300	0.0064	0.0401	72.51	15000
		VBL300	0.0074	0.0364	12.75	187.76
		VBAL300	0.0072	0.0379	10.51	152.94
		MCMC600	0.0063	0.0287	141.61	15000
		VBL600	0.0064	0.0268	33.12	239.58
		VBAL600	0.0064	0.0276	25.44	184.56
	0.8	MCMC300	0.0081	0.0511	72.45	15000
		VBL300	0.0082	0.0460	12.77	185.12
		VBAL300	0.0081	0.0475	11.14	162.40
		MCMC600	0.0073	0.0372	138.94	15000
		VBL600	0.0070	0.0335	30.97	230.52
		VBAL600	0.0071	0.0342	24.20	181.12
$\tau = 0.5$	0.2	MCMC300	0.0013	0.0237	72.87	15000
		VBL300	0.0013	0.0219	10.37	153.26
		VBAL300	0.0012	0.0217	8.49	125.64
		MCMC600	0.0005	0.0154	138.17	15000
		VBL600	0.0005	0.0147	26.83	201.04
		VBAL600	0.0005	0.0147	20.60	154.20
	0.5	MCMC300	0.0016	0.0297	72.52	15000
		VBL300	0.0014	0.0262	10.89	159.44
		VBAL300	0.0013	0.0259	8.68	127.62
		MCMC600	0.0006	0.0181	137.78	15000
		VBL600	0.0006	0.0169	28.41	203.06
		VBAL600	0.0006	0.0169	21.55	154.96
	0.8	MCMC300	0.0029	0.0398	72.54	15000
		VBL300	0.0022	0.0333	10.67	154.82
		VBAL300	0.0022	0.0343	8.62	125.64
		MCMC600	0.0014	0.0291	132.83	15000
		VBL600	0.0012	0.0266	25.46	193.94
		VBAL600	0.0012	0.0266	19.57	150.10
$\tau = 0.75$	0.2	MCMC300	0.0045	0.0341	72.32	15000
		VBL300	0.0041	0.0313	14.20	209.58
		VBAL300	0.0047	0.0332	11.77	171.98
		MCMC600	0.0054	0.0251	134.31	15000
		VBL600	0.0050	0.0242	32.51	247.12
		VBAL600	0.0053	0.0245	25.43	195.98
	0.5	MCMC300	0.0047	0.0365	72.38	15000
		VBL300	0.0041	0.0343	14.12	207.06
		VBAL300	0.0047	0.0353	11.68	170.26
		MCMC600	0.0057	0.0293	132.30	15000
		VBL600	0.0052	0.0279	32.49	249.02
		VBAL600	0.0056	0.0285	25.89	197.42
	0.8	MCMC300	0.0068	0.0509	72.32	15000
		VBL300	0.0057	0.0454	13.79	199.12
		VBAL300	0.0063	0.0464	11.40	166.48
		MCMC600	0.0059	0.0376	139.28	15000
		VBL600	0.0053	0.0342	33.33	248.56
		VBAL600	0.0057	0.0351	26.57	198.34

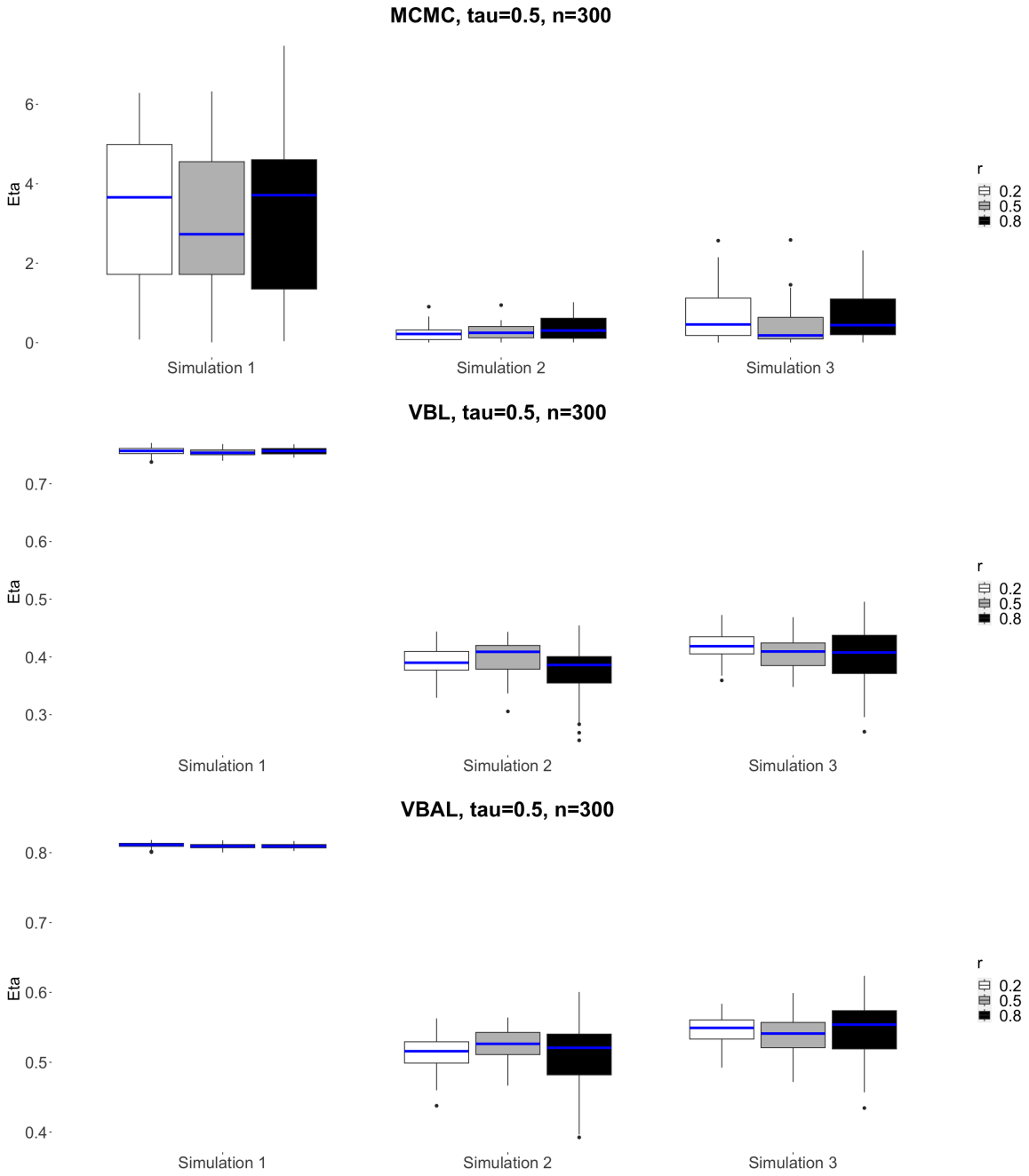
Table 4.3: Numerical results based on 50 replications in Simulation 2 at different quantile levels ($\tau = 0.25, 0.5, 0.75$), different correlation coefficients ($r = 0.2, 0.5, 0.8$) and sample sizes ($n = 300, 600$) for the proposed VB algorithm and the MCMC method.

	r	Methods	RMSE	MMAD	Time	Iterations
$\tau = 0.25$	0.2	MCMC300	0.0480	0.1440	72.53	15000
		VBL300	0.0527	0.1393	21.88	325.14
		VBAL300	0.0543	0.1412	17.17	251.62
		MCMC600	0.0373	0.1160	139.41	15000
		VBL600	0.0417	0.1201	85.30	629.26
		VBAL600	0.0433	0.1166	70.87	523.72
	0.5	MCMC300	0.0535	0.1495	73.38	15000
		VBL300	0.0564	0.1481	22.20	321.28
		VBAL300	0.0576	0.1521	17.77	258.00
		MCMC600	0.0399	0.1251	135.76	5000
		VBL600	0.0438	0.1284	74.98	556.92
		VBAL600	0.0456	0.1280	66.56	494.28
	0.8	MCMC300	0.0933	0.2084	71.79	15000
		VBL300	0.0931	0.1865	24.17	354.70
		VBAL300	0.0938	0.1910	17.29	254.38
		MCMC600	0.0694	0.1778	137.81	15000
		VBL600	0.0700	0.1662	80.76	598.78
		VBAL600	0.0714	0.1681	64.69	480.10
$\tau = 0.5$	0.2	MCMC300	0.0421	0.1330	70.85	15000
		VBL300	0.0408	0.1286	20.50	313.08
		VBAL300	0.0367	0.1279	16.41	245.78
		MCMC600	0.0244	0.1062	138.17	15000
		VBL600	0.0256	0.1046	82.89	611.04
		VBAL600	0.0227	0.1020	65.66	488.04
	0.5	MCMC300	0.0461	0.1496	72.49	15000
		VBL300	0.0442	0.1392	19.39	282.82
		VBAL300	0.0408	0.1392	16.63	242.96
		MCMC600	0.0282	0.1184	137.76	5000
		VBL600	0.0303	0.1181	68.27	508.12
		VBAL600	0.0281	0.1157	63.47	474.44
	0.8	MCMC300	0.0843	0.2031	71.94	5000
		VBL300	0.0848	0.1859	21.84	321.76
		VBAL300	0.0800	0.1928	16.65	245.64
		MCMC600	0.0465	0.1574	137.93	15000
		VBL600	0.0479	0.1495	68.13	506.86
		VBAL600	0.0447	0.1495	64.04	476.66
$\tau = 0.75$	0.2	MCMC300	0.0647	0.1416	72.47	15000
		VBL300	0.0916	0.1470	21.61	317.84
		VBAL300	0.0867	0.1478	16.66	245.18
		MCMC600	0.0464	0.1148	140.00	15000
		VBL600	0.0608	0.1224	74.53	550.96
		VBAL600	0.0572	0.1219	66.28	490.46
	0.5	MCMC300	0.0768	0.1666	72.90	15000
		VBL300	0.1073	0.1689	23.32	339.90
		VBAL300	0.1042	0.1719	17.18	250.80
		MCMC600	0.0555	0.1339	138.37	15000
		VBL600	0.0670	0.1342	75.13	559.92
		VBAL600	0.0652	0.1330	60.79	453.94
	0.8	MCMC300	0.0971	0.1984	72.16	15000
		VBL300	0.1256	0.2004	22.70	333.64
		VBAL300	0.1184	0.2004	16.83	248.30
		MCMC600	0.0718	0.1669	137.85	15000
		VBL600	0.0834	0.1681	72.53	535.92
		VBAL600	0.0800	0.1639	60.76	449.10

Table 4.4: Numerical results based on 50 replications in Simulation 3 at different quantile levels ($\tau = 0.25, 0.5, 0.75$), different correlation coefficients ($r = 0.2, 0.5, 0.8$) and sample sizes ($n = 300, 600$) for the proposed VB algorithm and the MCMC method.

	r	Methods	RMSE	MMAD	Time	Iterations
$\tau = 0.25$	0.2	MCMC300	0.0386	0.1232	70.36	15000
		VBL300	0.0372	0.1141	18.80	282.68
		VBAL300	0.0320	0.1112	15.10	228.14
		MCMC600	0.0203	0.0879	137.33	15000
		VBL600	0.0188	0.0849	70.35	522.86
		VBAL600	0.0165	0.0816	54.36	404.66
	0.5	MCMC300	0.0448	0.1338	70.85	15000
		VBL300	0.0425	0.1229	18.80	281.02
		VBAL300	0.0374	0.1261	15.83	234.88
		MCMC600	0.0259	0.0988	130.31	15000
		VBL600	0.0253	0.0904	71.29	548.46
		VBAL600	0.0220	0.0898	50.61	388.02
	0.8	MCMC300	0.0652	0.1692	71.40	15000
		VBL300	0.0609	0.1539	18.08	272.14
		VBAL300	0.0545	0.1545	13.71	205.02
		MCMC600	0.0373	0.1344	138.36	15000
		VBL600	0.0339	0.1186	63.69	468.38
		VBAL600	0.0309	0.1180	53.81	395.50
$\tau = 0.5$	0.2	MCMC300	0.0887	0.1521	72.66	15000
		VBL300	0.1174	0.1641	20.76	305.48
		VBAL300	0.1117	0.1641	15.55	229.02
		MCMC600	0.0700	0.1257	138.51	15000
		VBL600	0.0853	0.1329	72.23	535.76
		VBAL600	0.0817	0.1309	55.80	415.56
	0.5	MCMC300	0.1012	0.1657	70.30	15000
		VBL300	0.1318	0.1786	20.01	301.70
		VBAL300	0.1264	0.1786	16.08	242.36
		MCMC600	0.0766	0.1400	135.85	15000
		VBL600	0.0918	0.1433	60.59	449.96
		VBAL600	0.0885	0.1420	55.48	412.70
	0.8	MCMC300	0.1191	0.2062	71.46	15000
		VBL300	0.1449	0.2101	18.49	276.16
		VBAL300	0.1401	0.2106	16.65	249.38
		MCMC600	0.0949	0.1827	138.209	15000
		VBL600	0.1100	0.1776	65.84	487.86
		VBAL600	0.1039	0.1762	62.92	467.52
$\tau = 0.75$	0.2	MCMC300	0.2602	0.2157	72.96	15000
		VBL300	0.3162	0.2243	26.01	381.70
		VBAL300	0.3139	0.2215	25.84	377.84
		MCMC600	0.2397	0.1931	137.93	15000
		VBL600	0.2701	0.2055	88.10	656.32
		VBAL600	0.2724	0.2070	79.09	590.12
	0.5	MCMC300	0.2753	0.2337	71.23	15000
		VBL300	0.3277	0.2371	23.64	354.34
		VBAL300	0.3284	0.2342	21.40	318.50
		MCMC600	0.2508	0.2158	136.35	15000
		VBL600	0.2730	0.2182	88.71	663.50
		VBAL600	0.2735	0.2182	69.26	517.66
	0.8	MCMC300	0.3338	0.3024	70.63	15000
		VBL300	0.3858	0.2811	23.88	357.60
		VBAL300	0.3958	0.2945	20.09	301.40
		MCMC600	0.3005	0.2743	134.40	15000
		VBL600	0.3208	0.2666	87.75	669.14
		VBAL600	0.3272	0.2713	77.56	592.78

Figure 4.7: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 300$ ($\tau = 0.5$).



quantile level, when the correlation coefficient increased, the accuracy of parameter estimation decreased slightly. This suggested that the correlation between predictors could have an influence on variable selection. At the fixed value of correlation coefficient and quantile level, an increase in the sample size led to a decrease in RMSE and MMAD. This suggested that increasing the sample size is beneficial to parameter estimation and variable selection because it increases the amount of data information, reduces the uncertainty of parameters and thus, improves the effect of parameter estimation and variable selection.

Upon looking at Table 4.2, in Simulation 1, there was no outliers in generated datasets. All three methods were comparable in terms of RMSE and MMAD. However, the proposed algorithm was much faster than the MCMC method. In Simulation 2 with large outliers (Table 4.3), the proposed algorithms required more iterations to converge, whilst the MCMC method was consistent in computational time. Yet, the proposed algorithms remained faster. Similar results are also observed in Simulation 3 (Table 4.4) where there was skewed and heavy-tailed noise with multiple outliers. This indicated that the proposed algorithms are as robust to outliers as the MCMC method is. By comparing the proposed algorithms, the VBAL algorithm was generally faster than the VBL algorithm because the former required fewer iterations than the latter.

We reported boxplot performances of RMSE in Figures 4.4-4.6 for the sample size of $n = 300$. Evidently, there is no noticeable difference between boxplots of both MCMC method and proposed algorithms due to the similar trend at varying quantile levels for different correlation coefficients. likewise, Simulations 1-3 each saw the similar trend. Upon looking at correlation coefficients, the trend saw an increase in the box size and the amount of outliers produced within the boxplot, as the correlation coefficient increased from $r = 0.2$ to $r = 0.8$. This coincided with Tables 4.2-4.4 where they have observed that the accuracy of parameter estimation would see a decrease after an increase in the correlation coefficient. Similar performances are observed for boxplots of RMSE at $n = 600$ and MMAD at $n = 300, 600$, and the figures are provided in Appendix D.3 (see Figures D.15-D.23).

We have also presented boxplots of η that were produced from the MCMC method as posterior median and from the proposed algorithms as optimal estimate in Figure 4.7 at $\tau = 0.5$ for $n = 300$. In terms of method comparison, the boxplots of the MCMC methods have larger box size with longer whiskers unlike the proposed algorithms having the tighter boxplots with shorter whiskers. This coincided with Figure 4.2 in Section 4.4.1 where the proposed algorithms have proven to be more precise in producing estimates of η. In terms of correlation coefficients, they behaved similarly as those in Figures 4.4-4.6. Nevertheless,

Table 4.5: MSPE, MAPE, MHPE with $\delta = 1.345$ and MedSPE for the Boston Housing data, computed from 20-fold cross-validation.

	Methods	MSPE	MAPE	MedSPE	MHPE	Time	Iterations
	MCMC	0.3031	0.3396	0.0949	0.1294	106.65	15000
$\tau = 0.3$	VBL	0.3131	0.3420	0.0925	0.1329	14.25	201.15
	VBAL	0.3029	0.3360	0.0940	0.1285	6.37	90.10
	MCMC	0.2533	0.3113	0.0744	0.1080	104.90	15000
$\tau = 0.5$	VBL	0.2644	0.3175	0.0772	0.1128	11.64	167.85
	VBAL	0.2578	0.3112	0.0734	0.1090	6.74	98.45
	MCMC	0.2566	0.3348	0.0850	0.1124	109.98	15000
$\tau = 0.7$	VBL	0.2626	0.3370	0.0836	0.1145	10.17	140.60
	VBAL	0.2652	0.3301	0.0848	0.1152	6.70	93.60

when looking at Simulations 1-3, both MCMC method and proposed algorithms adaptively selected the value η where it is relatively large in the absence of outliers (Simulation 1) and relatively small in the presence of outliers (Simulations 2-3). Therefore, like the MCMC method, η was adaptively chosen for different scenarios via the proposed algorithms. Similar performances are observed for sample sizes of $n = 600$ at $\tau = 0.5$ and $n = 300, 600$ at $\tau = 0.25, 0.75$, and the figures are provided in Appendix D.3 (see Figures D.24-D.28).

To summarise the results, the proposed algorithms consistently performed well in terms of parameter estimation and variable selection like the simulation studies of the MCMC method in Chapter 3. They were also significantly faster than the MCMC method for different scenarios at varying quantile levels, particularly, the VBAL algorithm has the lowest computational time due to the nature of adaptive lasso. It is notable that an increase in the correlation coefficients would lead to a decrease in the accuracy of parameter estimation. Nevertheless, the proposed algorithms have proven to be robust to outliers under various noises including skewness and heavy tail like the MCMC method, yet they are faster.

4.5 Boston Housing Data Example

The computational performance and efficiency of the variational algorithm of Bayesian Huberised Lasso and adaptive Lasso quantile regression models were demonstrated via the analysis of a famous benchmarking dataset, Boston Housing data (Harrison Jr and Rubinfeld (1978)) in investigating the normality assumption of residuals for robust estimation methods. This dataset is available in R package 'spData' (Bivand et al. (2021)) and is found to have large outliers (Kawakami and Hashimoto (2023)). The response variable in

Figure 4.8: Posterior medians and 95% credible intervals of the coefficients for the MCMC method and optimal estimates of the coefficients for the proposed algorithms (VBL & VBAL), applied to the Boston Housing data at $\tau = 0.3, 0.5, 0.7$.

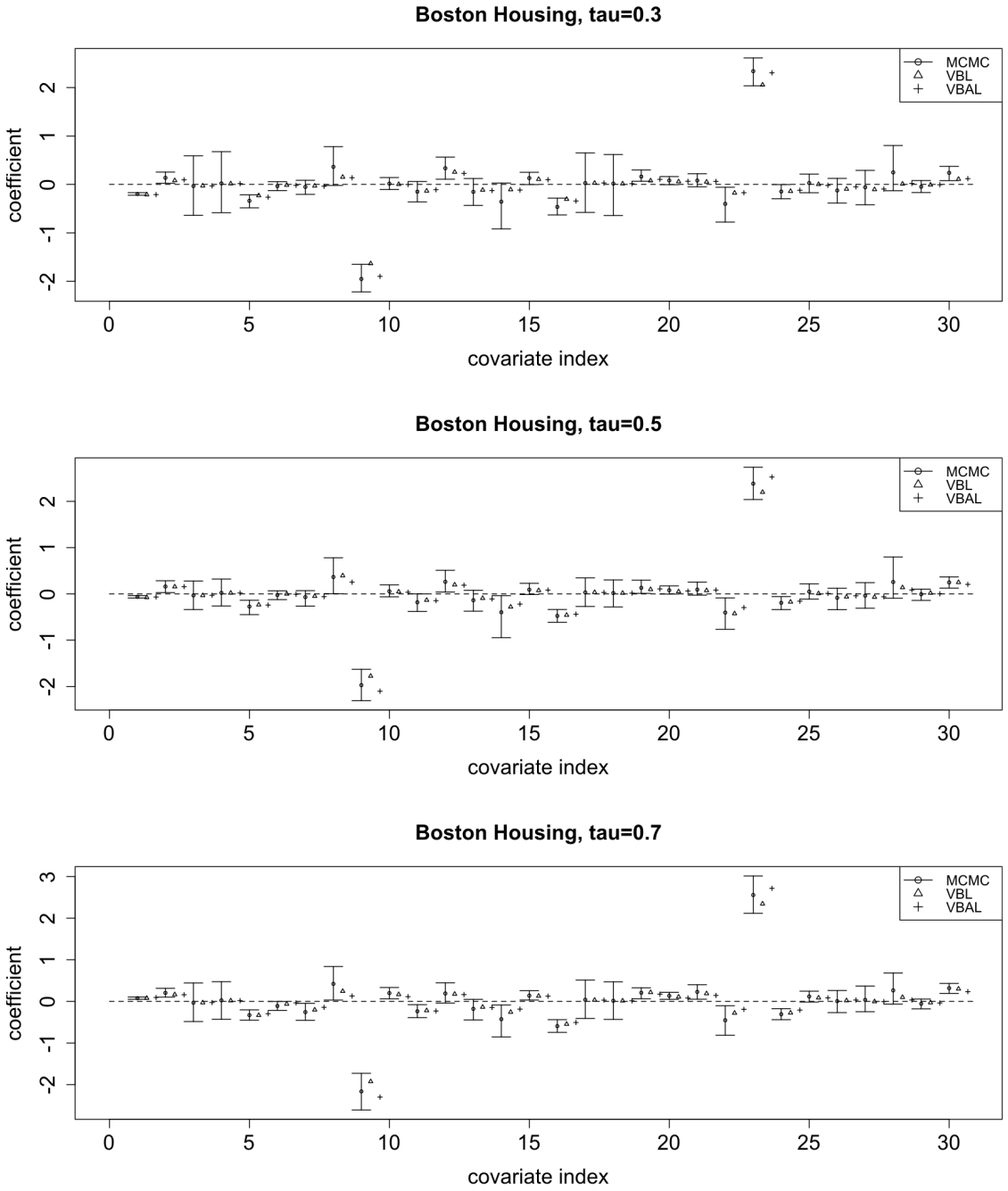


Table 4.6: Selection of important predictors by the MCMC method and the proposed algorithms for the Boston Housing data at $\tau = 0.3, 0.5, 0.7$ where the index 0 represents the intercept term.

	Methods	Predictor Index
$\tau = 0.3$	MCMC	0, 1, 4, 8, 11, 15, 18, 21, 22, 23, 29
	VBL	7, 11, 14, 22, 29
	VBAL	7, 11, 18, 22, 29
$\tau = 0.5$	MCMC	0, 1, 4, 7, 8, 10, 11, 13, 15, 18, 21, 22, 23, 29
	VBL	1, 7, 11, 22, 27, 29
	VBAL	1, 7, 11, 18, 22, 29
$\tau = 0.7$	MCMC	0, 1, 4, 5, 6, 7, 8, 9, 10, 13, 14, 15, 18, 19, 20, 21, 22, 23, 29
	VBL	1, 7, 9, 11, 14, 18, 19, 20, 22, 29
	VBAL	1, 7, 9, 11, 14, 18, 20, 22, 29

the Boston housing data is the corrected median value of owner-occupied homes in USD 1000's, and there are 15 predictors including one binary predictor. Similar to Hamura et al. (2022), Hashimoto and Sugasawa (2020) and Kawakami and Hashimoto (2023), we standardised 14 continuous predictors, and included squared values of these predictors, which resulted in 29 predictors in the design matrix for our models. We also centred response variables. The sample size is 506. Like in simulation studies, we also considered all the three methods of which we generated 15,000 posterior samples with a burn-in of 5,000 posterior samples for the MCMC method, and obtained the optimal estimates from the proposed algorithms after the convergence criterion was met with error threshold of 10^{-5}. Posterior medians were computed for the MCMC method. For all methods, we set the quantile levels as $\tau \in \{0.3, 0.5, 0.7\}$.

Since this dataset may contain outliers, we adopted the following four criteria as measures of predictive accuracy; MSPE, MAPE, MHPE for $\delta = 1.345$ and MedSPE via 20-fold cross validation. They are defined by $\mathrm{MSPE} = 20^{-1}\sum_{j=1}^{20}(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})^2$, $\mathrm{MAPE} = 20^{-1}\sum_{j=1}^{20}|\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)}|$, $\mathrm{MHPE} = 20^{-1}\sum_{j=1}^{20} L_\delta^{\mathrm{Huber}}(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})$ and $\mathrm{MedSPE} = \mathrm{median}_{1\le j\le 20}(\boldsymbol{y}_j - \boldsymbol{X}_j^T\hat{\boldsymbol{\beta}}^{(-j)})^2$, where $L_\delta^{\mathrm{Huber}}(\cdot)$ is given in Equation (1.1), $\hat{\boldsymbol{\beta}}^{(-j)}$ is the posterior median based on dataset except for j^{th} validation set, and $\boldsymbol{y}_j$ and $\boldsymbol{X}_j$ are the response variables and design matrix based on the j^{th} validation set, respectively.

The posterior medians and 95% credible intervals of the coefficients based on the MCMC method and the optimal estimates of the coefficients based on the VBL and VBAL algorithms are reported in Figure 4.8. The overall trend of the figure suggested that all the methods were comparable and their estimates were similar for all the quantile levels. For

the proposed algorithms, we followed a common practice of setting cutoff values for identifying the non-significant coefficients (Hoti and Sillanpää (2006), Guo et al. (2012), Feng et al. (2015), Feng et al. (2017), and Wang et al. (2023)). Specifically, if $|\beta_j| < 0.1$ for $j = 1, \ldots, k$, then the coefficient is suggested to be close to 0. The identification of important predictors for each method is summarised in Table 4.6. From the table, the MCMC method identified more important predictors than the proposed algorithms. Moreover, from the figure and table, we observed that more important predictors were selected, as the quantile level increased from $\tau = 0.3$ to $\tau = 0.7$. Table 4.5 also presents the predictive performance of the three methods for three different quantile levels. All the methods were comparable at each quantile in terms of MSPE, MAPE, MedSPE and MHPE. Looking at the computational performance, the VBAL algorithm was the fastest method and consistent for all the quantile levels unlike the VBL algorithm and the MCMC method. The VBL and VBAL algorithms were approximately 9 times and 15 times faster than the MCMC method, respectively.

4.6 Chapter Summary

We have proposed the novel VB Huberised Lasso and adaptive Lasso. By using the asymmetric Huberised loss function, they easily extended to VB Huberised Lasso quantile regression and VB Huberised adaptive Lasso quantile regression. We derived the approximate variational densities for the posterior distributions in corresponding hierarchical models using the mean-field VB and Laplace VI methods. The Laplace VI method retained the advantage of the data-dependent estimation of the tuning robustness parameter. The CAVI algorithm is utilised to solve the optimisation problem, in other words, to iteratively optimise each variational density. This resulted in the derivation of ELBO.

A variety of simulation studies and the Boston Housing data example showed that the proposed VB algorithms outperformed the MCMC method significantly in terms of parameter estimation, computational time and variable selection under various scenarios. In particular, the algorithm of VB Huberised adaptive Lasso quantile regression yielded the fastest computational speed amongst other methods because the adaptive Lasso penalty handles the over-fitting issue of the Lasso penalty efficiently and thus, it increases the convergence rate. Nevertheless, the proposed algorithms have proven to be successful in obtaining comparable statistical inference results, whilst saving significant amount of computational memory. Moreover, they retained the robustness of Bayesian Huberised regularisation, whilst enjoyed low computational cost and fast convergence rate, which indicated that the VB Huberised

Lasso and adaptive Lasso are much beneficial over the MCMC method.

Chapter 5

Conclusion and Future Research

In this chapter we discuss the main conclusions drawn from Chapters 2-4 and how these relate to the aim of our thesis, mentioned in Chapter 1. Several directions for further research are also outlined.

5.1 Conclusions

Throughout the thesis, we have presented the three distinctive contributions for the development of novel Bayesian methods in parametric statistical inference, whilst considering the robustness of methods. The first contribution is the new Bayesian non-linear quantile regression with the variable selection method for a specific medical application, which can easily be generalised to other applications. The second contribution is the novel loss function and Bayesian robust regularisation, which have seen a new variant of Bayesian regularised quantile regression with well-grounded theoretical properties in a high-dimensional setting. Yet, it faced some computational problems leading to the third contribution, that is the incorporation of the approximate-based technique to speed up the computational performance. The details of each contribution are as follows.

Chapter 2 has introduced the methodology of a new Bayesian non-linear quantile regression under the FP model and variable selection with quantile-dependent prior. The quantile regression analysis investigates how relationships differ across the median and upper quantile levels. The utility of FPs allows them to be non-linear parametrically. The variable selection investigates for important predictors that contribute to the non-linear relationships via the Bayesian paradigm. We have applied this methodology to the medical application in investigating the impact of BMI on the BP measures, including SBP and DBP using the

data extracted from the 2007-2008 NHANES database. The descriptive analysis showed that there are statistically significant associations between the BP measure and the risk factors of CVD. The model analysis suggested that the proposed method provides better estimates in terms of narrower credible intervals, autocorrelation plots with a faster decreasing rate of correlated posterior samples, and non-linearity. The variable selection has also identified important predictors, which contributed to the non-linear relationships under the FP model across all the quantile levels. Thus, the analysis has proven that the quantile-based FP approaches are adept at providing clearer statistical interpretations on the medical survey data, whilst also demonstrate the significance of considering non-linear relationships in the modelling process.

Chapter 3 has introduced a new loss function called the asymmetric Huberised loss function, and derived its probability density function that has the scale mixture of normal representation. In a high-dimensional setting, we proposed a new Bayesian robust regularisation, including Bayesian Huberised Lasso (Kawakami and Hashimoto (2023)) and Bayesian Huberised Elastic Net. Thus, the by-product of this research is the derivation of Bayesian Huberised regularised quantile regression. We have utilised the MCMC method with the approximate Gibbs sampler for the data-dependent estimation of the robustness parameter. The theorem and propositions are theoretically derived and new. Amongst all the simulation studies and real-life data examples, the proposed methods showed promising results in terms of robustness. In particular, they are effective in predictive accuracy being influenced by the robustness parameter under different error distributional assumptions.

An alternative approach is proposed in Chapter 4 for Bayesian Huberised regularised quantile regression for a fast computational performance. Both mean-field VI and Laplace VI methods, retaining the advantage of the data-dependent estimation of the robustness parameter, are utilised to propose both VB Huberised Lasso quantile regression and VB Huberised adaptive Lasso quantile regression. The CAVI algorithms and their ELBO are derived. They have proven to perform significantly better than the MCMC method in terms of computational time, whilst obtaining comparable statistical inference results for parameter estimation and variable selection under various error distributions. Notably, the algorithm of VB Huberised adaptive Lasso quantile regression is found to yield the fastest computational speed amongst other methods, as the adaptive Lasso penalty handles the over-fitting issue of the Lasso penalty efficiently that increases the convergence rate of the ELBO. To this extent, the VB algorithms are preferable over the MCMC method for a fast computational performance within a high-dimensional setting.

One could continue this investigation and even contemplate some of the alternative proposals in the next section.

5.2 Future Research

This is in a fascinating field of statistics, and there are still many open problems one could examine.

Chapter 2

The methodology is presented for fitting a Bayesian quantile-based FP model to the real-life data via the variable selection method. In this approach, a FP model is chosen at a fixed degree manually. Based on the power set $\mathrm{S}^{\mathrm{FP}} = \{-2, -1, -1/2, 0, 1/2, 1, 2, 3\}$, for a univariate FP model, the number of possible FPs with degree $m = 0$, 1, 2 and 3 is 1, 8, 36 and 120, respectively. The model space complexity increases exponentially as a function of the number of predictors. For example, for a maximum degree $m = 2$ and $k = 5$ predictors, there exist $(1 + 8 + 36)5 = 184,528,125$ different possible FP models. This is computationally burdensome in searching for the best fitted model. For Bayesian mean-based FP approach, Sabanés Bové and Held (2011) adopted a stochastic search algorithm and Bayesian model averaging in a trans-dimensional setting. We could follow this idea by extending our quantile-based FP approach to use reversible jump MCMC method, proposed by Green (1995), where one can automatically select the best fitted FP model according to different move types within a trans-dimensional setting. Given the multiple FP model,

$$\eta(\boldsymbol{x}) = \sum_{l=1}^{k} f_l^{m_l}(x_l; \boldsymbol{\alpha}_l, \boldsymbol{p}_l) = \sum_{l=1}^{k} \sum_{j=1}^{m_l} \alpha_{lj} h_{lj}(x_l),$$

we randomly select one of the following four move types:

- Birth: Randomly select one of the predictors with FP degree $m_l < m_{\max}$ $(l = 1, \ldots, k)$. Add a power to its $\boldsymbol{p}_l$ after randomly drawing it from S^{FP}.
- Death: Randomly select one of the predictors with FP degree $m_l > 0$. Remove a randomly chosen power from its $\boldsymbol{p}_l$.
- Move: Randomly select one of the predictors with FP degree $m_l > 0$. Remove a randomly chosen power from its $\boldsymbol{p}_l$, then randomly draw a power from S^{FP} and add it to $\boldsymbol{p}_l$.

- Switch: Randomly select one of the covariates with non-empty power vector $\boldsymbol{p}_l$. Randomly select one of the other covariates with power vector $\boldsymbol{p}_q$ $(q \neq l)$. Switch the power vectors $\boldsymbol{p}_l$ and $\boldsymbol{p}_q$.

The reversible jump MCMC method can be constructed by using the Metropolis-Hastings algorithm with four move types in a trans-dimensional setting. This iterative sampling algorithm would delete, add, switch and move FP powers to the model until the best FP model is obtained, whilst converging to a target posterior distribution.

Chapter 3

The work has several potential research directions for developing new models in theory and a wide variety of applications in many areas.

- The asymmetric Huberised loss function could replace the quantile loss function of quantile regression forests, proposed by Meinshausen and Ridgeway (2006), which present a new minimisation problem for random forests. Predictive intervals could also be derived to make statistical inference. Within the Bayesian paradigm, we add a prior distribution on each tree for random sampling of many decision trees, which is analogous to the work of Quadrianto and Ghahramani (2014). Then theoretically, we may develop new Bayesian robust quantile regression forests.

- Robust image analysis is the ongoing hot research topic in machine learning and is challenging to work with. The Soft Huber loss function (Equation (1.2)) and the Non-convex Huber loss function (Equation (1.3)) have been applied to the face recognition problems with noisy pixel data (Li et al. (2020)). We only need to replace these loss functions with our loss function and would develop an EM algorithm, consisting of Expectation and Maximisation steps, due to the scale mixture of normal property, with Bayesian robust regularisation. This is more robust due to the nature of conditional quantile function and robustness parameter. Other applications may also be explored, such as pattern recognition and speech recognition.

- The proposed models would easily be extended to a new type of time series models, such as quantile autoregressive models (Cai et al. (2012), and Yang et al. (2023)). In theory, we may develop a new Bayesian Huberised regularised quantile autoregressive model for robust high-dimensional time series estimation and forecasting, as well as apply to a wide variety of financial and economical applications.

- We may explore some medical applications, for example, the analysis of high-dimensional gene expression data. Currently, different variants of robust regularised regression models are used to analyse these data, whilst having the ability of being robust against outliers (see Ahdesmäki et al. (2007), Ren et al. (2023), and Algamal et al. (2018), amongst others). The proposed methods are better alternatives to provide comprehensive information about data based on the conditional quantile function and being robust to outliers simultaneously.

Chapter 4

This chapter has utilised an approximate-based technique to deal with slow-computational problems. Although the proposed algorithms are successful for moderate data, they may not scale to massive data because the modified Bessel function of the second kind involved in the algorithms grows exponentially as a function of predictors and sample size, and becomes infinite, which makes it infeasible in a software implementation. Some big data techniques may be explored, such as a divide-and-conquer strategy and bag of little bootstrap.

A divide-and-conquer approach is one of the most common big data strategies and is being conceptually and computationally appealing. It consists of three stages:

- First stage: partition the data independently.
- Second stage: perform local inference for each partitioned data.
- Third stage: combine the results to obtain a global approximation via aggregation.

The bootstrap method is a non-parametric method and a resampling technique for assessing uncertainty. This method is relatively simple and does not require making distributional assumptions, which makes inference easier. The bootstrap theory has been established for variational inference (Chen et al. (2018)). Bag of little bootstraps, introduced by Kleiner et al. (2012), extends the bootstrap method to the large-scale data setting. It has proven that it increases the computational speed by reducing the amount of information being passed between processors (Kleiner et al. (2012); Garnatz and Hardin (2021)).

This leaves an open problem for future research work.

Appendix A

Mathematical Proofs

Here we establish the proof for a range of propositions and theorem used in Chapter 3.

Proposition A.1. *If a random variable X follows the density function in Equation (3.2) then we have $P(X \leq \mu) = \tau$ and $P(X > \mu) = 1 - \tau$.*

Proof. We set $\mu = 0$ and we wish to calculate $P(X \leq 0)$, that is,

$$\begin{aligned} P(X \leq 0) &= \int_{-\infty}^{0} f_X(x)dx \\ &= \frac{\eta\tau(1-\tau)e^{\eta}}{2\rho^2(\eta+1)} \int_{-\infty}^{0} \exp\left\{-\sqrt{\eta\left(\eta - \frac{x(1-\tau)}{\rho^2}\right)}\right\} dx \\ &= \frac{\eta\tau(1-\tau)e^{\eta}}{2\rho^2(\eta+1)} \int_{0}^{\infty} \exp\left\{-\sqrt{\eta\left(\eta + \frac{x(1-\tau)}{\rho^2}\right)}\right\} dx. \end{aligned}$$

By letting $u = \sqrt{\eta\left(\eta + (x(1-\tau))/\rho^2\right)}$, we have

$$\begin{aligned} P(X \leq 0) &= \frac{\eta\tau(1-\tau)e^{\eta}}{2\rho^2(\eta+1)} \int_{\eta}^{\infty} e^{-u} \times \frac{2u\rho^2}{\eta(1-\tau)} du \\ &= \frac{\tau e^{\eta}}{(\eta+1)} \int_{\eta}^{\infty} u e^{-u} du \\ &= \frac{\tau e^{\eta}}{(\eta+1)} \left(\left[-ue^{-u}\right]_{\eta}^{\infty} + \int_{\eta}^{\infty} e^{-u} du \right) \\ &= \frac{\tau e^{\eta}}{(\eta+1)} \left(\eta e^{-\eta} + \left[-e^{-u}\right]_{\eta}^{\infty} \right) \\ &= \frac{\tau e^{\eta}}{(\eta+1)} \left(e^{-\eta}(\eta+1) \right) \\ &= \tau. \end{aligned}$$

On the other hand, it follows that $P(X > 0) = 1 - \tau$. This completes the proof.

□

Theorem A.1. *If the model error $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}$ follows the density function (Equation (3.2)), then we can represent ϵ_i as the scale mixture of normals given by*

$$
\begin{aligned}
&f(\epsilon_i; \tau, \eta, \rho^2) \\
&\quad \propto \iint N(\epsilon_i; (1-2\tau)v_i, 4v_i\sigma_i)\, Exp\left(v_i; \frac{\tau(1-\tau)}{2\sigma_i}\right) GIG\left(\sigma_i; \frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right) dv_i d\sigma_i, \\
&i = 1, \ldots, n,
\end{aligned}
$$

where GIG$(\cdot)$ denotes the GIG distribution and its density is specified by Equation (2.6), Exp$(\cdot)$ denotes the exponential distribution, and N$(\cdot)$ is the normal distribution.

Proof. Let a, b be some real constants. By using the equality

$$
\exp(-|ab|) = \int_0^\infty \frac{a}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}(a^2\sigma + b^2\sigma^{-1})\right\} d\sigma, \tag{A.1}
$$

(Andrews and Mallows (1974)) and let $a = \sqrt{\eta/\rho^2}$ and $b = \sqrt{\eta\rho^2 + \epsilon_i\,(\tau - \mathrm{I}(\epsilon_i < 0))}$, $f(\epsilon_i)$ can be expressed as the scale mixture of ALD and GIG densities:

$$
\begin{aligned}
&\frac{\eta\tau(1-\tau)e^\eta}{2\rho^2(\eta+1)} \exp\left\{-\sqrt{\eta\left(\eta + \frac{\epsilon_i}{\rho^2}\,(\tau - \mathrm{I}(\epsilon_i < 0))\right)}\right\} \\
&\qquad \propto \int_0^\infty \mathrm{AL}\,(\epsilon_i; 0, 2\sigma_i, \tau)\, \mathrm{GIG}\left(\sigma_i; \frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right) d\sigma_i
\end{aligned}
$$

where GIG$(\cdot)$ denotes the GIG distribution and its density is given by Equation (2.6), and AL$(\cdot)$ is the ALD and its density is given by Equation (1.8).

The ALD can be expressed as the scale mixture of normal and exponential densities using the equality (Equation (A.1)) by letting $a = 1/\sqrt{4\sigma_i}$, $b = \epsilon_i/\sqrt{4\sigma_i}$ and multiplying a factor of $\exp\{-((2\tau-1)\epsilon)/(4\sigma_i)\}$ (Kozumi and Kobayashi (2011)). Therefore, $f(\epsilon_i)$ is expressed as the scale mixture of normal, exponential and GIG densities:

$$
\begin{aligned}
&\frac{\eta\tau(1-\tau)e^\eta}{2\rho^2(\eta+1)} \exp\left\{-\sqrt{\eta\left(\eta + \frac{\epsilon_i}{\rho^2}\,(\tau - \mathrm{I}(\epsilon_i < 0))\right)}\right\} \\
&\qquad \propto \int_0^\infty \int_0^\infty \mathrm{N}(\epsilon_i; (1-2\tau)v_i, 4\sigma_i v_i) \mathrm{Exp}\left(v_i; \frac{\tau(1-\tau)}{2\sigma_i}\right) \mathrm{GIG}\left(\sigma_i; \frac{3}{2}, \frac{\eta}{\rho^2}, \eta\rho^2\right) d\sigma_i dv_i,
\end{aligned}
$$

where $\mathrm{N}(\cdot)$ and $\mathrm{Exp}(\cdot)$ are the normal and exponential densities, respectively.

□

Proposition A.2. *Let $\rho^2 \sim \pi(\rho^2) \propto 1/\rho^2$ (improper scale invariant prior). For fixed $\lambda_1 > 0$ and $\eta > 0$, the posterior distribution is proper for all n.*

Proof. The overall posterior distribution is given by

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{s}|\boldsymbol{y}) \\
&\quad = \frac{\pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})}{\int\int\int\int\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2}.
\end{aligned}$$

We show that the normalising constant of the posterior distribution is finite, that is,

$$\int\int\int\int\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2 < \infty.$$

First, we consider the integral with respect to $\boldsymbol{\beta}$. We have

$$\begin{aligned}
&\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2)d\boldsymbol{\beta} \\
&\qquad = \int \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\} \\
&\qquad\qquad \times \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\rho^2 s_j}} \exp\left\{-\frac{\beta_j^2}{2\rho^2 s_j}\right\} d\boldsymbol{\beta} \\
&\qquad = \int (8\pi)^{-n/2}(2\pi)^{-k/2}(\rho^2)^{-k/2} \left(\prod_{i=1}^{n}\sigma_i\right)^{-1/2} \left(\prod_{i=1}^{n} v_i\right)^{-1/2} \left(\prod_{j=1}^{k} s_j\right)^{-1/2} \\
&\qquad\qquad \times \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad\qquad \times \exp\left\{-\frac{1}{2\rho^2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}\right\} d\boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{V} = \text{diag}(4\sigma_1 v_1, \ldots, 4\sigma_n v_n)$ and $\boldsymbol{\Lambda} = \text{diag}(s_1, \ldots, s_k)$. In particular, we have

$$
\begin{aligned}
&\int \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})\right\} \times \exp\left\{-\frac{1}{2\rho^2}\boldsymbol{\beta}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\beta}\right\} d\boldsymbol{\beta} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \frac{1}{\rho^2}\boldsymbol{\Lambda}^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right)\right\} d\boldsymbol{\beta} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2}\left|\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} + \frac{1}{\rho^2}\boldsymbol{\Lambda}^{-1}\right)^{-1}\right|^{1/2} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2}\left|\frac{1}{\rho^2}\boldsymbol{\Lambda}^{-1}\right|^{-1/2} |\boldsymbol{V}|^{1/2} \left|\boldsymbol{V} + \rho^2\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^T\right|^{-1/2} \qquad \text{(A.2)} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2} 2^n (\rho^2)^{k/2}\left(\prod_{j=1}^{k} s_j\right)^{1/2}\left(\prod_{i=1}^{n}\sigma_i\right)^{1/2}\left(\prod_{i=1}^{n} v_i\right)^{1/2} \left|\boldsymbol{V} + \rho^2\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^T\right|^{-1/2}.
\end{aligned}
$$

The expression in Equation (A.2) is due to the identity of $|I + AB| = |I + BA|$ (Henderson and Searle (1981)).

Hence, we have

$$
\begin{aligned}
&\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2) d\boldsymbol{\beta} \\
&\qquad = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad\qquad \times \left|\boldsymbol{V} + \rho^2\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^T\right|^{-1/2}.
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s},\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2 \\
&= \iiiint (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \left|\boldsymbol{V}+\rho^2\boldsymbol{X}\boldsymbol{\Lambda}\boldsymbol{X}^T\right|^{-1/2}\prod_{i=1}^{n}\frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2 \\
&\leq \iiiint (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\}|\boldsymbol{V}|^{-1/2} \\
&\qquad \times \prod_{i=1}^{n}\frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2,
\end{aligned}
$$

by using the fact that $|A+B| \geq |A|$ implies $|A+B|^{-1/2} \leq |A|^{-1/2}$ for a positive definite matrix A and a semi-positive definite matrix B.

Next, we consider the integral with respect to $\boldsymbol{v}$. First, we have

$$
\begin{aligned}
&\int |\boldsymbol{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \prod_{i=1}^{n}\frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}d\boldsymbol{v} \\
&= \int \left(\frac{\tau(1-\tau)}{2}\right)^n 2^{-n}\left(\prod_{i=1}^{n}\sigma_i\right)^{-3/2}\left(\prod_{i=1}^{n}v_i\right)^{-1/2} \\
&\qquad \times \prod_{i=1}^{n}\exp\left\{-\frac{(y_i-(1-2\tau)v_i)^2}{8\sigma_i v_i}-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}d\boldsymbol{v} \\
&= \left(\frac{\tau(1-\tau)}{4}\right)^n\left(\prod_{i=1}^{n}\sigma_i\right)^{-3/2} \\
&\qquad \times \int \prod_{i=1}^{n}v_i^{-1/2}\exp\left\{-\frac{(y_i-(1-2\tau)v_i)^2}{8\sigma_i v_i}-\frac{\left(1-(1-2\tau)^2\right)v_i}{8\sigma_i}\right\}d\boldsymbol{v} \\
&= \left(\frac{\tau(1-\tau)}{4}\right)^n\left(\prod_{i=1}^{n}\sigma_i\right)^{-3/2} \\
&\qquad \times \int \prod_{i=1}^{n}v_i^{-1/2}\exp\left\{-\frac{y_i^2}{8\sigma_i v_i}-\frac{(1-2\tau)y_i}{4\sigma_i}-\frac{v_i}{8\sigma_i}\right\}d\boldsymbol{v} \\
&= \left(\frac{\tau(1-\tau)}{4}\right)^n\left(\prod_{i=1}^{n}\sigma_i\right)^{-3/2} \\
&\qquad \times \int \prod_{i=1}^{n}v_i^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{v_i}{4\sigma_i}+\frac{y_i^2}{4\sigma_i v_i}\right)\right\}\exp\left\{-\frac{(1-2\tau)y_i}{4\sigma_i}\right\}d\boldsymbol{v}.
\end{aligned}
$$

Letting $a^2 = 1/(4\sigma_i)$ and $b^2 = y_i^2/(4\sigma_i)$ and using the equality (Equation (A.1)), we have

$$\begin{aligned}
&\int |\boldsymbol{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\} d\boldsymbol{v} \\
&= \left(\frac{\tau(1-\tau)}{4}\right)^n \left(\prod_{i=1}^{n} \sigma_i^{-3/2} \exp\left\{-\frac{(1-2\tau)y_i}{4\sigma_i}\right\}\right) \\
&\qquad \times \int \prod_{i=1}^{n} v_i^{-1/2} \exp\left\{-\frac{1}{2}\left(a^2 v_i + b^2 v_i^{-1}\right)\right\} d\boldsymbol{v} \\
&= \left(\frac{\tau(1-\tau)}{4}\right)^n \left(\prod_{i=1}^{n} \sigma_i^{-3/2} \exp\left\{-\frac{(1-2\tau)y_i}{4\sigma_i}\right\}\right) \\
&\qquad \times \prod_{i=1}^{n} (2\pi)^{1/2} (4\sigma_i)^{1/2} \exp\left\{-\frac{|y_i|}{4\sigma_i}\right\} \\
&= (2\pi)^{n/2} \left(\frac{\tau(1-\tau)}{2}\right)^n \prod_{i=1}^{n} \sigma_i^{-1} \exp\left\{-\frac{|y_i| + (1-2\tau)y_i}{4\sigma_i}\right\}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma}) \pi(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2) \pi(\boldsymbol{v}|\boldsymbol{\sigma}) \pi(\sigma|\rho^2) \pi(\rho^2) \pi(\boldsymbol{s}) d\boldsymbol{\beta} d\boldsymbol{v} d\boldsymbol{\sigma} d\boldsymbol{s} d\rho^2 \\
&\qquad \leq \iiint \left(\frac{\tau(1-\tau)}{2}\right)^n \prod_{i=1}^{n} \sigma_i^{-1} \exp\left\{-\frac{|y_i| + (1-2\tau)y_i}{4\sigma_i}\right\} \\
&\qquad\qquad \times \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right)\right\} \pi(\rho^2) \pi(\boldsymbol{s}) d\boldsymbol{\sigma} d\boldsymbol{s} d\rho^2 \\
&\qquad = \iiint \left(\frac{\tau(1-\tau)}{2}\right)^n \left(\frac{1}{2\rho^2 K_{3/2}(\eta)}\right)^n \\
&\qquad\qquad \times \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right) - \frac{|y_i| + (1-2\tau)y_i}{4\sigma_i}\right\} \pi(\rho^2) \pi(\boldsymbol{s}) d\boldsymbol{\sigma} d\boldsymbol{s} d\rho^2.
\end{aligned}$$

Next, we consider the integral with respect to $\boldsymbol{\sigma}$. First, we have

$$\begin{aligned}
&\int \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right) - \frac{|y_i| + (1-2\tau)y_i}{4\sigma_i}\right\} d\boldsymbol{\sigma} \\
&\quad = \int \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\eta\sigma_i}{\rho^2} + \left(\eta\rho^2 + \frac{|y_i| + (1-2\tau)y_i}{2}\right)\frac{1}{\sigma_i}\right)\right\}) d\boldsymbol{\sigma}.
\end{aligned}$$

Letting $c^2 = \eta/\rho^2$ and $d^2 = \eta\rho^2 + (|y_i| + (1-2\tau)y_i)/2$ and using the equality (Equation

(A.1)), we have

$$\begin{aligned}
&\int \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2}+\frac{\rho^2}{\sigma_i}\right)-\frac{|y_i|+(1-2\tau)y_i}{4\sigma_i}\right\} d\boldsymbol{\sigma} \\
&\quad = \int \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{1}{2}\left(c^2\sigma_i + d^2\sigma_i^{-1}\right)\right\} d\boldsymbol{\sigma} \\
&\quad = \prod_{i=1}^{n} \left(\frac{\eta}{\rho^2}\right)^{-1/2} (2\pi)^{1/2} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s},\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s}) d\boldsymbol{\beta} d\boldsymbol{v} d\boldsymbol{\sigma} d\boldsymbol{s} d\rho^2 \\
&\quad \leq \iint (2\pi)^{n/2} \left(\frac{\tau(1-\tau)}{2}\right)^n \left(\frac{1}{2\sqrt{\eta\rho^2}K_{3/2}(\eta)}\right)^n \\
&\qquad \times \prod_{i=1}^{n} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{\lambda_1^2}{2} \exp\left\{-\frac{\lambda_1^2 s_j}{2}\right\} \pi(\rho^2) d\boldsymbol{s} d\rho^2 \\
&\quad = \int (2\pi)^{n/2} \left(\frac{\tau(1-\tau)}{2}\right)^n \left(\frac{1}{2\sqrt{\eta\rho^2}K_{3/2}(\eta)}\right)^n \\
&\qquad \times \prod_{i=1}^{n} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\} \times \frac{1}{\rho^2} d\rho^2 \\
&\quad = \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2(\eta)}}\right)^n \int (\rho^2)^{-n/2-1} \prod_{i=1}^{n} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\}.
\end{aligned} \tag{A.3}$$

In Equation (A.3), we note that the inequality

$$\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i|+(1-2\tau)y_i}{2}\right)} = \sqrt{\eta^2 + \eta\frac{|y_i|+(1-2\tau)y_i}{2\rho^2}} \geq \sqrt{\frac{\eta}{\rho^2}\left(\frac{|y_i|+(1-2\tau)y_i}{2}\right)}$$

holds for any $\eta > 0$ for $i = 1, \ldots, n$. Hence, we have

$$\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{s},\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2 \\
&\quad\leq \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2(\eta)}}\right)^n \int (\rho^2)^{-n/2-1} \prod_{i=1}^n \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\} d\rho^2 \\
&\quad= \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2(\eta)}}\right)^n \int \sqrt{\rho^2}^{-n-2} \exp\left\{-\frac{1}{\sqrt{\rho^2}}\sqrt{\frac{\eta}{2}}\sum_{i=1}^n \sqrt{|y_i|+(1-2\tau)y_i}\right\} d\rho^2.
\end{aligned}$$

By using the transformation $\sqrt{\rho^2} = x$, we have

$$\begin{aligned}
&\int \sqrt{\rho^2}^{-n-2} \exp\left\{-\frac{1}{\sqrt{\rho^2}}\sqrt{\frac{\eta}{2}}\sum_{i=1}^n \sqrt{|y_i|+(1-2\tau)y_i}\right\} d\rho^2 \\
&\quad= 2\int x^{-n-1} \exp\left\{-\frac{1}{x}\sqrt{\frac{\eta}{2}}\sum_{i=1}^n \sqrt{|y_i|+(1-2\tau)y_i}\right\} dx. \qquad \text{(A.4)}
\end{aligned}$$

Since the integrand is the kernel of IG $\left(n, \sqrt{\eta/2}\sum_{i=1}^n \sqrt{|y_i|+(1-2\tau)y_i}\right)$ where IG$(\cdot)$ is the inverse Gamma distribution, the integral is finite for any n. Hence, the posterior distribution under the improper prior $\pi(\rho^2) \propto 1/\rho^2$ is proper for any n.

□

Proposition A.3. *Under the conditional prior for $\boldsymbol{\beta}$ given ρ^2 and fixed $\lambda_1 > 0$ and $\eta > 0$, the joint posterior $(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal with respect to $(\boldsymbol{\beta}, \rho^2)$.*

Proof. The joint posterior density of $(\boldsymbol{\beta}, \rho^2)$ is expressed by

$$\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) = \iint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{v})\pi(\boldsymbol{\beta}|\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)d\boldsymbol{v}d\boldsymbol{\sigma}.$$

First, we consider the integral with respect to $\boldsymbol{v}$. We have

$$
\begin{aligned}
&\int \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\sigma},\boldsymbol{v})\pi(\boldsymbol{v}|\boldsymbol{\sigma})d\boldsymbol{v} \\
&\quad= \int \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\} \\
&\qquad\times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\} d\boldsymbol{v} \\
&\quad= (8\pi)^{-n/2} \left(\frac{\tau(1-\tau)}{2}\right)^n \left(\prod_{i=1}^{n}\sigma_i\right)^{-3/2} \\
&\qquad\times \int \prod_{i=1}^{n} v_i^{-1/2} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i} - \frac{\tau(1-\tau)v_i}{2\sigma_i}\right\} d\boldsymbol{v} \\
&\quad= \left(\frac{\tau(1-\tau)}{2}\right)^n \prod_{i=1}^{n} \sigma_i^{-1} \exp\left\{-\frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{4\sigma_i}\right\}.
\end{aligned}
$$

Hence, we have

$$
\begin{aligned}
&\pi(\boldsymbol{\beta},\rho^2|\boldsymbol{y}) \\
&\quad= \pi(\boldsymbol{\beta}|\rho^2)\pi(\rho^2) \int \left(\frac{\tau(1-\tau)}{2}\right)^n \prod_{i=1}^{n} \sigma_i^{-1} \exp\left\{-\frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{4\sigma_i}\right\} \\
&\qquad\times \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right)\right\} d\boldsymbol{\sigma} \\
&\quad= \pi(\boldsymbol{\beta}|\rho^2)\pi(\rho^2)(\rho^2)^{-n} \left(\frac{\tau(1-\tau)}{4K_{3/2}(\eta)}\right)^n \\
&\qquad\times \int \prod_{i=1}^{n} \sigma_i^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\eta}{\rho^2\sigma_i} + \left(\eta\rho^2 + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2}\right)\frac{1}{\sigma_i}\right)\right\} d\boldsymbol{\sigma} \\
&\quad= \pi(\boldsymbol{\beta}|\rho^2)\pi(\rho^2)(\rho^2)^{-n} \left(\frac{\tau(1-\tau)}{4K_{3/2}(\eta)}\right)^n \\
&\qquad\times \prod_{i=1}^{n} \left(\frac{\eta}{\rho^2}\right)^{-1/2} (2\pi)^{1/2} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2}\right)}\right\} \\
&\quad\propto (\rho^2)^{-1}(\rho^2)^{-k/2}(\rho^2)^{-n}(\rho^2)^{-n/2} \prod_{j=1}^{k} \exp\left\{-\frac{\lambda_1|\beta_j|}{\sqrt{\rho^2}}\right\} \\
&\qquad\times \prod_{i=1}^{n} \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2}\right)}\right\} \\
&\quad= (\rho^2)^{-n/2-k/2-1} \exp\left\{-\frac{\lambda_1}{\sqrt{\rho^2}} \sum_{j=1}^{k} |\beta_j|\right\} \\
&\qquad\times \prod_{i=1}^{n} \exp\left\{-\sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}\right\}.
\end{aligned}
$$

Then the log posterior density is given by

$$\log \pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) \propto -\left(\frac{n}{2}+\frac{k}{2}+1\right)\log\rho^2 - \frac{\lambda_1}{\sqrt{\rho^2}}\|\boldsymbol{\beta}\|_1 - \sum_{i=1}^{n}\sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}. \tag{A.5}$$

Like Kawakami and Hashimoto (2023) and Cai and Sun (2021), we consider the coordinate transformation $\Phi \leftrightarrow \boldsymbol{\beta}/\sqrt{\rho^2}$, $\xi \leftrightarrow 1/\sqrt{\rho^2}$. In the transformation coordinate, Equation (A.5) is given by

$$(n+k+2)\log\xi - \lambda_1\|\Phi\|_1 - \sum_{i=1}^{n}\sqrt{\eta\left(\eta + \frac{\xi}{2}\left(|\xi y_i - \boldsymbol{x}_i\Phi| + (1-2\tau)(\xi y_i - \boldsymbol{x}_i\Phi)\right)\right)}. \tag{A.6}$$

Since the three terms in Equation (A.6) are log-concave, the joint posterior $\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal. This completes the proof.

□

Proposition A.4. *Let $\rho^2 \sim \pi(\rho^2) \propto 1/\rho^2$ (improper scale invariant prior). For fixed $\lambda_3 > 0$, $\lambda_4 > 0$ and $\eta > 0$, the posterior distribution is proper for all n.*

Proof. We follow the proof of Proposition A.2 in the similar manner. The overall posterior distribution is given by

$$\pi(\boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})}{\int\int\int\int\int\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2}.$$

We show that the normalising constant of the posterior distribution is finite, that is,

$$\int\int\int\int\int \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t}, \rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 < \infty.$$

First, we consider the integral with respect to $\boldsymbol{\beta}$. We have

$$\begin{aligned}
&\int \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t},\rho^2)d\boldsymbol{\beta} \\
&\quad = \int (8\pi)^{-n/2}\pi^{-k/2}\lambda_4^{k/2}(\rho^2)^{-k/2}\left(\prod_{i=1}^n \sigma_i\right)^{-1/2}\left(\prod_{i=1}^n v_i\right)^{-1/2}\left(\prod_{j=1}^k \frac{t_j}{t_j-1}\right)^{1/2} \\
&\qquad \times \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \exp\left\{-\frac{\lambda_4}{\rho^2}\boldsymbol{\beta}^T\boldsymbol{T}^{-1}\boldsymbol{\beta}\right\} d\boldsymbol{\beta},
\end{aligned}$$

where $\boldsymbol{V} = \mathrm{diag}(4\sigma_1 v_1, \ldots, 4\sigma_n v_n)$ and $\boldsymbol{T} = \mathrm{diag}((t_1-1)t_1^{-1}, \ldots, (t_n-1)t_n^{-1})$. In particular, we have

$$\begin{aligned}
&\int \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-(1-2\tau)\boldsymbol{v})\right\} \times \exp\left\{-\frac{\lambda_4}{\rho^2}\boldsymbol{\beta}^T\boldsymbol{T}^{-1}\boldsymbol{\beta}\right\} d\boldsymbol{\beta} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2}\left|\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}+\frac{2\lambda_4}{\rho^2}\boldsymbol{T}^{-1}\right)^{-1}\right|^{1/2} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2}\left|\frac{2\lambda_4}{\rho^2}\boldsymbol{T}^{-1}\right|^{-1/2}|\boldsymbol{V}|^{1/2}\left|\boldsymbol{V}+\frac{\rho^2}{2\lambda_4}\boldsymbol{X}\boldsymbol{T}^{-1}\boldsymbol{X}^T\right|^{-1/2} \\
&\quad = \exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times (2\pi)^{k/2}2^n 2^{k/2}(\rho^2)^{k/2}\lambda_4^{-k/2}\left(\prod_{j=1}^k \frac{t_j}{t_j-1}\right)^{-1/2}\left(\prod_{i=1}^n \sigma_i\right)^{1/2}\left(\prod_{i=1}^n v_i\right)^{1/2} \\
&\qquad \times \left|\boldsymbol{V}+\frac{\rho^2}{2\lambda_4}\boldsymbol{X}\boldsymbol{T}^{-1}\boldsymbol{X}^T\right|^{-1/2}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\int \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t},\rho^2)d\boldsymbol{\beta} \\
&\quad = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \left|\boldsymbol{V}+\frac{\rho^2}{2\lambda_4}\boldsymbol{X}\boldsymbol{T}^{-1}\boldsymbol{X}^T\right|^{-1/2}.
\end{aligned}$$

Next, we have

$$\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t},\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 \\
&\quad = \iiiint (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \left|\boldsymbol{V}+\frac{\rho^2}{2\lambda_4}\boldsymbol{X}\boldsymbol{T}^{-1}\boldsymbol{X}^T\right|^{-1/2} \pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 \\
&\quad \leq \iiiint (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\}|\boldsymbol{V}|^{-1/2} \\
&\qquad \times \prod_{i=1}^{n}\frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{s})d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{s}d\rho^2.
\end{aligned}$$

Next, we consider the integral with respect to $\boldsymbol{v}$. We have

$$\begin{aligned}
&\int |\boldsymbol{V}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})^T\boldsymbol{V}^{-1}(\boldsymbol{y}-(1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \prod_{i=1}^{n}\frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}d\boldsymbol{v} \\
&\quad = (2\pi)^{n/2}\left(\frac{\tau(1-\tau)}{2}\right)^n\prod_{i=1}^{n}\sigma_i^{-1}\exp\left\{-\frac{|y_i|+(1-2\tau)y_i}{4\sigma_i}\right\}.
\end{aligned}$$

Hence, we have

$$
\begin{aligned}
&\iiiint \pi(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})\pi(\boldsymbol{\beta}|\boldsymbol{t},\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\sigma|\rho^2)\pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\beta}d\boldsymbol{v}d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 \\
&\quad\leq \iiint \left(\frac{\tau(1-\tau)}{2}\right)^n \prod_{i=1}^n \sigma_i^{-1} \exp\left\{-\frac{|y_i|+(1-2\tau)y_i}{4\sigma_i}\right\} \\
&\qquad\times \prod_{i=1}^n \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2}+\frac{\rho^2}{\sigma_i}\right)\right\} \pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 \\
&\quad= \iiint \left(\frac{\tau(1-\tau)}{4K_{3/2}(\eta)}\right)^n (\rho^2)^{-n} \\
&\qquad\times \prod_{i=1}^n \sigma_i^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{\eta\sigma_i}{\rho^2}+\left(\eta\rho^2+\frac{|y_i|+(1-2\tau)y_i}{2}\right)\frac{1}{\sigma_i}\right)\right\} \pi(\rho^2)\pi(\boldsymbol{t})d\boldsymbol{\sigma}d\boldsymbol{t}d\rho^2 \\
&\quad= \iint \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2}(\eta)}\right)^n (\rho^2)^{-n/2} \prod_{i=1}^n \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2+\frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\} \\
&\qquad\times \prod_{j=1}^k \Gamma^{-1}\left(\frac{1}{2},\tilde{\lambda}_3\right)\sqrt{\frac{\tilde{\lambda}_3}{t_j}} \exp\left\{-\tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j>1) \times \frac{1}{\rho^2} d\boldsymbol{t}d\rho^2 \\
&\quad= \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2}(\eta)}\right)^n \tilde{\lambda}_3^{-k}\Gamma^{-k}\left(\frac{1}{2},\tilde{\lambda}_3\right)\Gamma^k\left(\frac{1}{2},1\right) \\
&\qquad\times \int (\rho^2)^{-n/2-1} \prod_{i=1}^n \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2+\frac{|y_i|+(1-2\tau)y_i}{2}\right)}\right\} d\rho^2 \\
&\quad\leq \left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2}(\eta)}\right)^n \tilde{\lambda}_3^{-k}\Gamma^{-k}\left(\frac{1}{2},\tilde{\lambda}_3\right)\Gamma^k\left(\frac{1}{2},1\right) \\
&\qquad\times 2\int x^{-n-1} \exp\left\{-\frac{1}{x}\sqrt{\frac{\eta}{2}}\sum_{i=1}^n \sqrt{|y_i|+(1-2\tau)y_i}\right\} dx.
\end{aligned}
$$

As the integrand is the same as that in Equation (A.4), the integral is finite for any n. Hence, the posterior distribution under the improper prior $\pi(\rho^2) \propto 1/\rho^2$ is proper for any n.

□

Proposition A.5. *Under the conditional prior for $\boldsymbol{\beta}$ given ρ^2 and fixed $\lambda_3 > 0$, $\lambda_4 > 0$ and $\eta > 0$, the joint posterior $(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal with respect to $(\boldsymbol{\beta}, \rho^2)$.*

Proof. We follow the proof of Proposition A.3 in the similar manner. The joint posterior

density of $(\boldsymbol{\beta}, \rho^2)$ is expressed by

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) &= \iint \pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{v})\pi(\boldsymbol{\beta}|\rho^2)\pi(\boldsymbol{v}|\boldsymbol{\sigma})\pi(\boldsymbol{\sigma}|\rho^2)\pi(\rho^2)d\boldsymbol{v}d\boldsymbol{\sigma} \\
&= \pi(\boldsymbol{\beta}|\rho^2)\pi(\rho^2)\left(\frac{\sqrt{\pi}\tau(1-\tau)}{\sqrt{8\eta}K_{3/2}(\eta)}\right)^n (\rho^2)^{-n/2} \\
&\quad \times \prod_{i=1}^n \exp\left\{-\sqrt{\frac{\eta}{\rho^2}\left(\eta\rho^2 + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2}\right)}\right\} \\
&\propto (\rho^2)^{-n/2-k/2-1} \exp\left\{-\frac{\lambda_3}{\sqrt{\rho^2}}\sum_{j=1}^k |\beta_j| - \frac{\lambda_4}{\rho^2}\sum_{j=1}^k \beta_j^2\right\} \\
&\quad \times \prod_{i=1}^n \exp\left\{-\sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}\right\}.
\end{aligned}$$

Then the log posterior density is given by

$$\begin{aligned}
\log \pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y}) \propto &-\left(\frac{n}{2} + \frac{k}{2} + 1\right)\log\rho^2 - \frac{\lambda_3}{\sqrt{\rho^2}}\|\boldsymbol{\beta}\|_1 - \frac{\lambda_4}{\rho^2}\|\boldsymbol{\beta}\|_2^2 \\
&- \sum_{i=1}^n \sqrt{\eta\left(\eta + \frac{|y_i - \boldsymbol{x}_i\boldsymbol{\beta}| + (1-2\tau)(y_i - \boldsymbol{x}_i\boldsymbol{\beta})}{2\rho^2}\right)}. \qquad \text{(A.7)}
\end{aligned}$$

We also consider the coordinate transformation $\Phi \leftrightarrow \boldsymbol{\beta}/\sqrt{\rho^2}$, $\xi \leftrightarrow 1/\sqrt{\rho^2}$. In the transformation coordinate, Equation (A.7) is given by

$$\begin{aligned}
&(n + k + 2)\log\xi - \lambda_1\|\Phi\|_1 - \lambda_4\|\Phi\|_2^2 \\
&\quad - \sum_{i=1}^n \sqrt{\eta\left(\eta + \frac{\xi}{2}\left(|\xi y_i - \boldsymbol{x}_i\Phi| + (1-2\tau)(\xi y_i - \boldsymbol{x}_i\Phi)\right)\right)}. \qquad \text{(A.8)}
\end{aligned}$$

Since the four terms in Equation (A.8) are log-concave, the joint posterior of $\pi(\boldsymbol{\beta}, \rho^2|\boldsymbol{y})$ is unimodal. This completes the proof.

□

Appendix B

Details of Gibbs Sampling Algorithms

Here we provide the full Gibbs sampling algorithms of hierarchical models for Bayesian Huberised regularised quantile regression used in Chapter 3.

B.1 Bayesian Huberised Lasso Quantile Regression

The joint posterior distribution is as follows.

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \lambda_1, \boldsymbol{s} | \boldsymbol{y}) \\
&\quad = \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{ -\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i} \right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{ -\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i} \right) \right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{ -\frac{\tau(1-\tau)v_i}{2\sigma_i} \right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\rho^2 s_j}} \exp\left\{ -\frac{\beta_j^2}{2\rho^2 s_j} \right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{\lambda_1^2}{2} \exp\left\{ -\frac{\lambda_1^2 s_j}{2} \right\} \\
&\qquad \times \frac{b^a}{\Gamma(a)} (\lambda_1^2)^{a-1} \exp\left\{ -b\lambda_1^2 \right\} \\
&\qquad \times \frac{d^c}{\Gamma(c)} \eta^{c-1} \exp\left\{ -d\eta \right\} \\
&\qquad \times \frac{1}{\rho^2}.
\end{aligned}
$$

The full conditional posterior distribution of $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
&\pi(\boldsymbol{\beta} | \boldsymbol{y}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \lambda_1, \boldsymbol{s}) \\
&\quad \propto \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{ -\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i} \right\} \\
&\qquad \times \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\rho^2 s_j}} \exp\left\{ -\frac{\beta_j^2}{2\rho^2 s_j} \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v}) \right\} \\
&\qquad \times \exp\left\{ -\frac{1}{2\rho^2}\boldsymbol{\beta}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{\beta} \right\} \\
&\quad \propto \exp\left\{ -\frac{1}{2}\left(\boldsymbol{\beta}^T \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} + \frac{1}{\rho^2}\boldsymbol{\Lambda}^{-1} \right) \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - (1-2\tau)\boldsymbol{v}) \right) \right\} \\
&\quad \propto \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \right),
\end{aligned}
$$

where $\boldsymbol{V} = \mathrm{diag}(4\sigma_1 v_1, \ldots, 4\sigma_n v_n)$, $\boldsymbol{\Lambda} = \mathrm{diag}(s_1, \ldots, s_k)$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} + 1/\rho^2 \boldsymbol{\Lambda}^{-1} \right)^{-1}$ and $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \boldsymbol{X}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - (1-2\tau)\boldsymbol{v})$.

The full conditional posterior distribution of σ_i, $i = 1, \ldots, n$, is given by

$$
\begin{aligned}
&\pi(\sigma_i|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \lambda_1, \boldsymbol{s}) \\
&\quad \propto \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right)\right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\} \\
&\quad \propto \sigma_i^{-1} \exp\left\{-\frac{1}{2}\left(\frac{\eta}{\rho^2}\sigma_i + \left(\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{4v_i} + \tau(1-\tau)v_i + \eta\rho^2\right)\frac{1}{\sigma_i}\right)\right\} \\
&\quad \propto \text{GIG}\left(0, \frac{\eta}{\rho^2}, \frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{4v_i} + \tau(1-\tau)v_i + \eta\rho^2\right).
\end{aligned}
$$

The full conditional posterior distribution of v_i, $i = 1, \ldots, n$, is given by

$$
\begin{aligned}
&\pi(v_i|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{\sigma}, \lambda_1, \boldsymbol{s}) \\
&\qquad \propto \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\} \\
&\qquad\quad \times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\} \\
&\qquad \propto v_i^{-1/2} exp\left\{-\frac{1}{2}\left(\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{4\sigma_i v_i} + \frac{\tau(1-\tau)v_i}{\sigma_i}\right)\right\} \\
&\qquad \propto v_i^{-1/2} exp\left\{-\frac{1}{2}\left(\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{4\sigma_i}\frac{1}{v_i} + \left(\frac{(1-2\tau)^2}{4\sigma_i} + \frac{\tau(1-\tau)}{\sigma_i}\right)v_i\right)\right\} \\
&\qquad \propto \text{GIG}\left(\frac{1}{2}, \frac{(1-2\tau)^2}{4\sigma_i} + \frac{\tau(1-\tau)}{\sigma_i}, \frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2}{4\sigma_i}\right).
\end{aligned}
$$

The full conditional posterior distribution of ρ^2 is given by

$$\begin{aligned}
&\pi(\rho^2|\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma},\lambda_1,\boldsymbol{s}) \\
&\qquad\propto \prod_{i=1}^{n}\frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)}\exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2}+\frac{\rho^2}{\sigma_i}\right)\right\} \\
&\qquad\quad\times\prod_{j=1}^{k}\frac{1}{\sqrt{2\pi\rho^2 s_j}}\exp\left\{-\frac{\beta_j^2}{2\rho^2 s_j}\right\} \\
&\qquad\quad\times\frac{1}{\rho^2} \\
&\qquad\propto (\rho^2)^{-n-\frac{k}{2}-1}\exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\eta}{\sigma_i}\rho^2+\left(\sum_{i=1}^{n}\eta\sigma_i+\sum_{j=1}^{k}\frac{\beta_j^2}{s_j}\right)\frac{1}{\rho^2}\right)\right\} \\
&\qquad\propto \text{GIG}\left(-n-\frac{k}{2},\sum_{i=1}^{n}\frac{\eta}{\sigma_i},\sum_{i=1}^{n}\eta\sigma_i+\sum_{j=1}^{k}\frac{\beta_j^2}{s_j}\right).
\end{aligned}$$

The full conditional posterior distribution of s_j, $j=1,\ldots,k$, is given by

$$\begin{aligned}
&\pi(s_j|\boldsymbol{y},\boldsymbol{\beta},\rho^2,\boldsymbol{v},\boldsymbol{\sigma},\lambda_1) \\
&\qquad\propto \frac{1}{\sqrt{2\pi\rho^2 s_j}}\exp\left\{-\frac{\beta_j^2}{2\rho^2 s_j}\right\}\times\frac{\lambda_1^2}{2}\exp\left\{-\frac{\lambda_1^2 s_j}{2}\right\} \\
&\qquad\propto s_j^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{\beta_j^2}{\rho^2}\frac{1}{s_j}+\lambda_1^2 s_j\right)\right\} \\
&\qquad\propto \text{GIG}\left(\frac{1}{2},\lambda_1^2,\frac{\beta_j^2}{\rho^2}\right).
\end{aligned}$$

The full conditional posterior distribution of λ_1 is given by

$$\begin{aligned}
&\pi(\lambda_1|\boldsymbol{y},\boldsymbol{\beta},\rho^2,\boldsymbol{v},\boldsymbol{\sigma},\boldsymbol{s}) \\
&\qquad\propto \prod_{j=1}^{k}\frac{\lambda_1^2}{2}\exp\left\{-\frac{\lambda_1^2 s_j}{2}\right\} \\
&\qquad\quad\times\frac{b^a}{\Gamma(a)}(\lambda_1^2)^{a-1}\exp\left\{-b\lambda_1^2\right\} \\
&\qquad\propto (\lambda_1^2)^{a+k-1}\exp\left\{-\left(b+\sum_{j=1}^{k}\frac{s_j}{2}\right)\lambda_1^2\right\} \\
&\qquad\propto \text{Gamma}\left(a+k,b+\sum_{j=1}^{k}\frac{s_j}{2}\right).
\end{aligned}$$

B.2 Bayesian Huberised Elastic Net Quantile Regression

The joint posterior distribution is as follows.

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4 | \boldsymbol{y}) \\
&\quad = \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{ -\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i} \right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{ -\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i} \right) \right\} \\
&\qquad \times \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i} \exp\left\{ -\frac{\tau(1-\tau)v_i}{2\sigma_i} \right\} \\
&\qquad \times \prod_{j=1}^{k} \sqrt{\frac{\lambda_4 t_j}{\pi\rho^2(t_j - 1)}} \exp\left\{ -\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j - 1)} \right\} \\
&\qquad \times \prod_{j=1}^{k} \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3 \right) \sqrt{\frac{\tilde{\lambda}_3}{t_j}} \exp\left\{ -\tilde{\lambda}_3 t_j \right\} \mathrm{I}(t_j > 1) \\
&\qquad \times \frac{b_1^{a_1}}{\Gamma(a_1)} (\tilde{\lambda}_3)^{a_1 - 1} \exp\left\{ -b_1 \tilde{\lambda}_3 \right\} \\
&\qquad \times \frac{b_2^{a_2}}{\Gamma(a_2)} \lambda_2^{a_2 - 1} \exp\left\{ -b_2 \lambda_4 \right\} \\
&\qquad \times \frac{b_3^{a_3}}{\Gamma(a_3)} \eta^{a_3 - 1} \exp\left\{ -b_3 \eta \right\} \\
&\qquad \times \frac{1}{\rho^2}.
\end{aligned}
$$

Clearly, it is obvious to see that the full conditional posterior distributions of σ_i and v_i, $i = 1, \ldots, n$ are the same as those in the Bayesian Huberised Lasso quantile regression.

The full conditional posterior distribution of $\boldsymbol{\beta}$ is given by

$$
\begin{aligned}
&\pi(\boldsymbol{\beta}|\boldsymbol{y}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4) \\
&\quad = \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\} \\
&\qquad \times \prod_{j=1}^{k} \sqrt{\frac{\lambda_4 t_j}{\pi\rho^2(t_j-1)}} \exp\left\{-\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j-1)}\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})^T \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - (1-2\tau)\boldsymbol{v})\right\} \\
&\qquad \times \exp\left\{-\frac{\lambda_4}{\rho^2}\boldsymbol{\beta}^T\boldsymbol{T}^{-1}\boldsymbol{\beta}\right\} \\
&\quad \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}^T\left(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X} + \frac{2\lambda_4}{\rho^2}\boldsymbol{T}^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})\right)\right\} \\
&\quad \propto \mathrm{N}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right),
\end{aligned}
$$

where $\boldsymbol{V} = \mathrm{diag}(4\sigma_1 v_1, \ldots, 4\sigma_n v_n)$, $\boldsymbol{T} = \mathrm{diag}((t_1-1)t_1^{-1}, \ldots, (t_n-1)t_n^{-1})$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X} + 2\lambda_4/\rho^2\boldsymbol{T}^{-1}\right)^{-1}$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{X}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - (1-2\tau)\boldsymbol{v})$.

The full conditional posterior distribution of ρ^2 is given by

$$
\begin{aligned}
&\pi(\rho^2|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4) \\
&\quad \propto \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)} \exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right)\right\} \\
&\qquad \times \prod_{j=1}^{k} \sqrt{\frac{\lambda_4 t_j}{\pi\rho^2(t_j-1)}} \exp\left\{-\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j-1)}\right\} \\
&\qquad \times \frac{1}{\rho^2} \\
&\quad \propto (\rho^2)^{-n-\frac{k}{2}-1} \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\eta}{\sigma_i}\rho^2 + \left(\sum_{i=1}^{n}\eta\sigma_i + \sum_{j=1}^{k}\frac{2\lambda_4 t_j \beta_j^2}{t_j - 1}\right)\frac{1}{\rho^2}\right)\right\} \\
&\quad \propto \mathrm{GIG}\left(-n - \frac{k}{2}, \sum_{i=1}^{n}\frac{\eta}{\sigma_i}, \sum_{i=1}^{n}\eta\sigma_i + \sum_{j=1}^{k}\frac{2t_j\lambda_4\beta_j^2}{t_j - 1}\right).
\end{aligned}
$$

The full conditional posterior distribution of $t_j - 1$ is given by

$$\begin{aligned}
&\pi(t_j - 1|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \lambda_3, \lambda_4) \\
&\qquad \propto \sqrt{\frac{\lambda_4 t_j}{\pi\rho^2(t_j-1)}} \exp\left\{-\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j-1)}\right\} \\
&\qquad\quad \times \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right) \sqrt{\frac{\tilde{\lambda}_3}{t_j}} \exp\left\{-\tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j > 1) \\
&\qquad \propto (t_j - 1)^{-1/2} \exp\left\{-\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j-1)} - \tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j > 1) \\
&\qquad \propto (t_j - 1)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{2\lambda_4\beta_j^2}{\rho^2}\frac{1}{t_j - 1} + 2\tilde{\lambda}_3(t_j - 1)\right)\right\} \mathrm{I}(t_j - 1 > 0) \\
&\qquad \propto \mathrm{GIG}\left(\frac{1}{2}, 2\tilde{\lambda}_3, \frac{2\lambda_4\beta_j^2}{\rho^2}\right) \mathrm{I}(t_j - 1 > 0).
\end{aligned}$$

The full conditional posterior distribution of $\tilde{\lambda}_3$ is given by

$$\begin{aligned}
&\pi(\tilde{\lambda}_3|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4) \\
&\qquad \propto \prod_{j=1}^{k} \Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right) \sqrt{\frac{\tilde{\lambda}_3}{t_j}} \exp\left\{-\tilde{\lambda}_3 t_j\right\} \mathrm{I}(t_j > 1) \\
&\qquad\quad \times \frac{b_1^{a_1}}{\Gamma(a_1)} (\tilde{\lambda}_3)^{a_1 - 1} \exp\left\{-b_1\tilde{\lambda}_3\right\} \\
&\qquad \propto \Gamma^{-k}\left(\frac{1}{2}, \tilde{\lambda}_3\right) (\tilde{\lambda}_3)^{\frac{k}{2} + a_1 - 1} \exp\left\{-\left(\sum_{j=1}^{k} t_j + b_1\right)\tilde{\lambda}_3\right\}.
\end{aligned}$$

As it is infeasible to directly sample from $\pi(\tilde{\lambda}_3|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4)$, the one-step Metropolis-Hastings algorithm is employed. Following Li et al. (2010), the proposal distribution is $q(\tilde{\lambda}_3|\boldsymbol{t}) \sim \mathrm{Gamma}\left(k + a_1, b_1 + \sum_{j=1}^{k}(t_j - 1)\right)$. They showed that

$$\lim_{\tilde{\lambda}_3 \to \infty} \frac{\sqrt{\tilde{\lambda}_3}\exp(\tilde{\lambda}_3)}{\Gamma^{-1}\left(\frac{1}{2}, \tilde{\lambda}_3\right)} = 1$$

implies that

$$\lim_{\tilde{\lambda}_3 \to \infty} \frac{\pi(\tilde{\lambda}_3|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4)}{q(\tilde{\lambda}_3|\boldsymbol{t})}$$

exists and equals to some positive constant. Hence, the tail behaviours of $q(\tilde{\lambda}_3|\boldsymbol{t})$ and $\pi(\tilde{\lambda}_3|\boldsymbol{y}, \boldsymbol{\beta}, \rho^2, \boldsymbol{v}, \boldsymbol{\sigma}, \boldsymbol{t}, \lambda_3, \lambda_4)$ are similar.

The full conditional posterior distribution of λ_4 is given by

$$
\begin{aligned}
&\pi(\lambda_4|\boldsymbol{y},\boldsymbol{\beta},\rho^2,\boldsymbol{v},\boldsymbol{\sigma},\boldsymbol{t},\lambda_3) \\
&\qquad \propto \prod_{j=1}^{k} \sqrt{\frac{\lambda_4 t_j}{\pi\rho^2(t_j-1)}} \exp\left\{-\frac{\lambda_4 t_j \beta_j^2}{\rho^2(t_j-1)}\right\} \\
&\qquad\qquad \times \frac{b_2^{a_2}}{\Gamma(a_2)} \lambda_2^{a_2-1} \exp\left\{-b_2\lambda_4\right\} \\
&\qquad \propto \lambda_4^{\frac{k}{2}+a_2-1} \exp\left\{-\left(\sum_{j=1}^{k} \frac{t_j\beta_j^2}{\rho^2(t_j-1)} + b_2\right)\lambda_4\right\} \\
&\qquad \propto \text{Gamma}\left(\frac{k}{2}+a_2, \sum_{j=1}^{k} \frac{t_j\beta_j^2}{\rho^2(t_j-1)} + b_2\right).
\end{aligned}
$$

Appendix C

Evidence Lower Bound

Here we provide the derivations of evidence lower bound for VB Huberised Lasso quantile regression and VB Huberised adaptive Lasso quantile regression used in Chapter 4.

C.1 Variational Bayesian Huberised Lasso Quantile Regression

According to the VI principle, minimising the KL divergence from $q(\Theta)$ to true posterior distribution $p(\Theta|\boldsymbol{y})$ is equivalent to maximising the lower bound of evidence $\log p(\boldsymbol{y})$, that is $\mathrm{LB}(q) = \mathbb{E}[\log p(\Theta, \boldsymbol{y})] - \mathbb{E}[\log q(\Theta)]$. For some function $g(\theta)$ of parameter $\theta \in \Theta$, we use $\mathbb{E}[g(\theta)]$ to represent the expectation of $g(\theta)$ about the optimal variational posterior distribution $q^*(\theta)$. ELBO is as follows. Based on the hierarchical model of Bayesian Huberised lasso quantile regression, we have

$$
\begin{aligned}
p(\Theta, \boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma}) p(\boldsymbol{v}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}|\rho^2, \eta) p(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2) p(\boldsymbol{s}|\lambda^2) p(\lambda^2) p(\rho^2) p(\eta) p(\beta_0), \\
q(\Theta) &= q(\boldsymbol{\beta}) q(\boldsymbol{s}) q(\boldsymbol{v}) q(\boldsymbol{\sigma}) q(\rho^2) q(\eta) q(\lambda^2) q(\beta_0).
\end{aligned}
$$

For the former, they are computed as follows.

- $\mathbb{E}[\log p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{v},\boldsymbol{\sigma})]$

$$= \mathbb{E}\left[\log \prod_{i=1}^{n} \frac{1}{\sqrt{8\pi\sigma_i v_i}} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i\boldsymbol{\beta} - (1-2\tau)v_i)^2}{8\sigma_i v_i}\right\}\right]$$

$$= -\frac{n}{2}\log 8\pi - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[\log\sigma_i] - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[\log v_i]$$

$$- \frac{1}{8}\sum_{i=1}^{n}\Big\{\mathbb{E}[\sigma_i^{-1}]\mathbb{E}[v_i^{-1}]\left((y_i - \mathbb{E}[\beta_0] - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}])^2 + \mathrm{Var}(\beta_0) + \mathrm{Tr}(\boldsymbol{x}_i\boldsymbol{x}_i^T\mathrm{Var}(\boldsymbol{\beta}))\right)$$

$$+ \frac{(1-2\tau)^2}{4}\mathbb{E}[\sigma_i^{-1}]\mathbb{E}[v_i] - \frac{1-2\tau}{2}\mathbb{E}[\sigma_i^{-1}](y_i - \mathbb{E}[\beta_0] - \boldsymbol{x}_i\mathbb{E}[\boldsymbol{\beta}])^2\Big\}.$$

- $\mathbb{E}[\log p(\boldsymbol{v}|\boldsymbol{\sigma})] = \mathbb{E}\left[\log \prod_{i=1}^{n} \frac{\tau(1-\tau)}{2\sigma_i}\exp\left\{-\frac{\tau(1-\tau)v_i}{2\sigma_i}\right\}\right]$

$$= n\log\frac{\tau(1-\tau)}{2} - \sum_{i=1}^{n}\mathbb{E}[\log\sigma_i] - \frac{\tau(1-\tau)}{2}\sum_{i=1}^{n}\mathbb{E}[\sigma_i^{-1}]\mathbb{E}[v_i].$$

- $\mathbb{E}[\log p(\boldsymbol{\sigma}|\rho^2,\eta)] = \mathbb{E}\left[\log \prod_{i=1}^{n} \frac{\sqrt{\sigma_i}}{2\rho^2 K_{3/2}(\eta)}\exp\left\{-\frac{\eta}{2}\left(\frac{\sigma_i}{\rho^2} + \frac{\rho^2}{\sigma_i}\right)\right\}\right]$

$$= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[\log\sigma_i] - n\log 2 - n\mathbb{E}[\log\rho^2] - n\mathbb{E}[\log K_{3/2}(\eta)]$$

$$- \frac{1}{2}\mathbb{E}[\eta]\sum_{i=1}^{n}(\mathbb{E}[\sigma_i]\mathbb{E}[\rho^2] + \mathbb{E}[\sigma_i^{-1}]\mathbb{E}[\rho^2]).$$

- $\mathbb{E}[\log p(\boldsymbol{\beta}|\boldsymbol{s},\rho^2)] = \mathbb{E}\left[\log \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi s_j\rho^2}}\exp\left\{-\frac{\beta_j^2}{2s_j\rho^2}\right\}\right]$

$$= -\frac{k}{2}\log 2\pi - \frac{k}{2}\mathbb{E}[\log\rho^2] - \frac{1}{2}\sum_{j=1}^{k}\mathbb{E}[\log s_j]$$

$$- \frac{1}{2}\mathbb{E}[(\rho^2)^{-1}]\sum_{j=1}^{k}\mathbb{E}[s_j^{-1}]\mathbb{E}[\beta_j^2].$$

- $\mathbb{E}[\log p(\boldsymbol{s}|\lambda^2)] = \mathbb{E}\left[\log \prod_{j=1}^{k} \frac{\lambda^2}{2}\exp\left\{-\frac{\lambda^2 s_j}{2}\right\}\right]$

$$= -k\log 2 + k\mathbb{E}[\log\lambda^2] - \frac{1}{2}\mathbb{E}[\lambda^2]\sum_{j=1}^{k}\mathbb{E}[s_j].$$

- $\mathbb{E}[\log p(\lambda^2)] = \mathbb{E}\left[\log\left(\frac{b^a}{\Gamma(a)}(\lambda^2)^{a-1}\exp\left\{-b\lambda^2\right\}\right)\right]$

$$= a\log b - \log\Gamma(a) + (a-1)\mathbb{E}[\log\lambda^2] - b\mathbb{E}[\lambda^2].$$

- $\mathbb{E}[\log p(\rho^2)] = \mathbb{E}\left[\log\frac{1}{\rho^2}\right] = -\mathbb{E}[\log\rho^2]$

- $\mathbb{E}[\log p(\eta)] = \mathbb{E}\left[\log\left(\frac{d^c}{\Gamma[c]}\eta^{c-1}\exp\{-d\eta\}\right)\right]$
$= c\log d - \log\Gamma(c) + (c-1)\mathbb{E}[\log\eta] - d\mathbb{E}[\eta].$
- $\mathbb{E}[\log p(\beta_0)] = 0.$

For the latter, they are also computed as follows.

- $\mathbb{E}[\log q^*(\boldsymbol{\beta})] = \mathbb{E}\left[\log\left((2\pi)^{-k/2}|\Sigma_{\boldsymbol{\beta}}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mu_{\boldsymbol{\beta}})^T\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\mu_{\boldsymbol{\beta}})\right\}\right)\right]$
$= -\frac{k}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| - \frac{k}{2}.$
- $\mathbb{E}[\log q^*(\boldsymbol{s})]$
$= \mathbb{E}\left[\log\prod_{j=1}^{k}\frac{1}{2}K_{1/2}^{-1}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right)\left(\frac{\hat{a}_{s_j}}{\hat{b}_{s_j}}\right)^{1/4}s_j^{-1/2}\exp\left\{-\frac{1}{2}\left(\hat{a}_{s_j}s_j + \hat{b}_{s_j}s_j^{-1}\right)\right\}\right]$
$= -k\log 2 - \sum_{j=1}^{k}\log K_{1/2}\left(\sqrt{\hat{a}_{s_j}\hat{b}_{s_j}}\right) + \frac{1}{4}\sum_{j=1}^{k}\left(\log\hat{a}_{s_j} - \log\hat{b}_{s_j}\right)$
$- \frac{1}{2}\sum_{j=1}^{k}\mathbb{E}[\log s_j] - \frac{1}{2}\sum_{j=1}^{k}\left(\hat{a}_{s_j}\mathbb{E}[s_j] + \hat{b}_{s_j}\mathbb{E}[s_j^{-1}]\right).$
- $\mathbb{E}[\log q^*(\boldsymbol{v})]$
$= \mathbb{E}\left[\log\prod_{i=1}^{n}\frac{1}{2}K_{1/2}^{-1}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right)\left(\frac{\hat{a}_{v_i}}{\hat{b}_{v_i}}\right)^{1/4}v_i^{-1/2}\exp\left\{-\frac{1}{2}\left(\hat{a}_{v_i}v_i + \hat{b}_{v_i}v_i^{-1}\right)\right\}\right]$
$= -n\log 2 - \sum_{i=1}^{n}\log K_{1/2}\left(\sqrt{\hat{a}_{v_i}\hat{b}_{v_i}}\right) + \frac{1}{4}\sum_{i=1}^{n}\left(\log\hat{a}_{v_i} - \log\hat{b}_{v_i}\right)$
$- \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[\log v_i] - \frac{1}{2}\sum_{i=1}^{n}\left(\hat{a}_{v_i}\mathbb{E}[v_i] + \hat{b}_{v_i}\mathbb{E}[v_i^{-1}]\right).$
- $\mathbb{E}[\log q^*(\boldsymbol{\sigma})]$
$= \mathbb{E}\left[\log\prod_{i=1}^{n}\frac{1}{2}K_0^{-1}\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right)\sigma_i^{-1}\exp\left\{-\frac{1}{2}\left(\hat{a}_{\sigma_i}\sigma_i + \hat{b}_{\sigma_i}\sigma_i^{-1}\right)\right\}\right]$
$= -n\log 2 - \sum_{i=1}^{n}\log K_0\left(\sqrt{\hat{a}_{\sigma_i}\hat{b}_{\sigma_i}}\right) - \sum_{i=1}^{n}\mathbb{E}[\log\sigma_i]$
$- \frac{1}{2}\sum_{i=1}^{n}\left(\hat{a}_{\sigma_i}\mathbb{E}[\sigma_i] + \hat{b}_{\sigma_i}\mathbb{E}[\sigma_i^{-1}]\right).$

- $\mathbb{E}[\log q^*(\rho^2)]$

$$
\begin{aligned}
&= \mathbb{E}\Bigg[\log \frac{1}{2} K^{-1}_{-n-\frac{k}{2}} \left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}} \right) \left(\frac{\hat{a}_{\rho^2}}{\hat{b}_{\rho^2}} \right)^{-\frac{n}{2}-\frac{k}{4}} \\
&\qquad \times (\rho^2)^{-n-\frac{k}{2}-1} \exp\left\{ -\frac{1}{2}\left(\hat{a}_{\rho^2}\rho^2 + \hat{b}_{\rho^2}(\rho^2)^{-1} \right) \right\} \Bigg] \\
&= -\log 2 - \log K_{-n-\frac{k}{2}} \left(\sqrt{\hat{a}_{\rho^2}\hat{b}_{\rho^2}} \right) - \left(\frac{n}{2} + \frac{k}{4} \right) \left(\log \hat{a}_{\rho^2} - \log \hat{b}_{\rho^2} \right) \\
&\quad - \left(n + \frac{k}{2} + 1 \right) \mathbb{E}[\log \rho^2] - \frac{1}{2}\left(\hat{a}_{\rho^2}\mathbb{E}[\rho^2] + \hat{b}_{\rho^2}\mathbb{E}[(\rho^2)^{-1}] \right).
\end{aligned}
$$

- $\mathbb{E}[\log q^*(\eta)]$

$$
\begin{aligned}
&= \mathbb{E}\left[\log \left(\frac{1}{K_{3/2}(\eta)^n} \eta^{c-1} \exp\left\{ -\frac{\eta}{2} \sum_{i=1}^{n} \left(\mathbb{E}[(\rho^2)^{-1}]\mathbb{E}[\sigma_i] + \mathbb{E}[\rho^2]\mathbb{E}[\sigma_i^{-1}] \right) \right\} \exp\{-\eta d\} \right) \right] \\
&= -n\mathbb{E}[\log K_{3/2}(\eta)] + (c-1)\mathbb{E}[\log \eta] - \mathbb{E}[\eta] \left(\frac{1}{2} \sum_{i=1}^{n} (\mathbb{E}[\sigma_i]\mathbb{E}[\rho^2] + \mathbb{E}[\sigma_i^{-1}]\mathbb{E}[\rho^2]) + d \right).
\end{aligned}
$$

- $\mathbb{E}[\log q^*(\lambda^2)] = \mathbb{E}\left[\log \left(\frac{\hat{b}_{\lambda^2}^{\hat{a}_{\lambda^2}}}{\Gamma(\hat{a}_{\lambda^2})} (\lambda^2)^{\hat{a}_{\lambda^2}-1} \exp\left\{ -\hat{b}_{\lambda^2}\lambda^2 \right\} \right) \right]$

$$
= \hat{a}_{\lambda^2} \log \hat{b}_{\lambda^2} - \log \Gamma(\hat{a}_{\lambda^2}) + (\hat{a}_{\lambda^2} - 1)\mathbb{E}[\log \lambda^2] - \hat{b}_{\lambda^2}\mathbb{E}[\lambda^2].
$$

- $\mathbb{E}[\log q^*(\beta_0)] = \mathbb{E}\left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2_{\beta_0}}} \exp\left\{ -\frac{(\beta_0 - \mu_{\beta_0})^2}{2\sigma^2_{\beta_0}} \right\} \right) \right]$

$$
= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2_{\beta_0} - \frac{1}{2}.
$$

From computing the ELBO above, we are required to calculate the expectation of the following functions of parameters: $\mathbb{E}[\log \sigma_i]$, $\mathbb{E}[\log v_i]$, $\mathbb{E}[\log s_j]$, $\mathbb{E}[\log \rho^2]$, $\mathbb{E}[\log \lambda^2]$, $\mathbb{E}[\log \eta]$ and $\mathbb{E}[\log K_{3/2}(\eta)]$. Because $q^*(\lambda^2) \sim \text{Gamma}\left(\hat{a}_{\lambda^2}, \hat{b}_{\lambda^2}\right)$, we have $\mathbb{E}[\log \lambda^2] = \psi(\hat{a}_{\lambda^2}) - \log \hat{b}_{\lambda^2}$, where $\psi(\cdot)$ is the digamma function. For $\mathbb{E}[\log \eta]$ and $\mathbb{E}[\log K_{3/2}(\eta)]$, the Taylor approximation is used to produce a second-order expansion at the expected value of each variable for approximation, in other words, $g(x) \approx f(x_0) + (x - x_0)g'(x_0) + \frac{1}{2}(x - x_0)^2 g"(x_0)$. Specifically, for $E[\log \eta]$, we generate a second-order expansion of the expectation of $\log \eta$ and let $x_0 = \mu_\eta$ then we have

$$
\begin{aligned}
E[\log \eta] &= \mathbb{E}\left[\log \mu_\eta + \frac{\eta - \mu_\eta}{\mu_\eta} - \frac{1}{2}\frac{(\eta - \mu_\eta)^2}{\mu_\eta^2} \right] \\
&= \log \mu_\eta - \frac{1}{2\mu_\eta^2}\mathbb{E}[(\eta - \mu_\eta)^2] \\
&= \log \mu_\eta - \frac{\text{Var}(\eta)}{2\mu_\eta^2}.
\end{aligned}
$$

Because $q^*(\eta) \sim \mathrm{N}(\mu_\eta, \sigma_\eta^2)$, we have $\mathbb{E}[\eta] = \mu_\eta$ and $\mathrm{Var}(\eta) = \sigma_\eta^2$.

Similarly, we have

$$\mathbb{E}[\log K_{3/2}(\eta)] = \log K_{3/2}(\mu_\eta) + \frac{1}{2}\mathrm{Var}(\eta)\frac{d^2}{d\eta^2}\log K_{3/2}(\mu_\eta).$$

Because $q^*(\sigma_i) \sim \mathrm{GIG}\left(0, \hat{a}_{\sigma_i}, \hat{b}_{\sigma_i}\right)$, $q^*(v_i) \sim \mathrm{GIG}\left(\frac{1}{2}, \hat{a}_{v_i}, \hat{b}_{v_i}\right)$, $q^*(s_j) \sim \mathrm{GIG}\left(\frac{1}{2}, \hat{a}_{s_j}, \hat{b}_{s_j}\right)$ and $q^*(\rho^2) \sim \mathrm{GIG}\left(-n - \frac{k}{2}, \hat{a}_{\rho^2}, \hat{b}_{\rho^2}\right)$, let $x \sim \mathrm{GIG}(\nu, a, b)$ be a GIG random variable then we have $\mathbb{E}[\log x] = \partial/\partial\nu \left[K_\nu\left(\sqrt{ab}\right)\right] K_\nu^{-1}\left(\sqrt{ab}\right) - 1/2 \log a/b$. Thus, $\mathbb{E}[\log \sigma_i]$, $\mathbb{E}[\log v_i]$, $\mathbb{E}[\log s_j]$ and $\mathbb{E}[\log \rho^2]$ are computed in the similar way.

By substituting the expectation of each parameter and the approximate expectation of the function, the approximate ELBO can be obtained.

C.2 Variational Bayesian Huberised Adaptive Lasso Quantile Regression

Based on the hierarchical model of Bayesian Huberised adaptive lasso quantile regression, we have

$$\begin{aligned} p(\Theta, \boldsymbol{y}) &= p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{v}, \boldsymbol{\sigma}) p(\boldsymbol{v}|\boldsymbol{\sigma}) p(\boldsymbol{\sigma}|\rho^2, \eta) p(\boldsymbol{\beta}|\boldsymbol{s}, \rho^2) p(\boldsymbol{s}|\boldsymbol{\lambda}^2) p(\boldsymbol{\lambda}^2) p(\rho^2) p(\eta) p(\beta_0), \\ q(\Theta) &= q(\boldsymbol{\beta}) q(\boldsymbol{s}) q(\boldsymbol{v}) q(\boldsymbol{\sigma}) q(\rho^2) q(\eta) q(\boldsymbol{\lambda}^2) q(\beta_0). \end{aligned}$$

For the former, they can be computed in the same way except

- $$\begin{aligned} \mathbb{E}[\log p(\boldsymbol{s}|\boldsymbol{\lambda}^2)] &= \mathbb{E}\left[\log \prod_{j=1}^{k} \frac{\lambda_j^2}{2} \exp\left\{-\frac{\lambda_j^2 s_j}{2}\right\}\right] \\ &= -k \log 2 + \sum_{j=1}^{k} \mathbb{E}[\log \lambda_j^2] - \frac{1}{2}\sum_{j=1}^{k} \mathbb{E}[\lambda_j^2]\mathbb{E}[s_j]. \end{aligned}$$
- $$\begin{aligned} \mathbb{E}[\log p(\boldsymbol{\lambda}^2)] &= \mathbb{E}\left[\log \prod_{j=1}^{k} \frac{b^a}{\Gamma(a)} (\lambda_j^2)^{a-1} \exp\left\{-b\lambda_j^2\right\}\right] \\ &= ka \log b - k \log \Gamma(a) + (a-1)\sum_{j=1}^{k} \mathbb{E}[\log \lambda_j^2] - b\sum_{j=1}^{k} \mathbb{E}[\lambda_j^2]. \end{aligned}$$

For the latter, they are also computed in the same way except

- $\mathbb{E}[\log q^*(\boldsymbol{\lambda}^2)] = \mathbb{E}\left[\log \prod_{j=1}^{k} \frac{\hat{b}_{\lambda_j^2}^{\hat{a}_{\lambda_j^2}}}{\Gamma(\hat{a}_{\lambda_j^2})} (\lambda_j^2)^{\hat{a}_{\lambda_j^2}-1} \exp\left\{-\hat{b}_{\lambda_j^2}\lambda_j^2\right\}\right]$
$= \sum_{j=1}^{k} \left\{\hat{a}_{\lambda_j^2} \log \hat{b}_{\lambda_j^2} - \log\Gamma(\hat{a}_{\lambda_j^2}) + (\hat{a}_{\lambda_j^2}-1)\mathbb{E}[\log\lambda_j^2] - \hat{b}_{\lambda_j^2}\mathbb{E}[\lambda_j^2]\right\}.$

Because $q^*(\lambda_j^2) \sim \text{Gamma}\left(\hat{a}_{\lambda_j^2}, \hat{b}_{\lambda_j^2}\right)$, we have $\mathbb{E}[\log\lambda_j^2] = \psi(\hat{a}_{\lambda_j^2}) - \log\hat{b}_{\lambda_j^2}$ for $j = 1, \ldots, k$.

Appendix D

Figures

D.1 Chapter 2 Data Analysis

Figure D.1: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.5$ under the Bayesian quantile regression model with FPs.

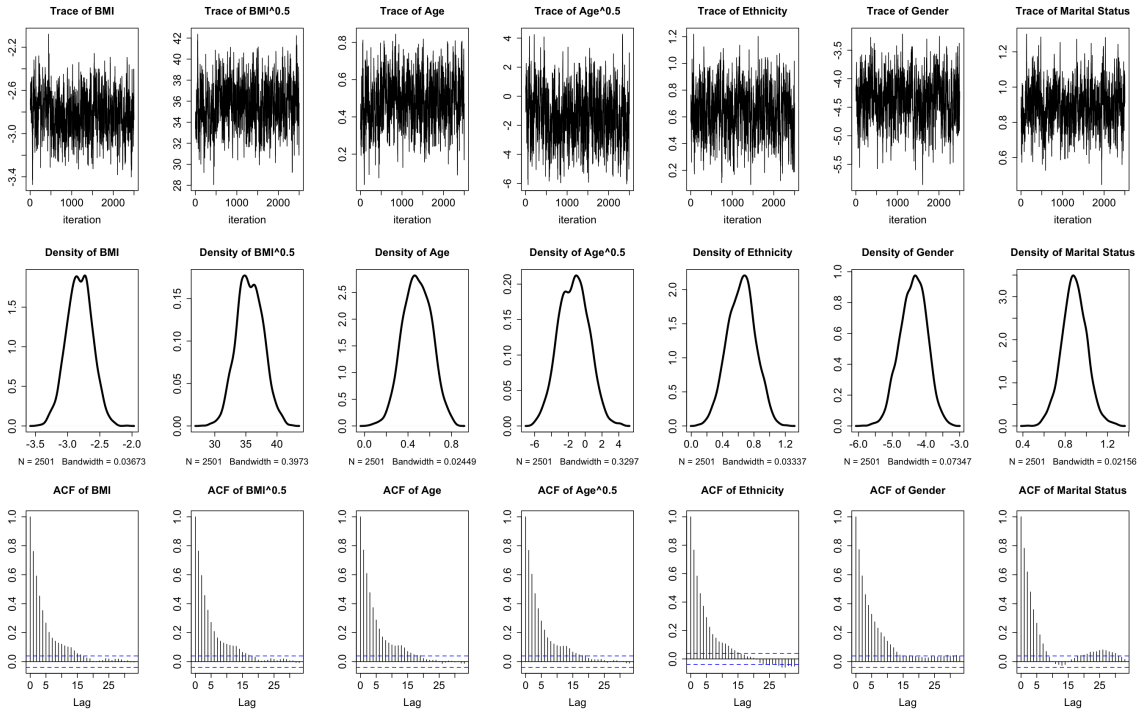


Figure D.2: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.5$ under the Bayesian quantile regression model with FPs.

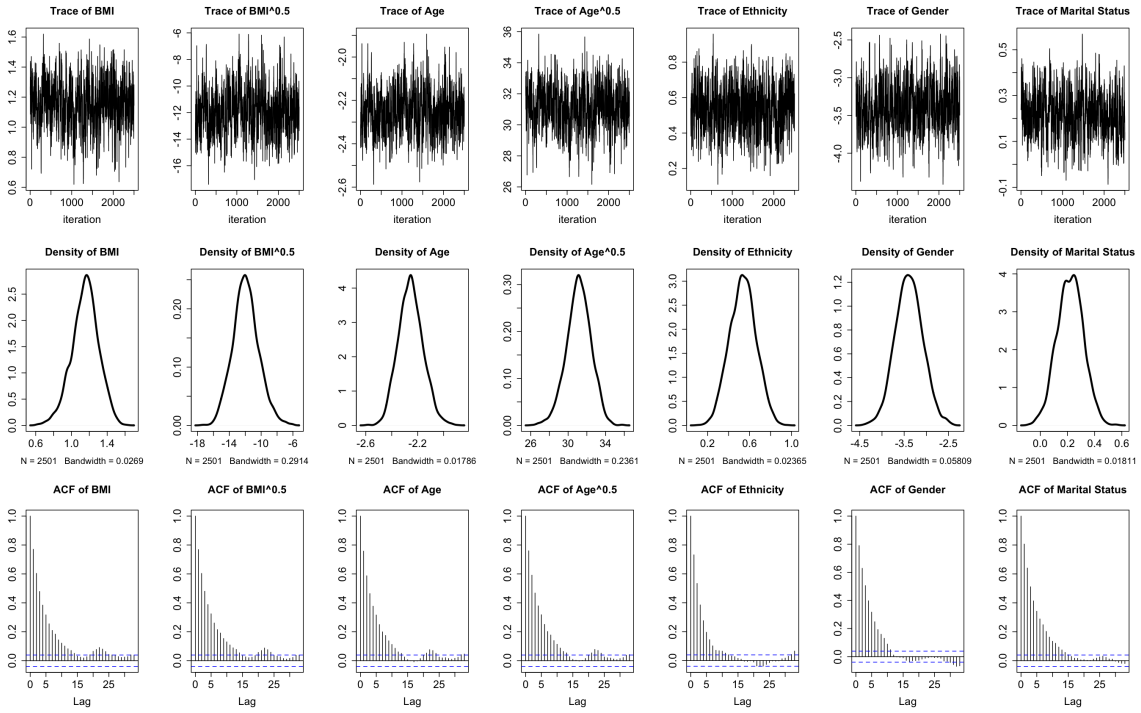


Figure D.3: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.5$ under the Bayesian quantile regression model with FPs and variable selection.

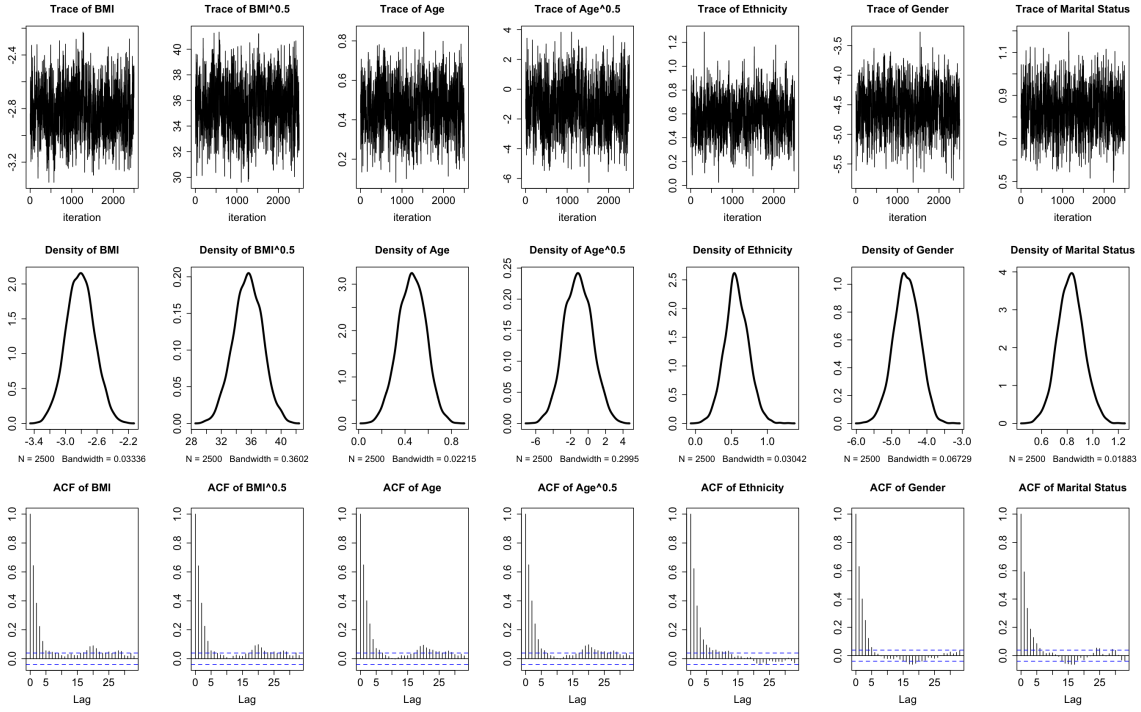


Figure D.4: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.5$ under the Bayesian quantile regression model with FPs and variable selection.

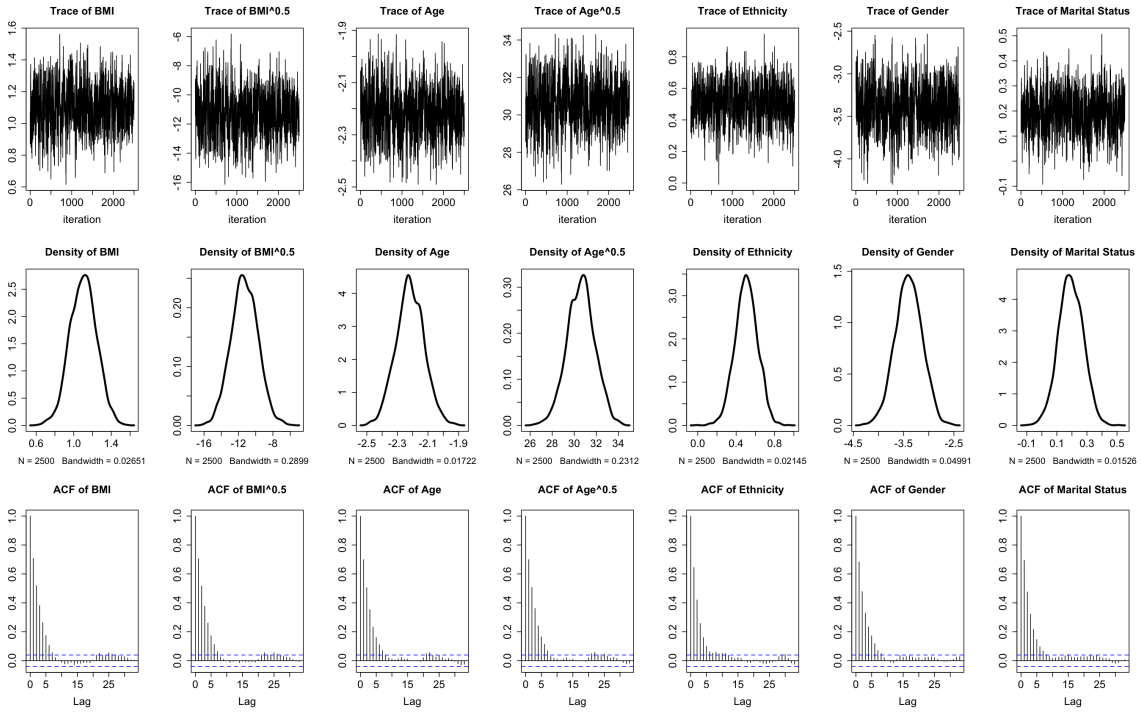


Figure D.5: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.75$ under the Bayesian quantile regression model with FPs.

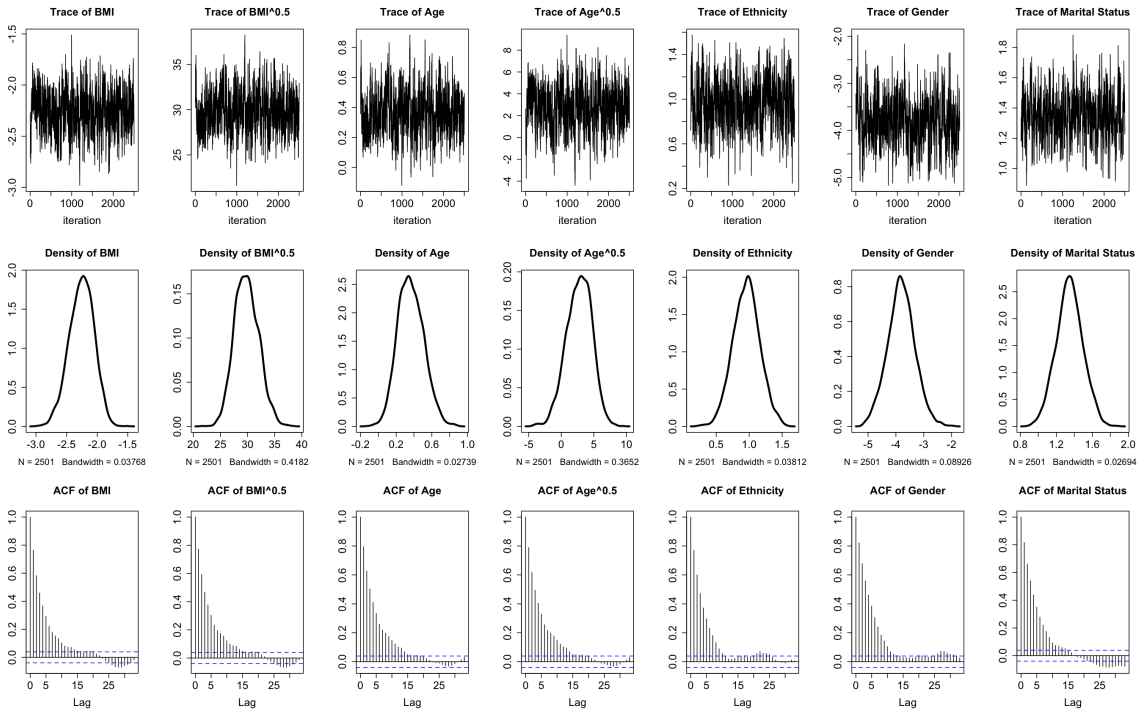


Figure D.6: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.75$ under the Bayesian quantile regression model with FPs.

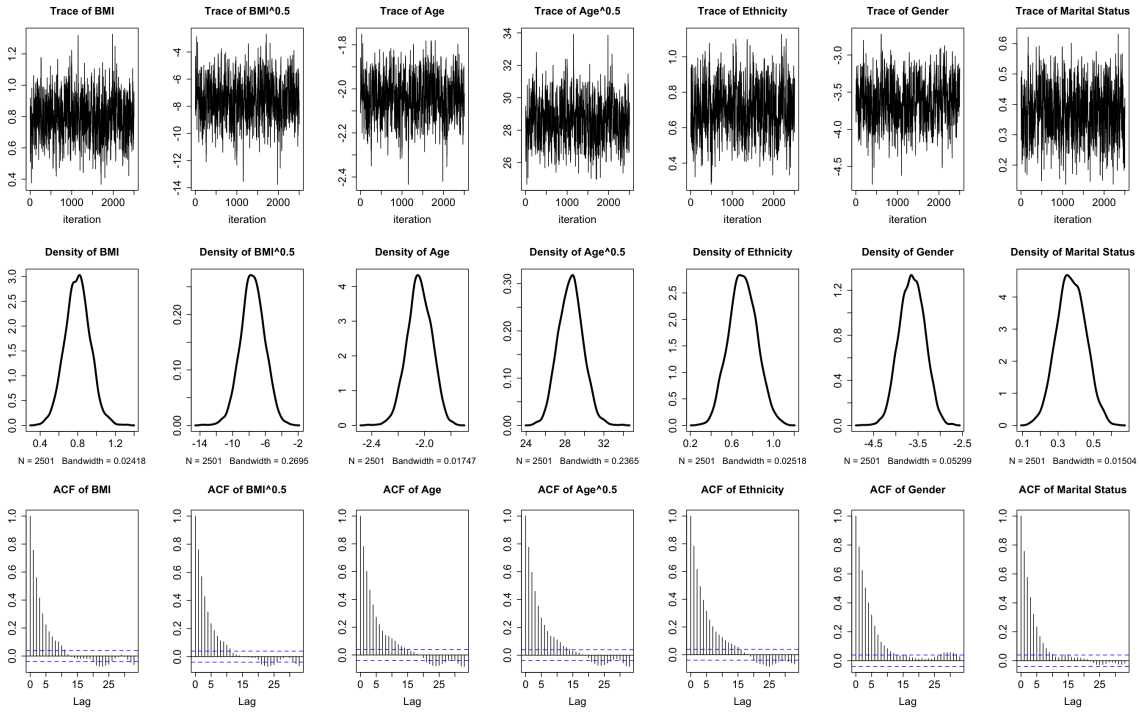


Figure D.7: Trace, density and autocorrelation plots for the risk factors of SBP at $\tau = 0.75$ under the Bayesian quantile regression model with FPs and variable selection.

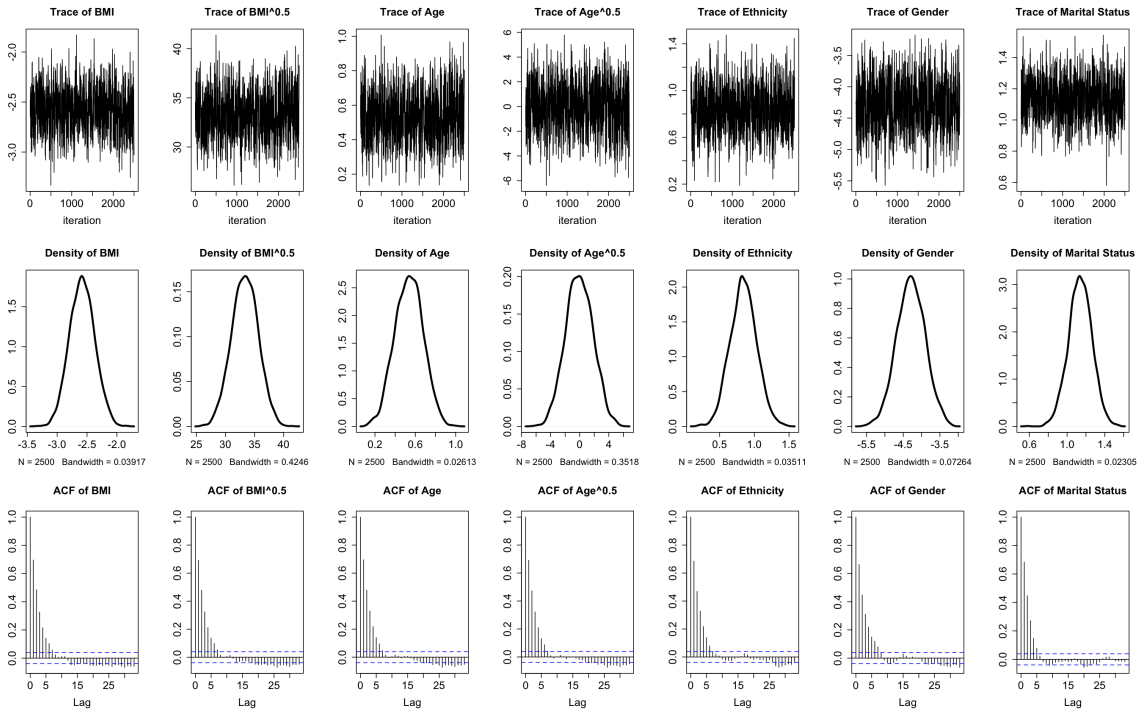
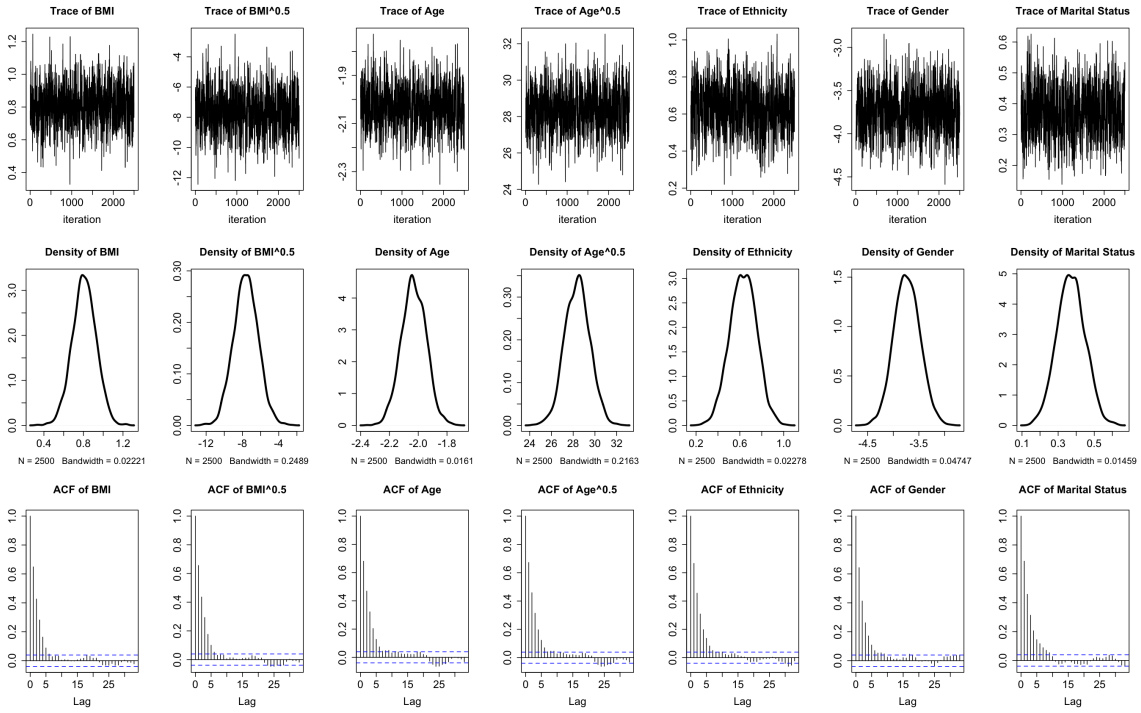


Figure D.8: Trace, density and autocorrelation plots for the risk factors of DBP at $\tau = 0.75$ under the Bayesian quantile regression model with FPs and variable selection.



D.2 Chapter 3 Simulation Studies

Figure D.9: Boxplots of RMSE based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.75$).

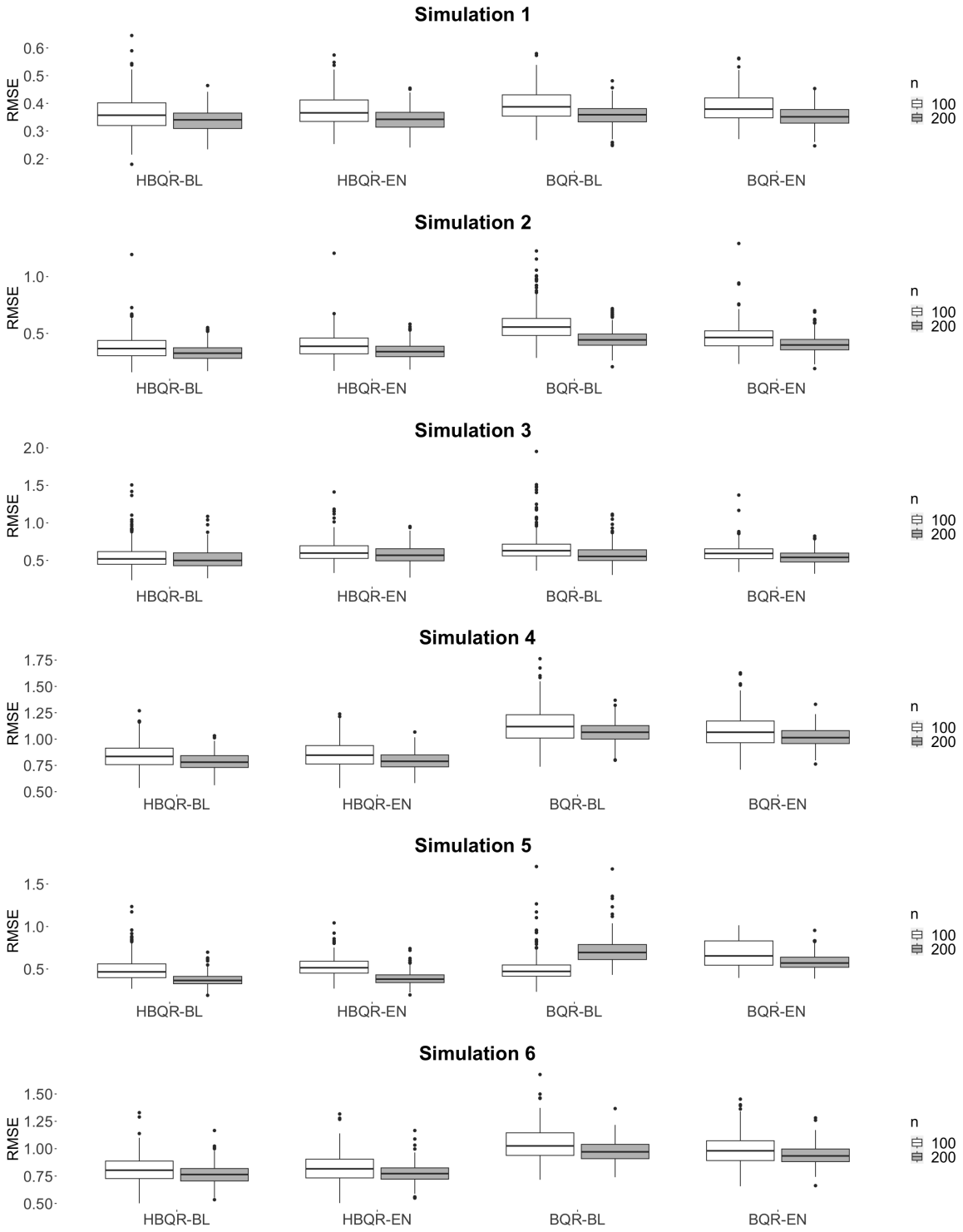


Figure D.10: Boxplots of MMAD based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.75$).

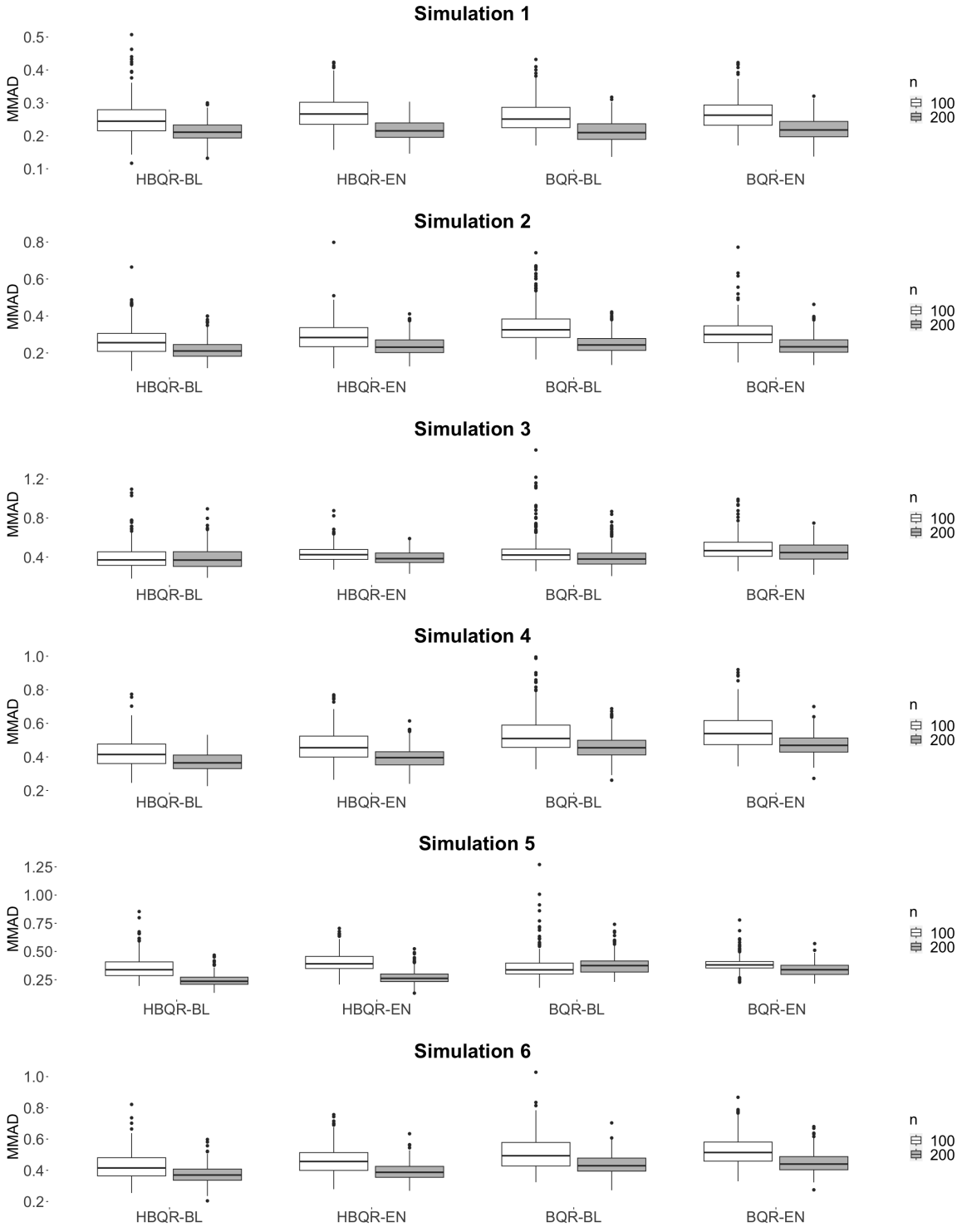


Figure D.11: Boxplots of AL based on 300 replications in six simulation scenarios for HBQR-BL, HBQR-EN, BQR-BL and BQR-EN ($\tau = 0.75$).

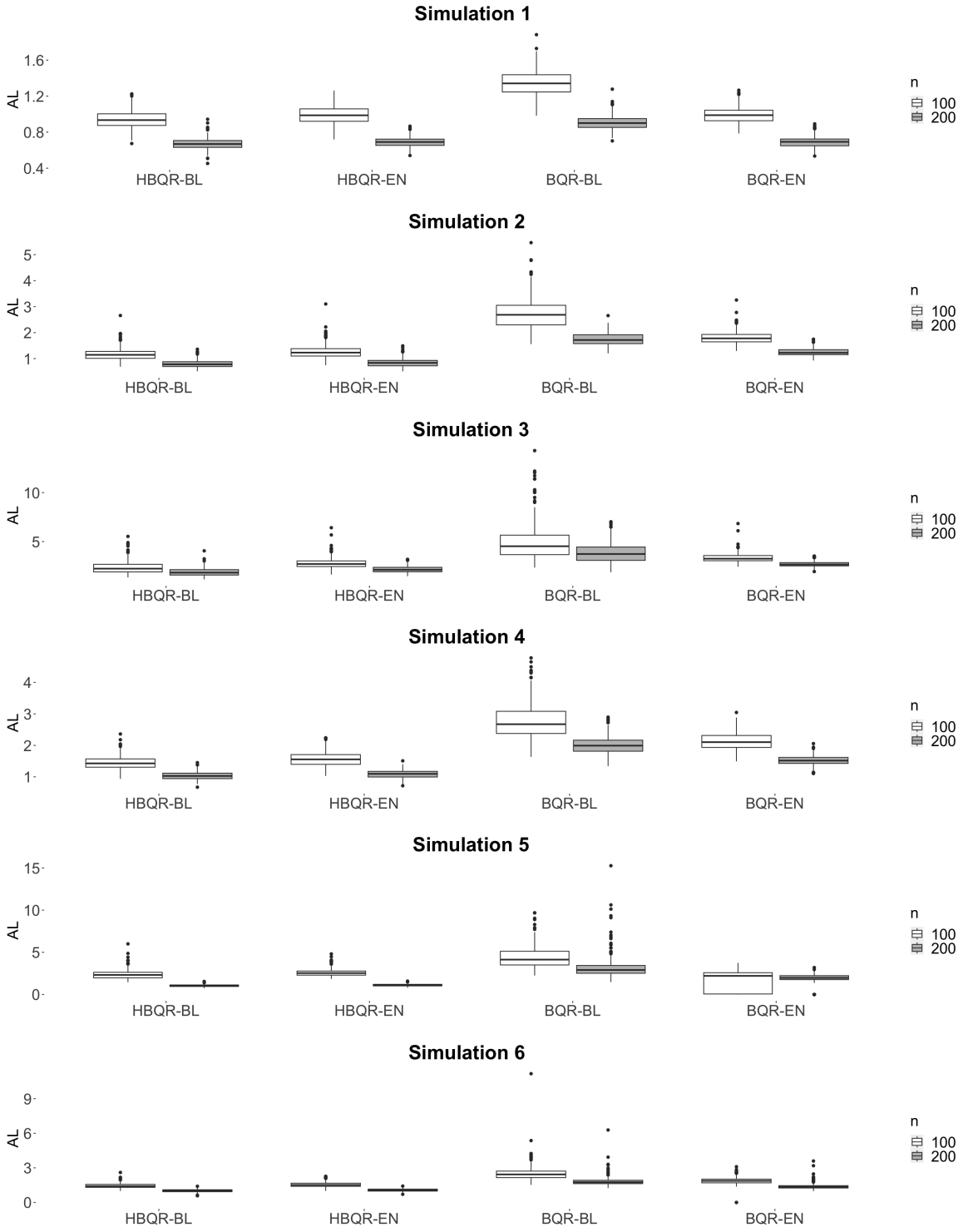
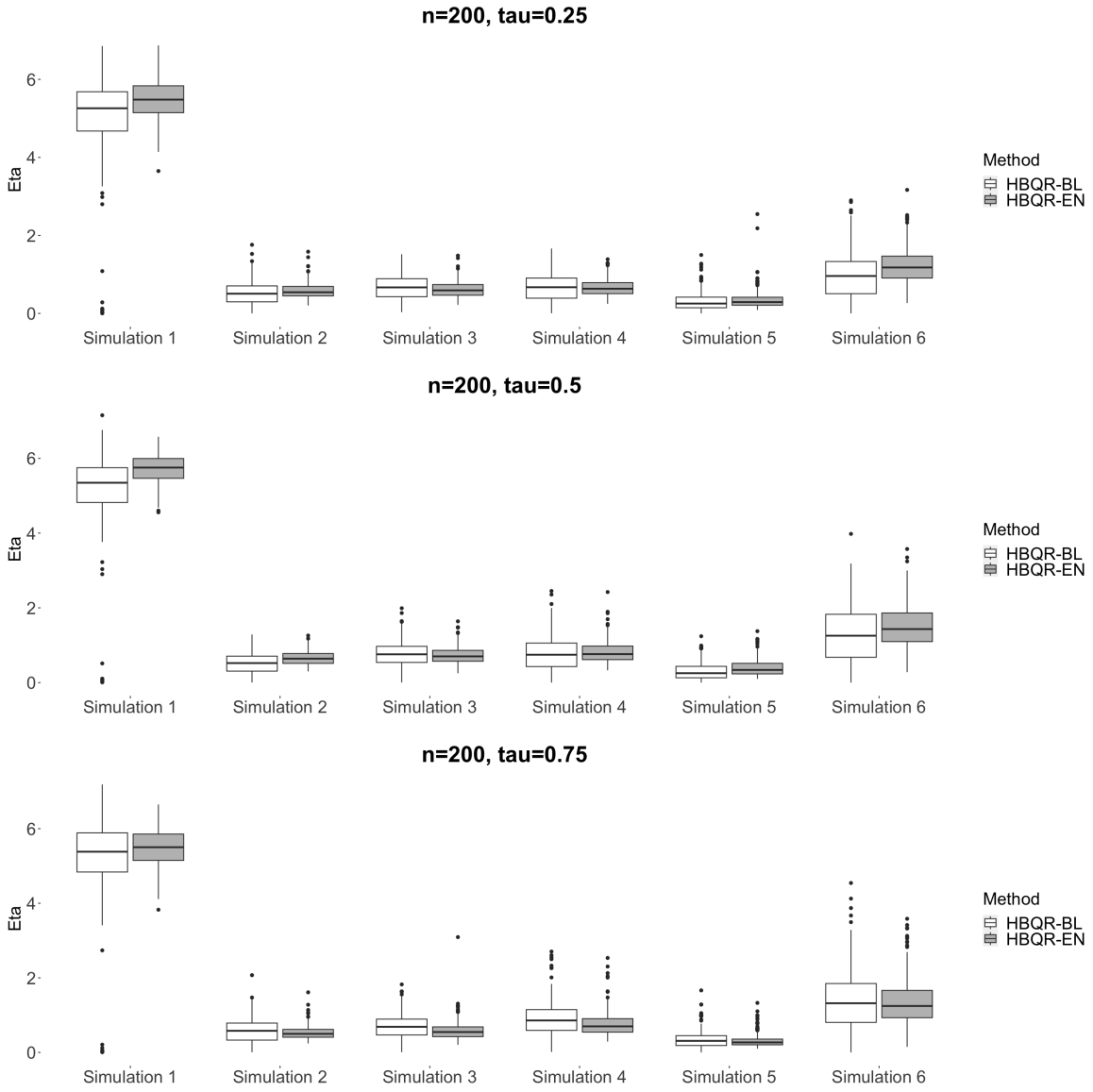


Figure D.12: Boxplots of posterior median of η based on 300 replications in six simulation scenarios for HBQR-BL and HBQR-EN ($n = 200$).



D.3 Chapter 4 Simulation Studies

Figure D.13: Boxplots of RMSE & MMAD based on 50 replications in parameter estimation simulation for the proposed VB algorithms and MCMC method ($\tau = 0.25$).

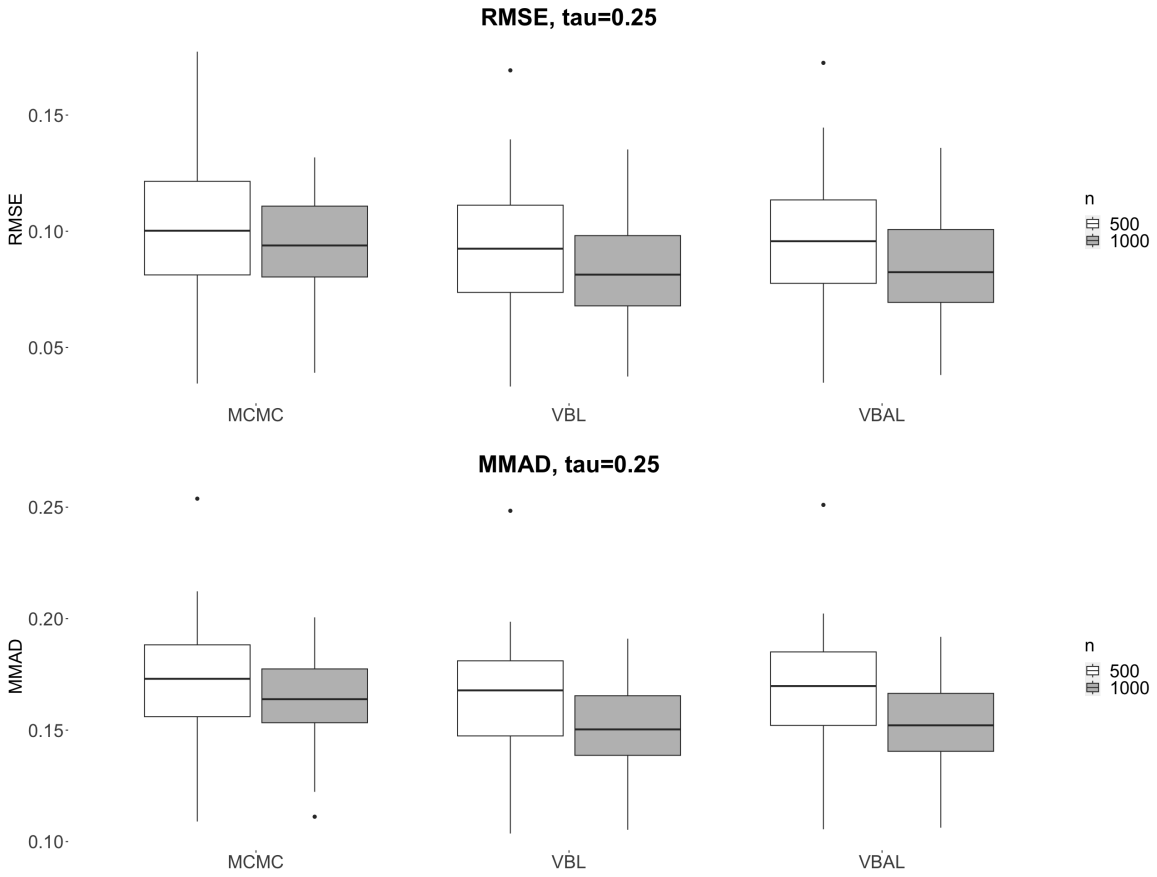


Figure D.14: Boxplots of RMSE & MMAD based on 50 replications in parameter estimation simulation for the proposed VB algorithms and MCMC method ($\tau = 0.75$).

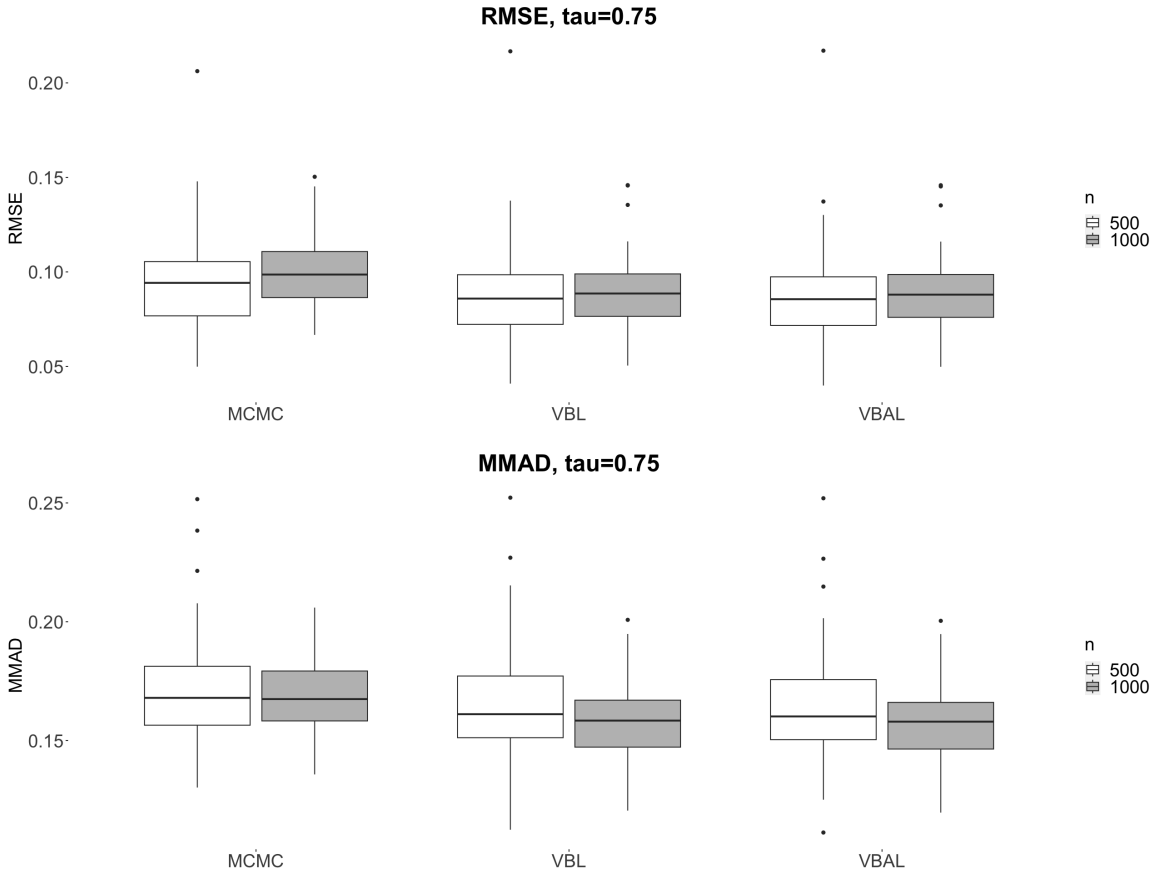


Figure D.15: Boxplots of RMSE based on 50 replications in Simulation 1 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

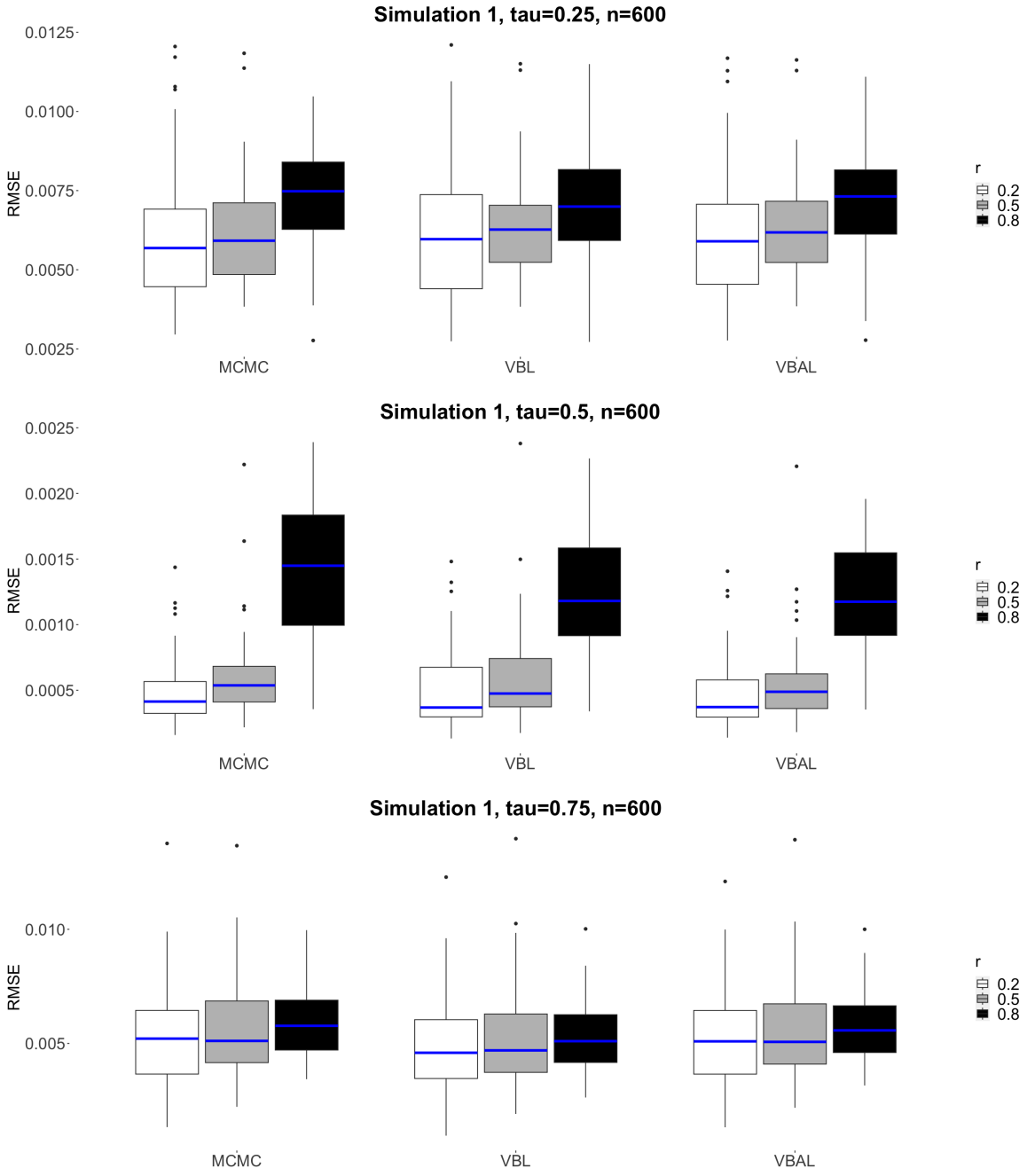


Figure D.16: Boxplots of RMSE based on 50 replications in Simulation 2 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

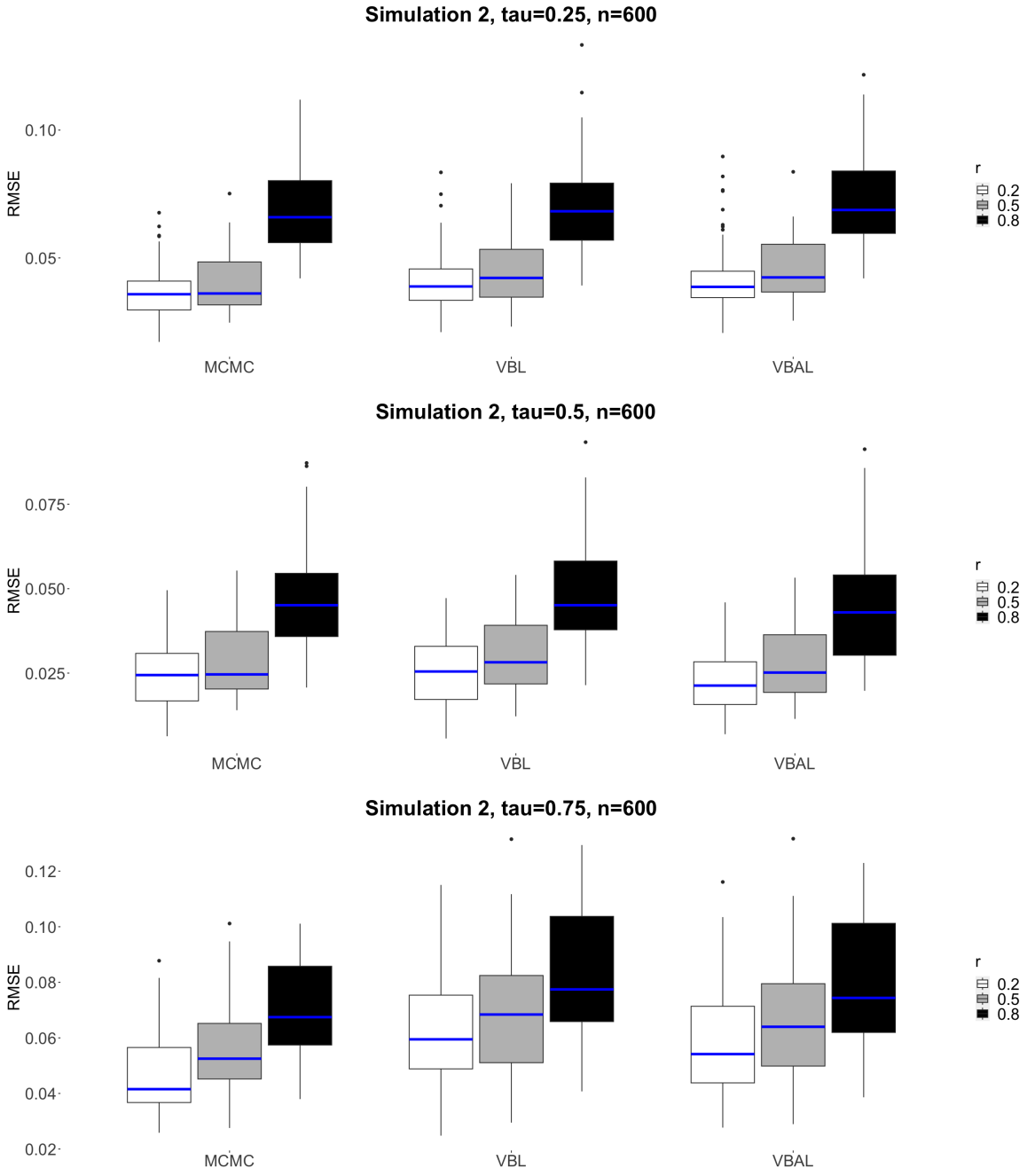


Figure D.17: Boxplots of RMSE based on 50 replications in Simulation 3 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

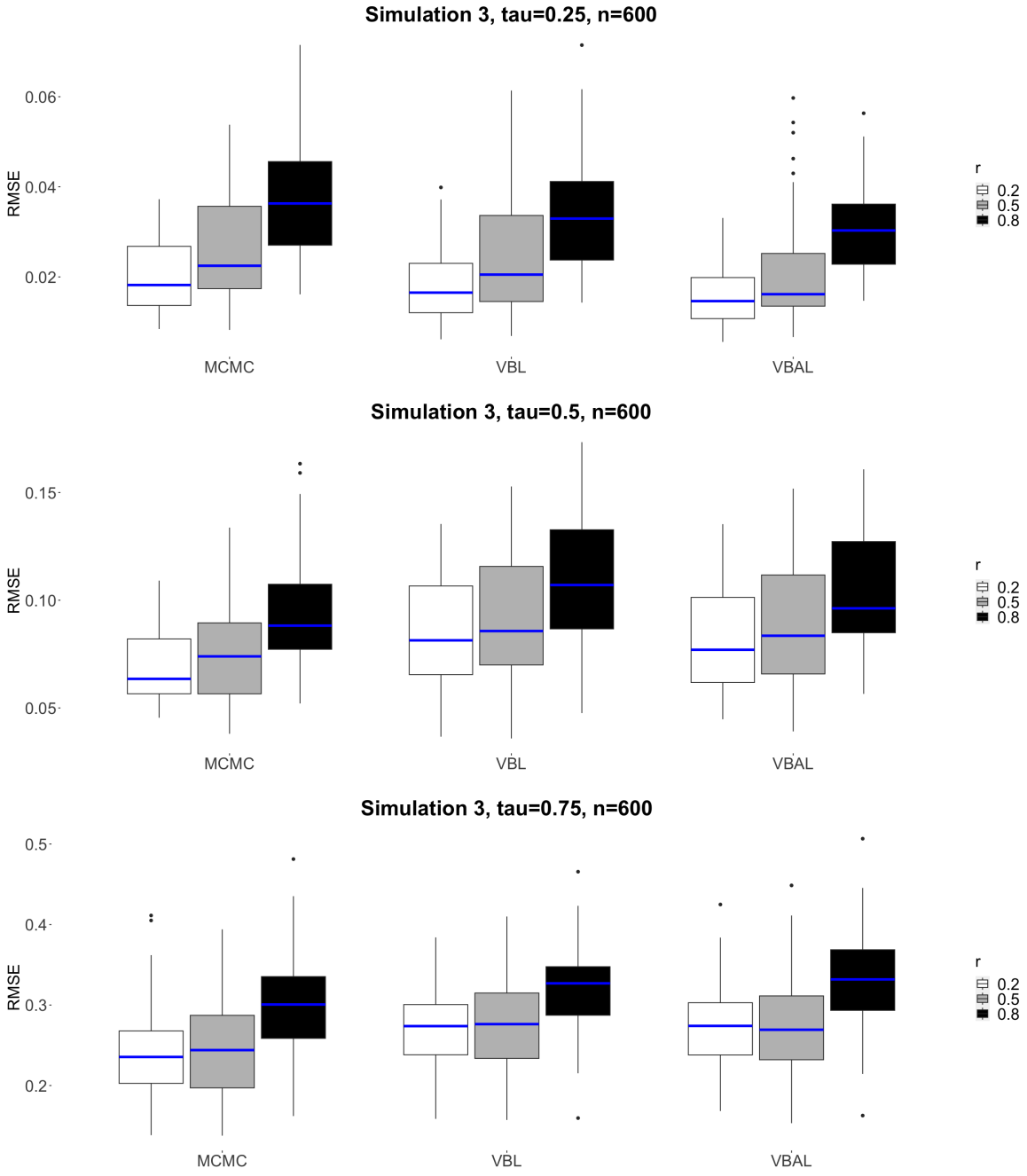


Figure D.18: Boxplots of MMAD based on 50 replications in Simulation 1 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

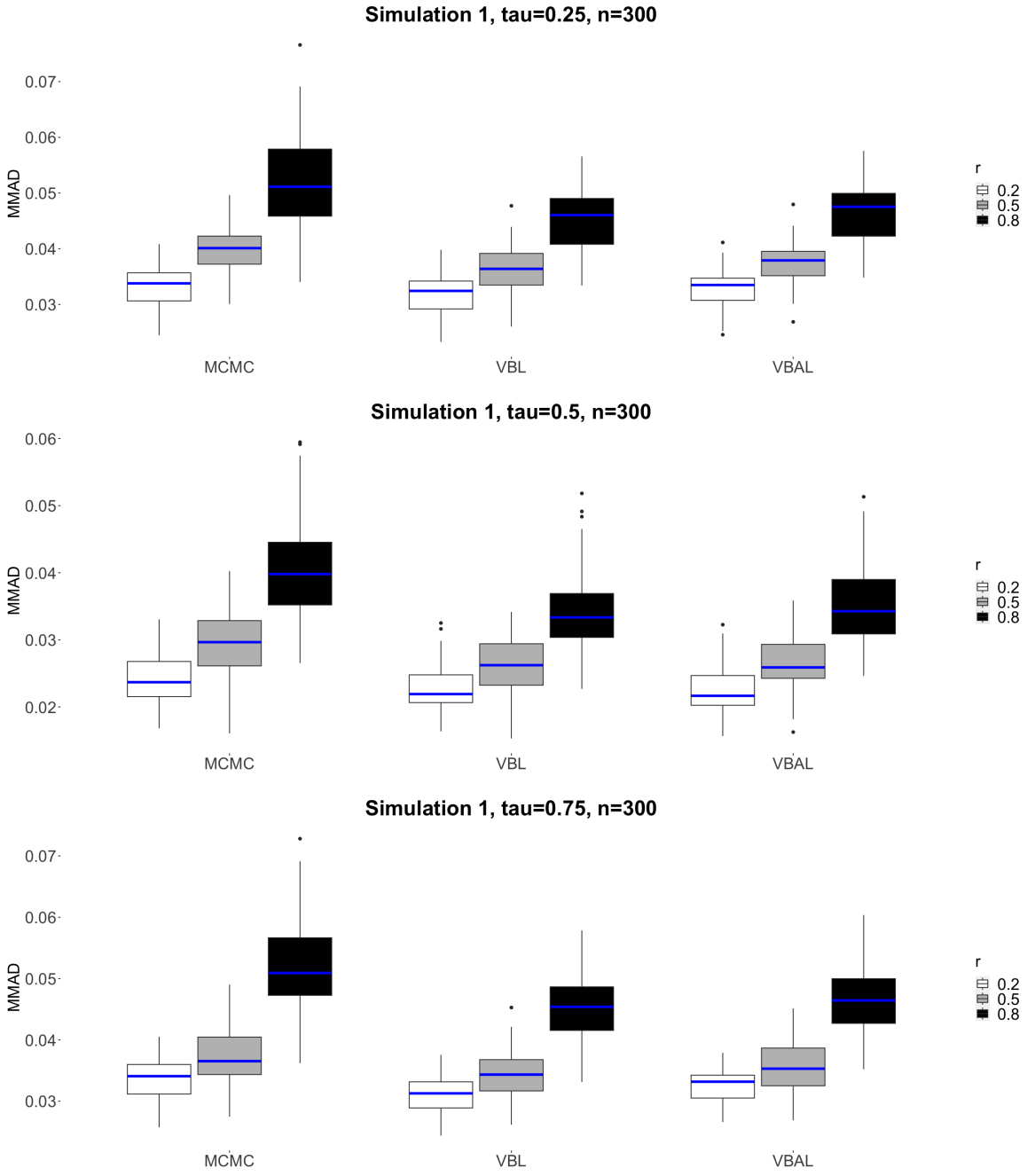


Figure D.19: Boxplots of MMAD based on 50 replications in Simulation 2 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

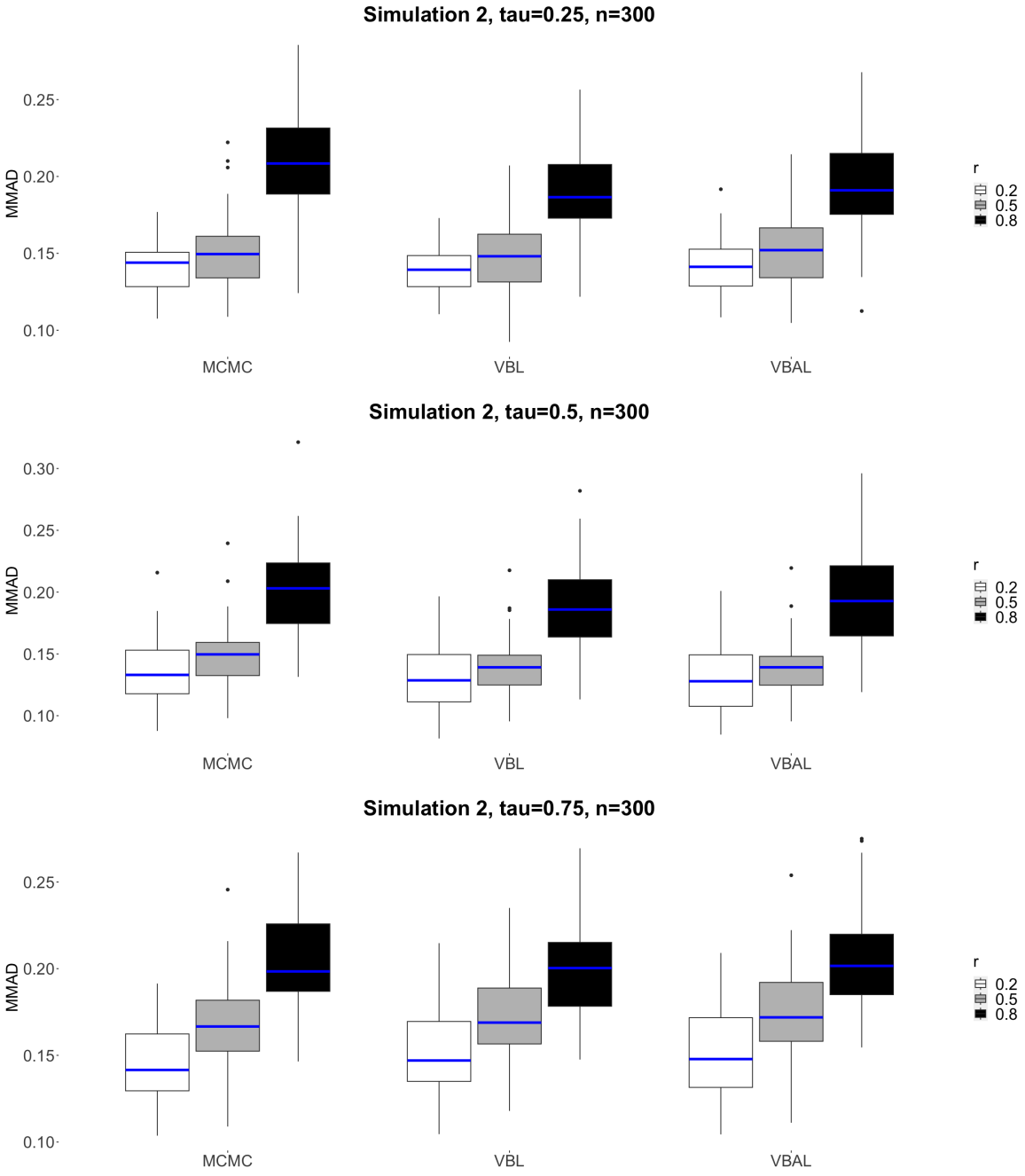


Figure D.20: Boxplots of MMAD based on 50 replications in Simulation 3 for the proposed VB algorithms and MCMC method for sample size of $n = 300$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

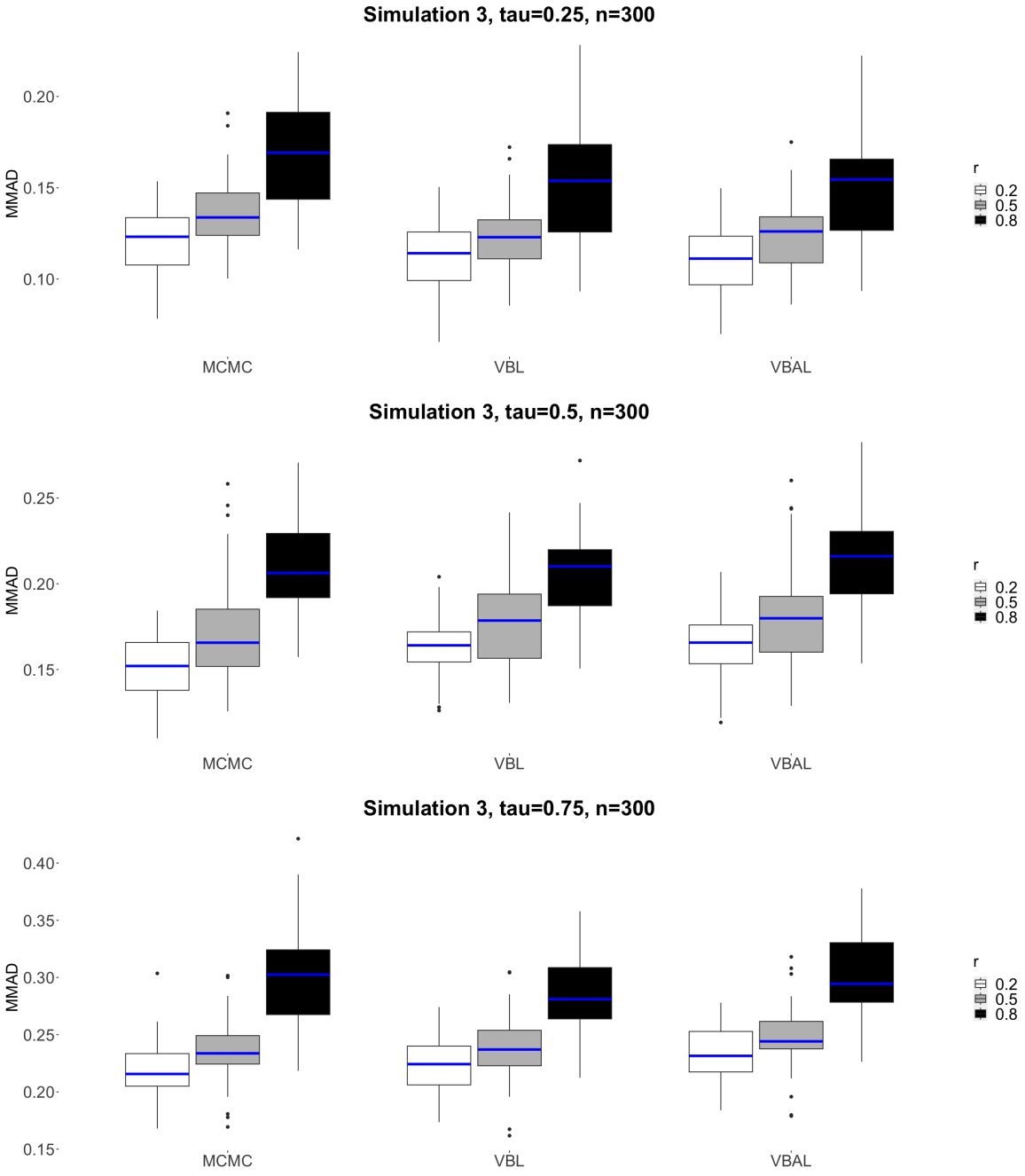


Figure D.21: Boxplots of MMAD based on 50 replications in Simulation 1 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

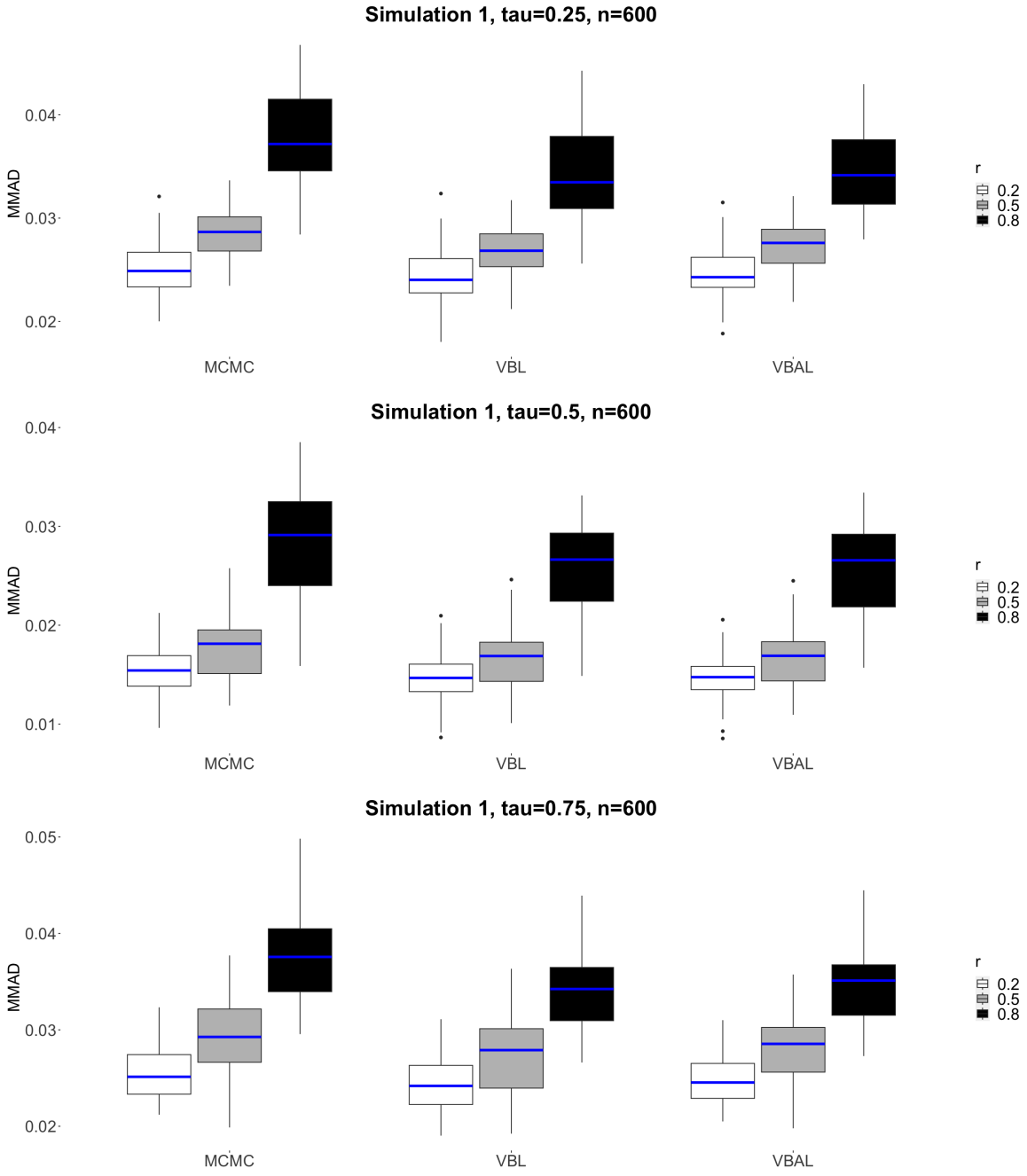


Figure D.22: Boxplots of MMAD based on 50 replications in Simulation 2 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

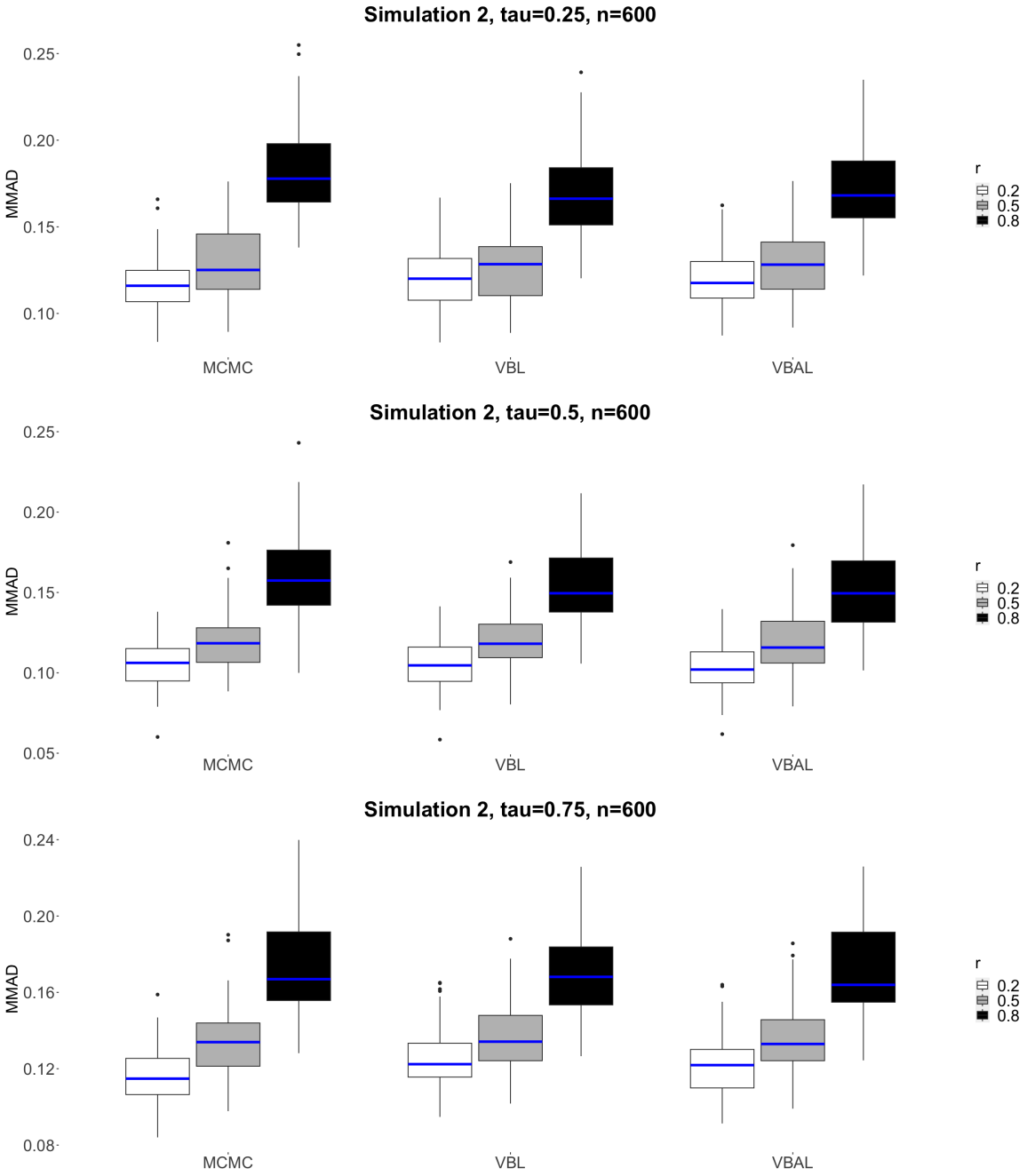


Figure D.23: Boxplots of MMAD based on 50 replications in Simulation 3 for the proposed VB algorithms and MCMC method for sample size of $n = 600$ at different quantile levels ($\tau = 0.25, 0.5, 0.75$).

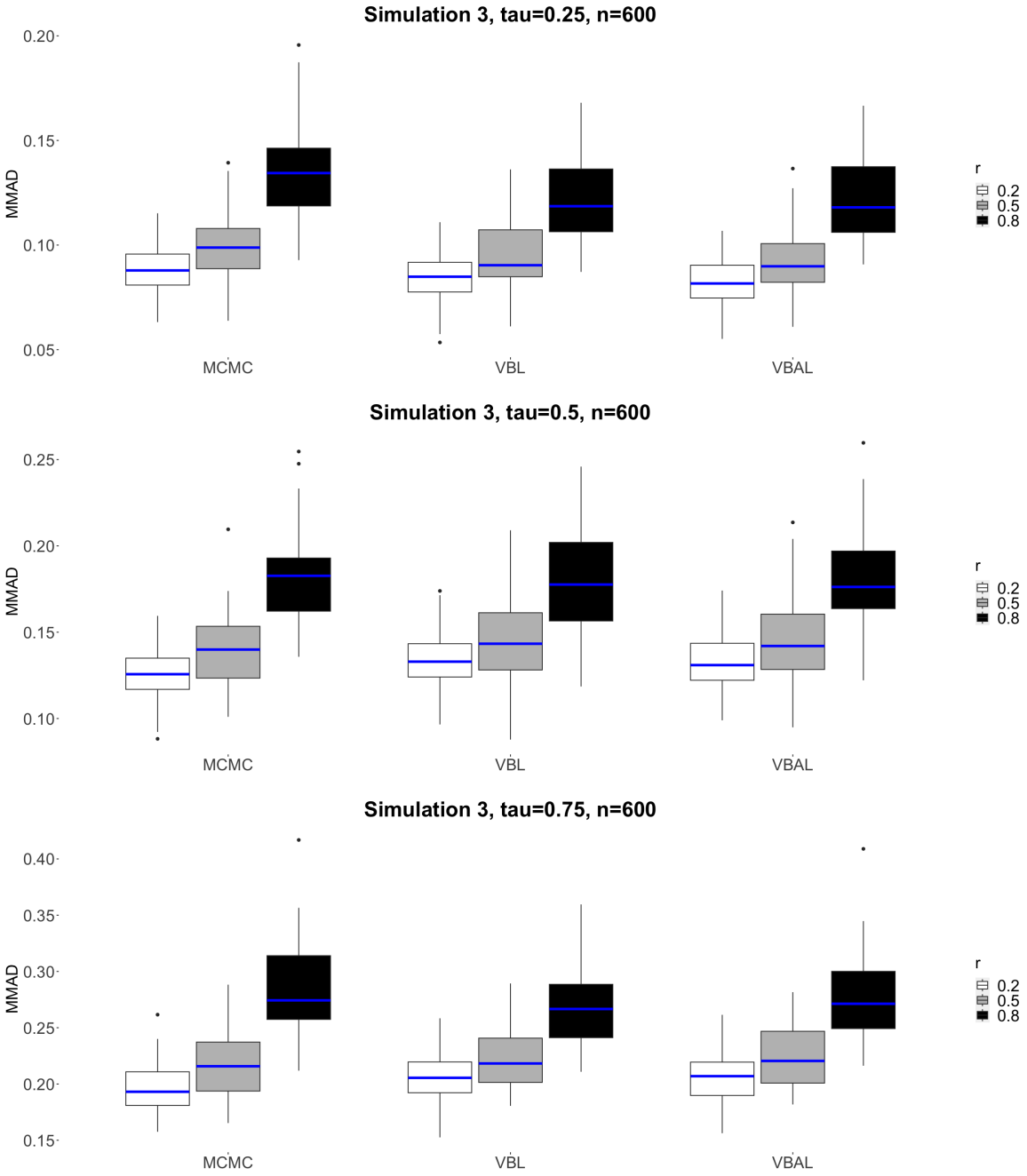


Figure D.24: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 300$ ($\tau = 0.25$).

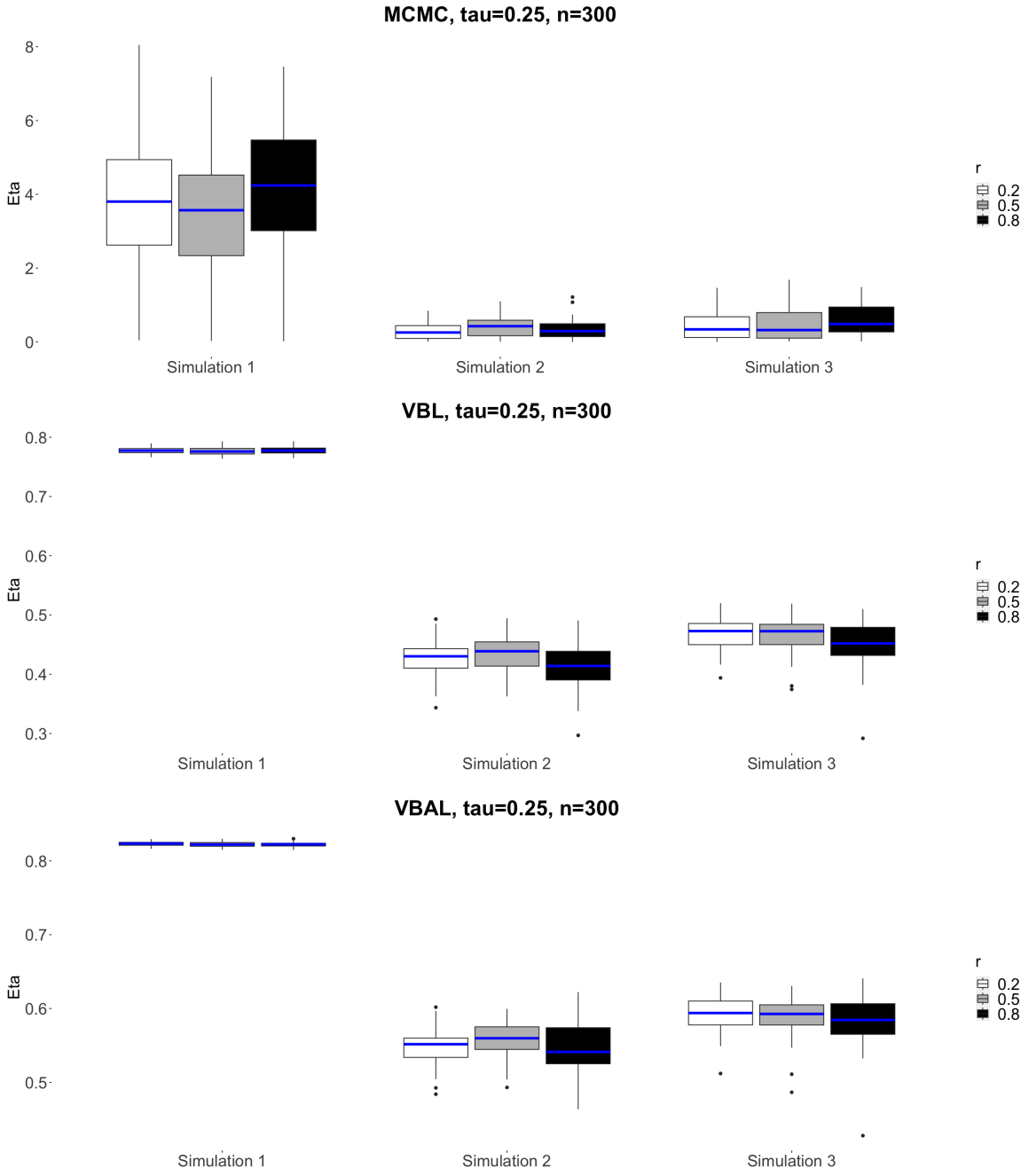


Figure D.25: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 600$ ($\tau = 0.25$).

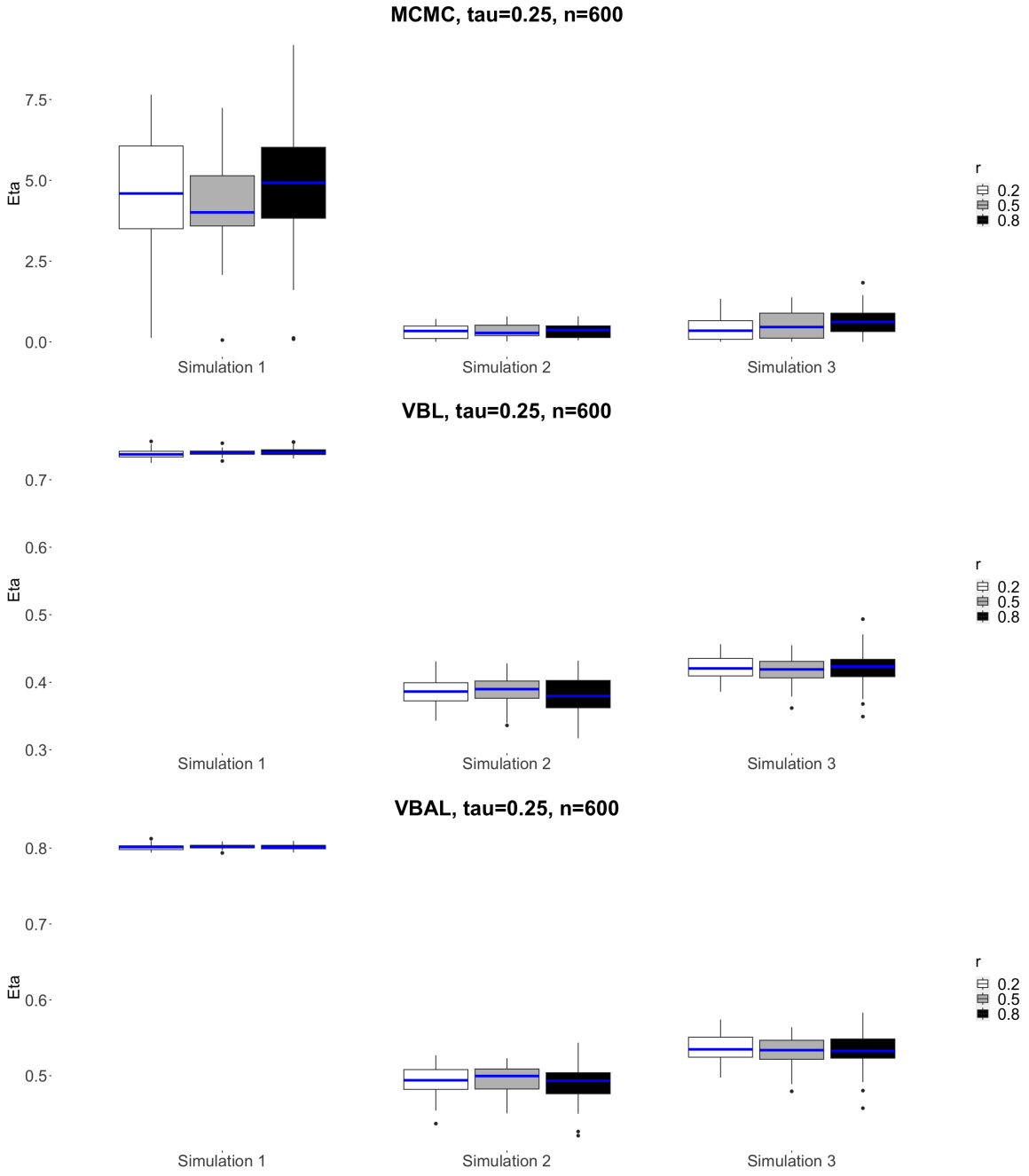


Figure D.26: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 600$ ($\tau = 0.5$).

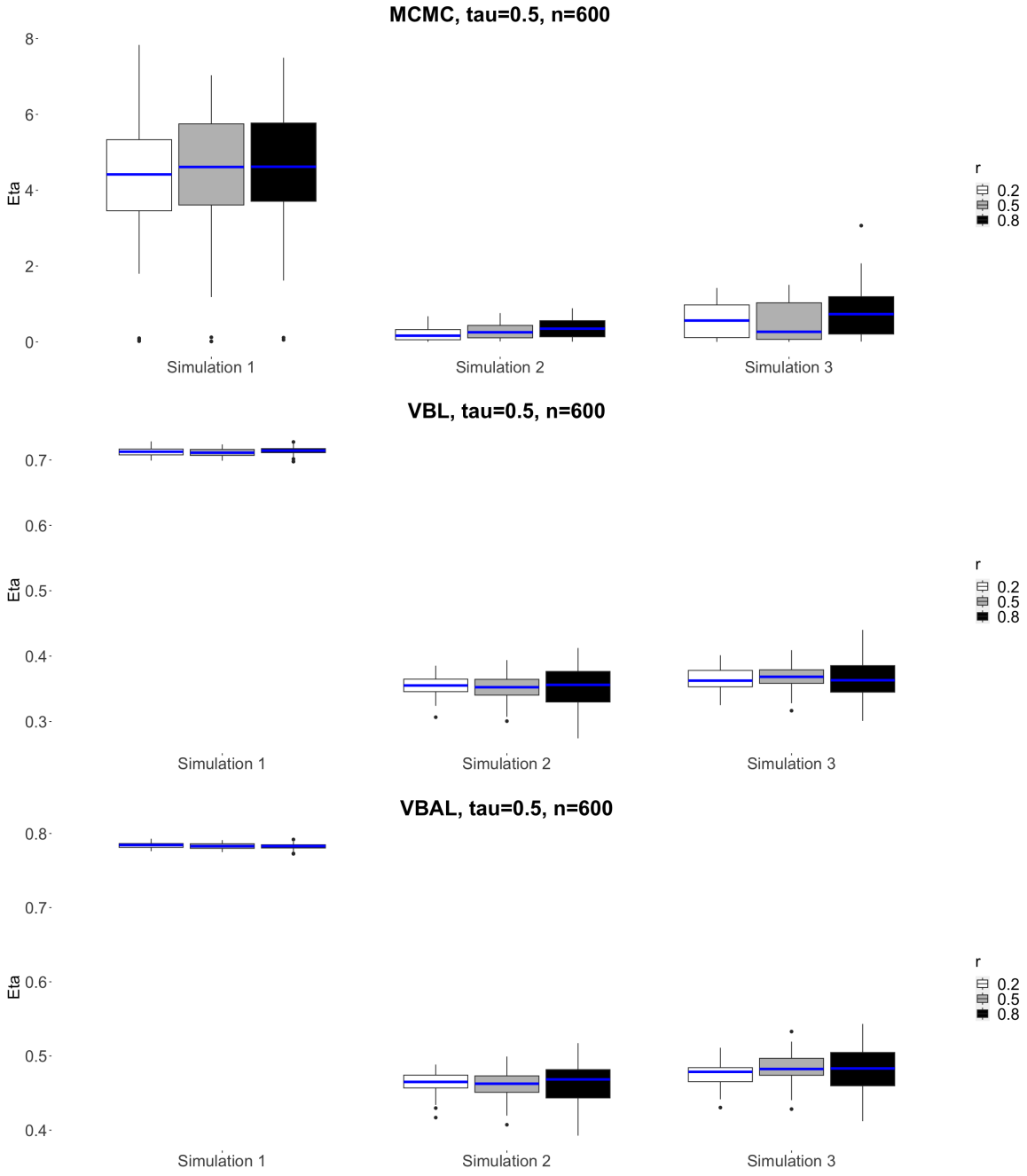


Figure D.27: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 300$ ($\tau = 0.75$).

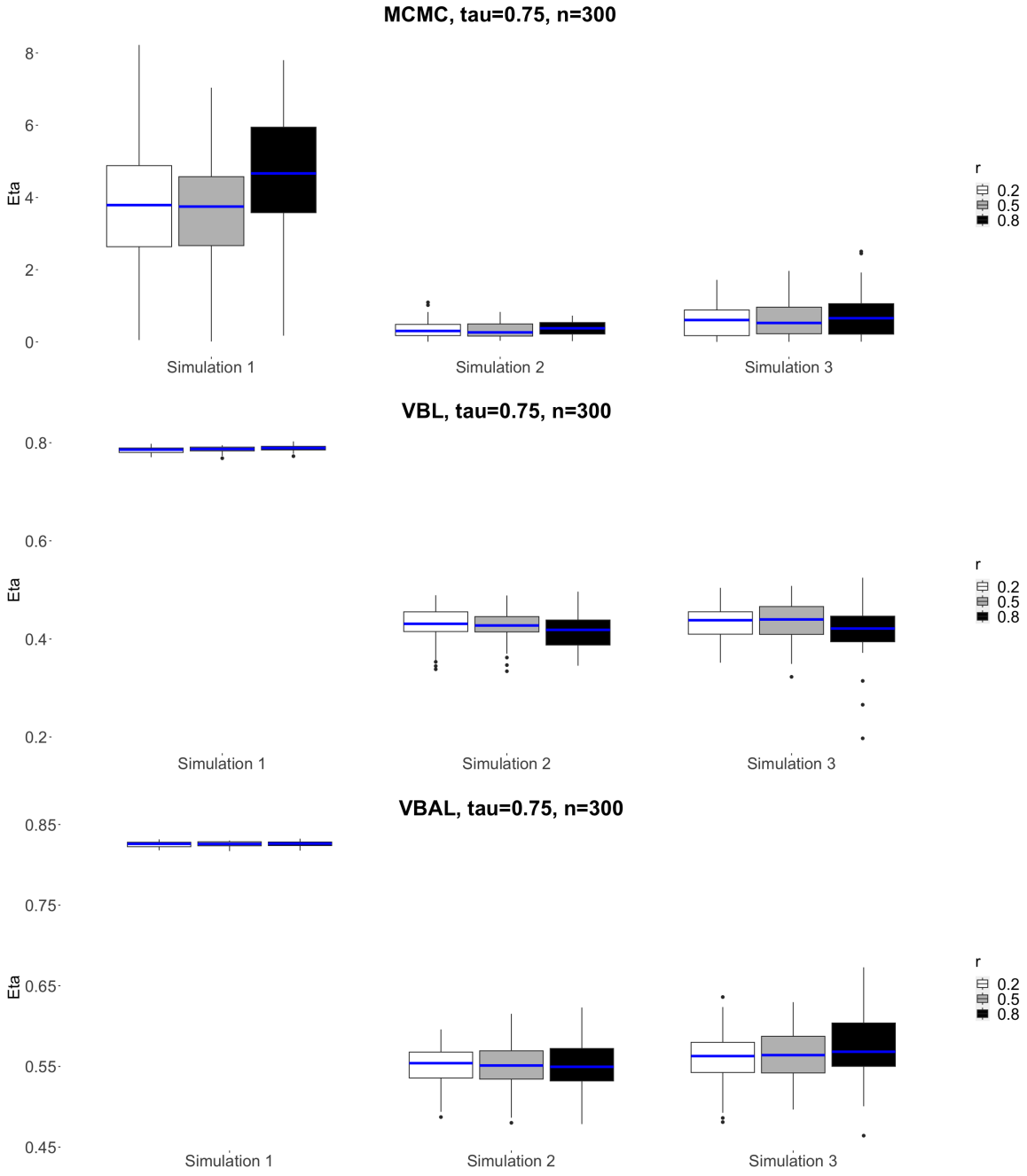
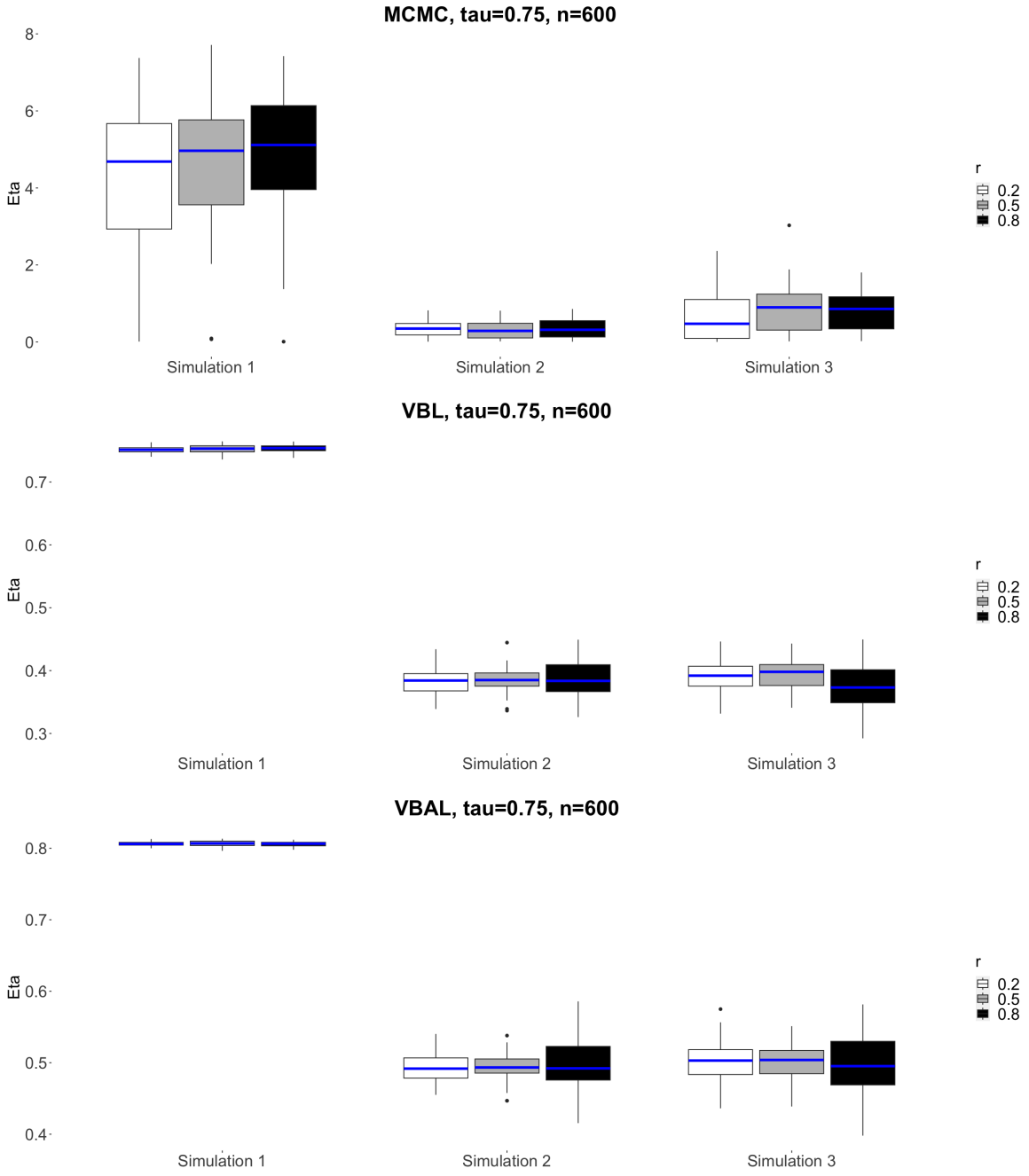


Figure D.28: Posterior median of η from the MCMC method and optimal estimate of η from the proposed VB algorithms for sample size $n = 600$ ($\tau = 0.75$).



Bibliography

Abramowitz, M. and Stegun, I. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, U.S. Government Printing Office, Volume 55 of Applied Mathematics Series.

Adlouni, S. E., Salaou, G. and St-Hilaire, A. (2018), *Regularized Bayesian quantile regression*, Communications in Statistics - Simulation and Computation, **47**(1), 277–293. https://doi.org/10.1080/03610918.2017.1280830.

Agresti, A. and Finlay, B. (1997), *Statistical Methods for Social Sciences*, **3rd ed.**, Prentice Hall.

Ahdesmäki, M., Lähdesmäki, H., Gracey, A. and Yli-Harja, O. (2007), *Robust regression for periodicity detection in non-uniformly sampled time-course gene expression data*, BMC Bioinformatics, **8**(233), 1–16. https://doi.org/10.1186/1471-2105-8-233.

Alfons, A. (2021), *robustHD: An R package for robust regression with high-dimensional data*, Journal of Open Source Software, **6**(67). Article no. 3786. https://doi.org/10.21105/joss.03786.

Alfons, A., Croux, C. and Gelper., S. (2016), *Robust groupwise least angle regression*, Computational Statistics & Data Analysis, **93**, 421–435. https://doi.org/10.1016/j.csda.2015.02.007.

Algamal, Z. Y., Alhamzawi, R. and Ali, H. T. M. (2018), *Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression*, Computers in Biology and Medicine, **97**, 145–152. https://doi.org/10.1016/j.compbiomed.2018.04.018.

Alhamzawi, R. (2016), *Bayesian Elastic Net Tobit Quantile Regression*, Communications in Statistics-Simulation and Computation, **45**(7), 2409–2427. https://doi.org/10.1080/03610918.2014.904341.

Alhamzawi, R., Alhamzawi, A. and Ali, H. T. M. (2019), *New Gibbs sampling methods for Bayesian regularized quantile regression*, Computers in Biology and Medicine, **110**, 52–65. https://doi.org/10.1016/j.compbiomed.2019.05.011.

Alhamzawi, R. and Yu, K. (2012), *Variable selection in quantile regression via Gibbs sampling*, Journal of Applied Statistics, **39**(4), 799–813. https://doi.org/10.1080/02664763.2011.620082.

Alhamzawi, R. and Yu, K. (2013), *Conjugate priors and variable selection for Bayesian quantile regression*, Computational Statistics & Data Analysis, **64**, 209–219. https://doi.org/10.1016/j.csda.2012.01.014.

Alhamzawi, R., Yu, K. and Benoit, D. F. (2012), *Bayesian adaptive Lasso quantile regression*, Statistical Modelling, **12**(3), 279–297. https://doi.org/10.1177/1471082X1101200304.

Alshaybawee, T., Midi, H. and Alhamzawi, R. (2017), *Bayesian Elastic Net single index quantile regression*, Journal of Applied Statistics, **44**(5), 853–871. https://doi.org/10.1080/02664763.2016.1189515.

Alves, L., Dias, R. and Migon, H. S. (2021), *Variational full Bayes Lasso: Knots selection in regression splines*, arXiv preprint arXiv:2102.13548. https://doi.org/10.48550/arXiv.2102.13548.

Andrews, D. F. and Mallows, C. L. (1974), *Scale Mixtures of Normal Distributions*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **36**(1), 99–102. https://doi.org/10.1111/j.2517-6161.1974.tb00989.x.

Barndorff-Nielsen, O. E. and Shephard, N. (2001), *Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **63**(2), 167–241. https://doi.org/10.1111/1467-9868.00282.

Beal, M. J. (2003), *Variational algorithms for approximate Bayesian inference*, Ph.D. thesis, University of London, University College London, UK.

Beal, M. J. and Ghahramani, Z. (2003), *The Variational Bayesian EM Algorithm for Incomplete Data: With Application to Scoring Graphical Model Structures*, Bayesian Statistics, **7**, 453–463. https://doi.org/10.1093/oso/9780198526155.003.0025.

Beal, M. J. and Ghahramani, Z. (2004), *Variational Bayesian learning of directed graphical models with hidden variables*, Bayesian Analysis, **1**(1), 1–44. https://doi.org/10.1214/06-BA126.

Bedogni, G., Giannone, G., Maghnie, M., Giacomozzi, C., Di Iorgi, N., Pedicelli, S., Peschiaroli, E., Melioli, G., Muraca, M., Cappa, M. and Cianfarani, S. (2012), *Serum insulin-like growth factor-I (IGF-I) reference ranges for chemiluminescence assay in childhood and adolescence. Data from a population of in-and out-patients*, Growth Hormone & IGF Research, **22**(3-4), 134–138. https://doi.org/10.1016/j.ghir.2012.04.005.

Bell, M. L., van Roode, T., Dickson, N. P., Jiang, Z. J. and Paul, C. (2010), *Consistency and reliability of self-reported lifetime number of heterosexual partners by gender and age in a cohort study*, Sexually transmitted diseases, **37**(7), 425–431. https://doi.org/10.1097/OLQ.0b013e3181d13ed8.

Belloni, A. and Chernozhukov, V. (2011), *ℓ_1-penalized quantile regression in high-dimensional sparse models*, Annals of Statistics, **39**(1), 82–130. https://doi.org/10.1214/10-AOS827.

Benoit, D. F. and Van den Poel, D. (2017), *bayesQR: A Bayesian approach to quantile regression*, Journal of Statistical Software, **76**(7), 1–32. https://doi.org/10.18637/jss.v076.i07.

Berger, J. O. (1984), *The robust Bayesian viewpoint (with discussion)*, In Robustness of Bayesian Analysis (J. Kadane, ed.), Amsterdam: North-Holland.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, **2nd ed.**, Springer-Verlag Inc.

Bishop, C. M. and Nasrabadi, N. M. (2006), *Pattern Recognition and Machine Learning*, Springer New York, New York, USA.

Bivand, R., Novosad, J., Lovelace, R., Monmonier, M. and Snow, G. (2021), *Package "spData"*. Available at: `https://cloud.r-project.org/web/packages/spData/spData.pdf` (Accessed: August 22, 2023).

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017), *Variational Inference: A Review for Statisticians*, Journal of the American Statistical Association, **112**(518), 859–877. `https://doi.org/10.1080/01621459.2017.1285773`.

Box, G. and Tidwell, P. (1962), *Transformation of the Independent Variables*, Technometrics, **4**(4), 531–550. `https://doi.org/10.1080/00401706.1962.10490038`.

Brent, R. (1973), *Algorithms for Minimization without Derivatives*, Englewood Cliffs N.J.: Prentice-Hall.

Bundy, J. D., Li, C., Stuchlik, P., Bu, X., Kelly, T. N., Mills, K. T., He, H., Chen, J., Whelton, P. K. and He, J. (2017), *Systolic Blood Pressure Reduction and Risk of Cardiovascular Disease and Mortality: A Systematic Review and Network Meta-analysis*, Journal of American Medical Association Cardiology, **2**(7), 775–781. `https://doi.org/10.1001/jamacardio.2017.1421`.

Cai, T., Karlaftis, V., Hearps, S., Matthews, S., Burgess, J., Monagle, P., Ignjatovic, V. and HAPPI Kids study team (2020), *Reference intervals for serum cystatin C in neonates and children 30 days to 18 years old*, Pediatric Nephrology, **35**, 1959–1966. `https://doi.org/10.1007/s00467-020-04612-5`.

Cai, Y., Stander, J. and Davies, N. (2012), *A new Bayesian approach to quantile autoregressive time series model estimation and forecasting*, Journal of Time Series Analysis, **33**(4), 684–698. `https://doi.org/10.1111/j.1467-9892.2012.00800.x`.

Cai, Z. and Sun, D. (2021), *Prior conditioned on scale parameter for Bayesian quantile LASSO and its generalizations*, Statistics and Its Interface, **14**(4), 459–474. `https://dx.doi.org/10.4310/21-SII666`.

Casati, D., Pellegrino, M., Cortinovis, I., Spada, E., Lanna, M., Faiola, S., Cetin, I. and Rustico, M. A. (2019), *Longitudinal Doppler references for monochorionic twins and comparison with singletons*, PLoS One, **14**(12). e0226090. `https://doi.org/10.1371/journal.pone.0226090`.

Casella, G. (1985), *An Introduction to Empirical Bayes Data Analysis*, The American Statistician, **39**(2), 83–87. `https://doi.org/10.1080/00031305.1985.10479400`.

Casella, G. (2001), *Empirical Bayes Gibbs Sampling*, Biostatistics, **2**(4), 485–500. `https://doi.org/10.1093/biostatistics/2.4.485`.

Casella, G. and George, E. I. (1992), *Explaining the Gibbs sampler*, The American Statistician, **46**(3), 167–174. `https://doi.org/10.1080/00031305.1992.10475878`.

Centers for Disease Control and Prevention (2022), *Defining Adult Overweight & Obesity*, Centers for Disease Control and Prevention. Available at: `https://www.cdc.gov/obesity/basics/adult-defining.html` (Accessed: March 17, 2023).

Charbonnier, P., Blanc-Féraud, L., Aubert, G. and Barlaud, M. (1997), *Deterministic edge-preserving regularization in computed imaging*, IEEE Transactions on Image Processing, **6**(2), 298–311. `https://doi.org/10.1109/83.551699`.

Chen, C. W., Dunson, D. B., Reed, C. and Yu, K. (2013), *Bayesian variable selection in quantile regression*, Statistics and its Interface, **6**(2), 261–274. https://dx.doi.org/10.4310/SII.2013.v6.n2.a9.

Chen, Y. C., Wang, Y. S. and Erosheva, E. A. (2018), *On the Use of Bootstrap with Variational Inference: Theory, Interpretation, and A Two-sample Test Example*, The Annals of Applied Statistics, **12**(2), 846–876. https://dx.doi.org/10.1214/18-AOAS1169.

Chen, Y., Yang, J., Wang, C. and Park, D. (2016), *Variational Bayesian extreme learning machine*, Neural Computing and Applications, **27**, 185–196. https://dx.doi.org/10.1007/s00521-014-1710-1.

Chitty, L. S. and Altman, D. G. (2003), *Charts of fetal size: kidney and renal pelvis measurements*, Prenatal Diagnosis: Published in Affiliation With the International Society for Prenatal Diagnosis, **23**(11), 891–897. https://doi.org/10.1002/pd.693.

Clark, D., Colantonio, L., Min, Y., Hall, M., Zhao, H., Mentz, R., Shimbo, D., Ogedegbe, G., Howard, G., Levitan, E. and Jones, D. (2019), *Population-Attributable Risk for Cardiovascular Disease Associated With Hypertension in Black Adults*, Journal of American Medical Association Cardiology, **4**(12), 1194–1202. https://doi.org/10.1001/jamacardio.2019.3773.

da Silva Ferreira, C., Bolfarine, H. and Lachos, V. H. (2011), *Skew scale mixtures of normal distributions: Properties and estimation*, Statistical Methodology, **8**(2), 154–171. https://doi.org/10.1016/j.stamet.2010.09.001.

Dao, M., Wang, M., Ghosh, S. and Ye, K. (2022), *Bayesian variable selection and estimation in quantile regression using a quantile-specific prior*, Computational Statistics, **37**(3), 1339–1368. https://doi.org/10.1007/s00180-021-01181-5.

Daouia, A., Gijbels, I. and Stupfler, G. (2019), *Extremiles: A New Perspective on Asymmetric Least Squares*, Journal of the American Statistical Association, **114**(527), 1366–1381. https://doi.org/10.1080/01621459.2018.1498348.

Daouia, A., Girard, S. and Stupfler, G. (2018), *Estimation of Tail Risk Based on Extreme Expectiles*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **80**(2), 263–292. https://doi.org/10.1111/rssb.12254.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), *Maximum Likelihood from Incomplete Data Via the EM Algorithm*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **39**(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Dong, B., Wang, Z., Arnold, L., Song, Y., Wang, H. J. and Ma, J. (2016), *The association between blood pressure and grip strength in adolescents: does body mass index matter?*, Hypertension Research, **39**(12), 919–925. https://doi.org/10.1038/hr.2016.84.

Ehm, W., Gneiting, T., Jordan, A. and Krüger, F. (2016), *Of Quantiles and Expectiles: Consistent Scoring Functions, Choquet Representations and Forecast Rankings*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **78**(3), 505–562. https://doi.org/10.1111/rssb.12154.

Ettehad, D., Emdin, C. A., Kiran, A., Anderson, S. G., Callender, T., Emberson, J., Chalmers, J., Rodgers, A. and Rahimi, K. (2016), *Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis*, The Lancet, **387**(10022), 957–967. https://doi.org/10.1016/S0140-6736(15)01225-8.

Feng, X. N., Wang, G. C., Wang, Y. F. and Song, X. Y. (2015), *Structure detection of semiparametric structural equation models with Bayesian adaptive group Lasso*, Statistics in Medicine, **34**(9), 1527–1547. https://doi.org/10.1002/sim.6410.

Feng, X. N., Wang, Y., Lu, B. and Song, X. Y. (2017), *Bayesian regularized quantile structural equation models*, Journal of Multivariate Analysis, **154**, 234–248. https://doi.org/10.1016/j.jmva.2016.11.002.

Fox, C. W. and Roberts, S. J. (2012), *A tutorial on variational Bayesian inference*, Artificial Intelligence Review, **38**, 85–95. https://doi.org/10.1007/s10462-011-9236-8.

Frangou, S., Modabbernia, A., Williams, S., Papachristou, E., Doucet, G., Agartz, I., Aghajani, M., Akudjedu, T., Albajes-Eizagirre, A., Alnæs, D. and Alpert, K. (2021), *Cortical thickness across the lifespan: Data from 17,075 healthy individuals aged 3–90 years*, Human Brain Mapping, **3**(1), 431-451. https://doi.org/10.1002/hbm.25364.

Fu, L. and Wang, Y.-G. (2021), *Robust regression with asymmetric loss functions*, Statistical Methods in Medical Research, **30**(8), 1800–1815. https://doi.org/10.1177/09622802211012012.

Garnatz, C. and Hardin, J. (2015), *Trusting the Black Box: Confidence with Bag of Little Bootstraps*, Senior Thesis in Mathematics, Pomona College, California, USA.

Gelman, A. and Vehtari, A. (2021), *What are the Most Important Statistical Ideas of the Past 50 Years?*, Journal of the American Statistical Association, **116**(536), 2087–2097. https://doi.org/10.1080/01621459.2021.1938081.

Geman, S. and Geman, D. (1984), *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence,, **PAMI-6**(6), 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Geraci, M. and Bottai, M. (2007), *Quantile regression for longitudinal data using the asymmetric Laplace distribution*, Biostatistics, **8**(1), 140–154. https://doi.org/10.1093/biostatistics/kxj039.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. (1995), *Markov Chain Monte Carlo in Practice*, Chapman and Hall/CRC, New York, USA.

Green, P. J. (1995), *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination*, Biometrika, **82**(4), 711–732. https://doi.org/10.1093/biomet/82.4.711.

Griffin, J. and Brown, P. (2007), *Bayesian adaptive Lassos with non-convex penalization*, Technical Report, University of Kent, Kent, UK. http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2007/paper07-2/.

Guo, R., Zhu, H., Chow, S. M. and Ibrahim, J. G. (2012), *Bayesian Lasso for Semiparametric Structural Equation Models*, Biometrics, **68**(2), 567–577. https://doi.org/10.1111/j.1541-0420.2012.01751.x.

Hampel, F. R. (1968), *Contribution to the Theory of Robust Estimation*, Ph.D Thesis, University of California, Berkeley, USA.

Hampel, F., Ronchetti, E., Rousseeuw, P. and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, USA.

Hamra, G., MacLehose, R. and Richardson, D. (2013), *Markov chain Monte Carlo: an introduction for epidemiologists*, International Journal of Epidemiology, **42**(2), 627–634. https://doi.org/10.1093/ije/dyt043.

Hamura, Y., Irie, K. and Sugasawa, S. (2022), *Log-regularly varying scale mixture of normals for robust regression*, Computational Statistics & Data Analysis, **173**. Article no. 107517. https://doi.org/10.1016/j.csda.2022.107517.

Harrison Jr, D. and Rubinfeld, D. L. (1978), *Hedonic housing prices and the demand for clean air*, Journal of Environmental Economics and Management, **5**(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2.

Hashimoto, S. and Sugasawa, S. (2020), *Robust Bayesian Regression with Synthetic Posterior Distributions*, Entropy, **22**(6). Article no. 661. https://doi.org/10.3390/e22060661.

Hastings, W. K. (1970), *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, Biometrika, **57**(1), 97–109. https://doi.org/10.1093/biomet/57.1.97.

Henderson, H. V. and Searle, S. R. (1981), *On Deriving the Inverse of a Sum of Matrices*, SIAM Review, **23**(1), 53–60. https://doi.org/10.1137/1023004.

Hespanhol, L., Vallio, C. S., Costa, L. M. and Saragiotto, B. T. (2019), *Understanding and interpreting confidence and credible intervals around effect estimates*, Brazilian Journal of Physical Therapy, **23**(4), 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006.

Hobert, J. P. and Casella, G. (1996), *The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models*, Journal of the American Statistical Association, **91**(436), 1461–1473. https://doi.org/10.1080/01621459.1996.10476714.

Hoti, F. and Sillanpää, M. J. (2006), *Bayesian mapping of genotype× expression interactions in quantitative and qualitative traits*, Heredity, **97**(1), 4–18. https://doi.org/10.1038/sj.hdy.6800817.

Huang, M. L., Han, Y. and Marshall, W. (2023), *An Algorithm of Nonparametric Quantile Regression*, Journal of Statistical Theory and Practice, **17**(2). Article no. 32. https://doi.org/10.1007/s42519-023-00325-8.

Huber, P. (1981), *Robust Statistics*, **1st ed.**, Wiley, New York, USA.

Huber, P. J. (1964), *Robust Estimation of a Location Parameter*, The Annals of Mathematical Statistics, **35**(1), 73–101. https://doi.org/10.1214/aoms/1177703732.

Huber, P. and Ronchetti, E. M. (2009), *Robust Statistics*, **2nd ed.**, Wiley, New York, USA.

Jensen, J. L. W. V. (1906), *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, Acta Mathematica, **30**(1), 175–193. https://doi.org/10.1007/BF02418571.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999), *An Introduction to Variational Methods for Graphical Models*, Machine Learning, **37**(2), 183–233. https://doi.org/10.1023/A:1007665907178.

Juhan, N., Zubairi, Y. Z., Mohd Khalid, Z. and Mahmood Zuhdi, A. S. (2020), *A Comparison Between Bayesian and Frequentist Approach in the Analysis of Risk Factors for Female Cardiovascular Disease Patients in Malaysia*, ASM Science Journal, **13**, 1–7. https://doi.org/10.32802/asmscj.2020.sm26(1.1).

Jullion, A. and Lambert, P. (2007), *Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models*, Computational Statistics and Data Analysis, **51**(5), 2542–2558. https://doi.org/10.1016/j.csda.2006.09.027.

Kawakami, J. and Hashimoto, S. (2023), *Approximate Gibbs sampler for Bayesian Huberized Lasso*, Journal of Statistical Computation and Simulation, **93**(1), 128–162. https://doi.org/10.1080/00949655.2022.2096886.

Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M. L. (2023), *The Big Data Bootstrap*, arXiv:1206.6415. https://doi.org/10.48550/arXiv.1206.6415.

Koenker, R. and Bassett, G. (1978), *Regression Quantiles*, Econometrica,, **46**(1), 33–50. https://doi.org/10.2307/1913643.

Koh, H. B., Heo, G. Y., Kim, K. W., Ha, J., Park, J. T., Han, S. H., Yoo, T.-H., Kang, S.-W. and Kim, H. W. (2022), *Trends in the association between body mass index and blood pressure among 19-year-old men in Korea from 2003 to 2017*, Scientific Reports, **12**(1). Article no. 6767. https://doi.org/10.1038/s41598-022-10570-9.

Kozubowski, T. J. and Podgórski, K. (2001), *Asymmetric Laplace laws and modeling financial data*, Mathematical and Computer Modelling, **34**(9-11), 1003–1021. https://doi.org/10.1016/S0895-7177(01)00114-5.

Kozumi, H. and Kobayashi, G. (2011), *Gibbs sampling methods for Bayesian quantile regression*, Journal of Statistical Computation and Simulation, **81**(11), 1565–1578. https://doi.org/10.1080/00949655.2010.496117.

Kroon, F. P., Ramiro, S., Royston, P., Le Cessie, S., Rosendaal, F. R. and Kloppenburg, M. (2017), *Reference curves for the Australian/Canadian Hand Osteoarthritis Index in the middle-aged Dutch population*, Rheumatology, **56**(5), 745–752. https://doi.org/10.1093/rheumatology/kew483.

Kuhudzai, A. G., Van Hal, G., Van Dongen, S. and Hoque, M. (2022), *Modelling of South African Hypertension: Comparative Analysis of the Classical and Bayesian Quantile Regression Approaches*, INQUIRY: The Journal of Health Care Organization, Provision and Financing, **59**, 1–9. https://doi.org/10.1177/00469580221082356.

Kullback, S. and Leibler, R. A. (1951), *On Information and Sufficiency*, The Annals of Mathematical Statistics, **22**(1), 79–86. https://www.jstor.org/stable/2236703.

Kyung, M., Gill, J., Ghosh, M. and Casella, G. (2010), *Penalized regression, standard errors, and Bayesian Lassos*, Bayesian Analysis, **5**(2), 369–412. https://doi.org/10.1214/10-BA607.

Lambert-Lacroix, S. and Zwald, L. (2011), *Robust regression through the Huber's criterion and adaptive Lasso penalty*, Eletronic Journal of Statistics, **5**, 1015–1053. https://doi.org/10.1214/11-EJS635.

Lesaffre, E. and Lawson, A. B. (2012), *Bayesian Biostatistics*, John Wiley & Sons.

Li, J., Chen, Q., Leng, J., Zhang, W. and Guo, M. (2020), *Probabilistic robust regression with adaptive weights – a case study on face recognition*, Frontiers of Computer Science, **14**(5), 1–12. https://doi.org/10.1007/s11704-019-9097-x.

Li, Q. and Lin, N. (2010), *The Bayesian Elastic Net*, Bayesian Analysis, **5**(1), 151–170. https://doi.org/10.1214/10-BA506.

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

Li, Y. and Zhu, J. (2008), *L_1-Norm Quantile Regression*, Journal of Computational and Graphical Statistics, **17**(1), 163–185. https://doi.org/10.1198/106186008X289155.

Lim, S., Vos, T., Flaxman, A., Danaei, G., Shibuya, K., Adair-Rohani, H., AlMazroa, M., Amann, M., Anderson, H., Andrews, K. and Aryee, M. (2012), *A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010*, The Lancet, **380**(9859), 2224–2260. https://doi.org/10.1016/S0140-6736(12)61766-8.

Liu, Y., Wu, M., Xu, B. and Kang, L. (2022), *Association between the urinary nickel and the diastolic blood pressure in general population*, Chemosphere, **286**(Part 1). Article no. 131900. https://doi.org/10.1016/j.chemosphere.2021.131900.

Loef, M., Kroon, F. P. B., Böhringer, S., Roos, E. M., Rosendaal, F. R. and Kloppenburg, M. (2020), *Percentile curves for the knee injury and osteoarthritis outcome score in the middle-aged Dutch population*, Osteoarthritis and Cartilage, **28**(8), 1046–1054. https://doi.org/10.1016/j.joca.2020.03.014.

MacKay, D. J. (1992), *Information-Based Objective Functions for Active Data Selection*, Neural computation, **4**(4), 590–604. https://doi.org/10.1162/neco.1992.4.4.590.

Maidman, A. and Wang, L. (2018), *New semiparametric method for predicting high-cost patients*, Biometrics, **74**(3), 1104–1111. https://doi.org/10.1111/biom.12834.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Wiley, New York, USA.

Meinshausen, N. and Ridgeway, G. (2006), *Quantile Regression Forests*, Journal of Machine Learning Research, **7**(6), 983–999. http://jmlr.org/papers/v7/meinshausen06a.html.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), *Equation of State Calculations by Fast Computing Machines*, Journal of Chemical Physics, **21**, 1087–1091. https://doi.org/10.1063/1.1699114.

Miller, J. W. (2019), *Fast and Accurate Approximation of the Full Conditional for Gamma Shape Parameters*, Journal of Computational and Graphical Statistics, **28**(2), 476–480. https://doi.org/10.1080/10618600.2018.1537929.

Minka, T. (2005), *Divergence measures and message passing*, Technical report. MSR-TR-2005-173, Microsoft Research Ltd, Cambridge, UK. https://api.semanticscholar.org/CorpusID:7585417.

Morris, C. N. (1983), *Parametric Empirical Bayes Inference: Theory and Applications*, Journal of the American Statistical Association, **78**(381), 47–55. https://doi.org/10.1080/01621459.1983.10477920.

Natarajan, R. and McCulloch, C. E. (1995), *A note on the existence of the posterior distribution for a class of mixed models for binomial responses*, Biometrika, **82**(3), 639–643. https://doi.org/10.1093/biomet/82.3.639.

Navar, A. M., Peterson, E. D., Wojdyla, D., Sanchez, R. J., Sniderman, A. D., D'Agostino, R. B. and Pencina, M. J. (2016), *Temporal Changes in the Association Between Modifiable Risk Factors and Coronary Heart Disease Incidence*, Journal of American Medical Association, **316**(19), 2041–2043. https://doi.org/10.1001/jama.2016.13614.

Ormerod, J. T. and Wand, M. P. (2010), *Explaining Variational Approximations*, The American Statistician, **64**(2), 140–153. https://doi.org/10.1198/tast.2010.09058.

Ostwald, D., Kirilina, E., Starke, L. and Blankenburg, F. (2014), *A tutorial on variational Bayes for latent linear stochastic time-series models*, Journal of Mathematical Psychology, **60**, 1–19. https://doi.org/10.1016/j.jmp.2014.04.003.

Park, T. and Casella, G. (2008), *The Bayesian Lasso*, Journal of the American Statistical Association, **103**(482), 681–686. https://doi.org/10.1198/016214508000000337.

Prospective Studies Collaboration (2002), *Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies*, The Lancet, **360**(9349), 1903–1913. https://doi.org/10.1016/S0140-6736(02)11911-8.

Quadrianto, N. and Ghahramani, Z. (2014), *A Very Simple Safe-Bayesian Random Roorest*, IEEE Transactions on Pattern Analysis and Machine Intelligence, **37**(6), 1297–1303. https://doi.org/10.1109/TPAMI.2014.2362751.

Ravaghi, V., Durkan, C., Jones, K., Girdler, R., Mair-Jenkins, J., Davies, G., Wilcox, D., Dermont, M., White, S., Dailey, Y. and Morris, A. (2020), *Area-level deprivation and oral cancer in England 2012–2016*, Cancer Epidemiology, **69**. Article no. 101840. https://doi.org/10.1016/j.canep.2020.101840.

Rea, L. M. and Parker, R. A. (2014), *Designing and Conducting Survey Research: A Comprehensive Guide*, John Wiley & Sons.

Reed, C. and Yu, K. (2009), *An Efficient Gibbs Sampler for Bayesian Quantile Regression*, Technical Report, Brunel University London, UK. http://bura.brunel.ac.uk/handle/2438/3593.

Reich, B. J. and Smith, L. B. (2013), *Bayesian Quantile Regression for Censored Data*, Biometrics, **69**(3), 651–660. https://doi.org/10.1111/biom.12053.

Ren, J., Zhou, F., Li, X., Ma, S., Jiang, Y. and Wu, C. (2023), *Robust Bayesian variable selection for gene–environment interactions*, Biometrics, **79**(2), 684–694. https://doi.org/10.1111/biom.13670.

Royston, P. and Altman, D. (1994), *Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling*, Journal of the Royal Statistical Society Series C: Applied Statistics, **46**(3), 429–467. https://doi.org/10.2307/2986270.

Royston, P. and Altman, D. (1997), *Approximating Statistical Functions by Using Fractional Polynomial Regression*, Journal of the Royal Statistical Society Series D: The Statistician, **46**(3), 411–422. https://doi.org/10.1111/1467-9884.00093.

Royston, P. and Sauerbrei, W. (2008), *Multivariable Model-building: A Pragmatic Approach to Regression Analysis based on Fractional Polynomials for Modelling Continuous Variables*, Wiley Series in Probability and Statistics. Wiley, Chichester.

Ryoo, J. H., Konold, T. R., Long, J. D., Molfese, V. J. and Zhou, X. (2017), *Nonlinear Growth Mixture Models With Fractional Polynomials: An Illustration With Early Childhood Mathematics Ability*, Structural Equation Modeling: A Multidisciplinary Journal , **24**(6), 897–910. https://doi.org/10.1080/10705511.2017.1335206.

Sabanés Bové, D. and Held, L. (2011), *Bayesian fractional polynomials*, Statistics and Computing, **21**, 309–324. https://doi.org/10.1007/s11222-010-9170-7.

Saul, L. K., Jaakkola, T. and Jordan, M. I. (1996), *Mean Field Theory for Sigmoid Belief Networks*, Journal of Artificial Intelligence Research, **4**, 61–76. https://doi.org/10.1613/jair.251.

Sinharay, S. (2003), *Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review*, ETS Research Report Series, **2003**(1), i–52. https://doi.org/10.1002/j.2333-8504.2003.tb01899.x.

Sriram, K., Ramamoorthi, R. and Ghosh, P. (2013), *Posterior Consistency of Bayesian Quantile Regression Based on the Misspecified Asymmetric Laplace Density*, Bayesian Analysis, **8**(2), 479–504. https://doi.org/10.1214/13-BA817.

Stamey, T. A. and Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and Yang, N. (1989), *Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate. II. Radical Prostatectomy Treated Patients*, The Journal of Urology, **141**(5), 1076–1083. https://doi.org/10.1016/S0022-5347(17)41175-X.

Su, M. and Wang, W. (2021), *Elastic Net penalized quantile regression model*, Journal of Computational and Applied Mathematics, **392**. Article no. 113462. https://doi.org/10.1016/j.cam.2021.113462.

Sun, Q., Zhou, W.-X. and Fan, J. (2020), *Adaptive Huber Regression*, Journal of the American Statistical Association, **115**(529), 254–265. https://doi.org/10.1080/01621459.2018.1543124.

Sun, W., Ibrahim, J. G. and Zou, F. (2010), *Genomewide Multiple-Loci Mapping in Experimental Crosses by Iterative Adaptive Penalized Regression*, Genetics, **185**(1), 349–359. https://doi.org/10.1534/genetics.110.114280.

Takagi, H. and Umemoto, T. (2013), *The Lower, the Better? Fractional Polynomials Meta-Regression of Blood Pressure Reduction on Stroke Risk*, High Blood Pressure & Cardiovascular Prevention, **20**, 135–138. https://doi.org/10.1007/s40292-013-0016-1.

Tan, Q., Thomassen, M., Hjelmborg, J. V. B., Clemmensen, A., Andersen, K. E., Petersen, T. K., McGue, M., Christensen, K. and Kruse, T. A. (2011), *A Growth Curve Model with Fractional Polynomials for Analysing Incomplete Time-Course Data in Microarray Gene Expression Studies*, Advances in Bioinformatics, **2011**. Article no. 261514. https://doi.org/10.1155/2011/261514.

Thompson, M. L., Williams, M. A. and Miller, R. S. (2009), *Modelling the association of blood pressure during pregnancy with gestational age and body mass index*, Paediatric and Perinatal Epidemiology, **23**(3), 254–263. https://doi.org/10.1111/j.1365-3016.2009.01027.x.

Tibshirani, R. (1996), *Regression Shrinkage and Selection Via the Lasso*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **58**(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tierney, L., Kass, R. and Kadane, J. (1989), *Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions*, Journal of American Statistical Association, **84**(407), 710–716. https://doi.org/10.1080/01621459.1989.10478824.

Tilling, K., Macdonald-Wallis, C., Lawlor, D. A., Hughes, R. A. and Howe., L. D. (2014), *Modelling Childhood Growth Using Fractional Polynomials and Linear Splines*, Annals of Nutrition & Metabolism, **65**(2–3), 129–138. https://doi.org/10.1159/000362695.

Tran, M.-N., Nguyen, T.-N. and Dao, V.-H. (2021), *A practical tutorial on Variational Bayes*, arXiv preprint, arXiv:2103.01327. https://doi.org/10.48550/arXiv.2103.01327.

Tukey, J. W. (1960), *A Survey of Sampling from Contaminated Distributions*, Contributions to Probability and Statistics, 448–485. https://cir.nii.ac.jp/crid/1570291226404846720.

U.S. Census Bureau (2006), *Statistical Abstract of the United States*, United States. Retrieved from the Library of Congress: https://www.loc.gov/item/lcwaN0014424/.

Wang, C. and Blei, D. M. (2013), *Variational Inference in Nonconjugate Models*, Journal of Machine Learning Research, **14**, 1005–1031. https://jmlr.org/papers/volume14/wang13b/wang13b.pdf.

Wong, E. S., Wang, B. C., Garrison, L. P., Alfonso-Cristancho, R., Flum, D. R., Arterburn, D. E. and Sullivan, S. D. (2011), *Examining the BMI-mortality relationship using fractional polynomials*, BMC Medical Research Methodology, **11**(1), 1–11. https://doi.org/10.1186/1471-2288-11-175.

Wang, J. (2012), *Bayesian quantile regression for parametric nonlinear mixed effects models*, Statistical Methods & Applications, **21**, 279–295. https://doi.org/10.1007/s10260-012-0190-7.

Wang, Z., Wu, Y. and Cheng, W. (2023), *Variational inference on a Bayesian adaptive Lasso Tobit quantile regression model*, Stat, **12**(1). e563. https://doi.org/10.1002/sta4.563.

West, M. (1987), *On scale mixtures of normal distributions*, Biometrika, **74**(3), 646–648. https://doi.org/10.1093/biomet/74.3.646.

Whelton, P., Carey, R., Aronow, W., Casey, D., Collins, K., Dennison Himmelfarb, C., DePalma, S., Gidding, S., Jamerson, K., Jones, D. and MacLaughlin, E. (2018), *2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines*, Journal of the American College of Cardiology, **71**(19), e127–e248. https://doi.org/10.1016/j.jacc.2017.11.006.

Wilcox, R. R. (2003), *Applying Contemporary Statistical Techniques*, Elsevier.

World Health Organisation (2013a), *A Global Brief on Hypertension: Silent Killer, Global Public Health Crisis*, World Health Organisation, 1–39.

World Health Organization (2013b), *Global Introduction of Hypertension*, Geneva: World Health Organization.

Wu, P., Dupuis, J. and Liu, C. T. (2023), *Identifying important gene signatures of BMI using network structure-aided nonparametric quantile regression*, Statistics in Medicine, **42**(10), 1625–1639. https://doi.org/10.1002/sim.9691.

Wu, Y. and Liu, Y. (2009), *Variable Selection in Quantile Regression*, Statistica Sinica, **19**(2), 801–817. http://www.jstor.org/stable/24308857.

Wu, Y. and Tang, N. (2021), *Variational Bayesian partially linear mean shift models for high-dimensional Alzheimer's disease neuroimaging data*, Statistics in Medicine, **40**(15), 3604–3624. https://doi.org/10.1002/sim.8985.

Yang, K., Peng, B. and Dong, X. (2023), *Bayesian inference for quantile autoregressive model with explanatory variables*, Communications in Statistics-Theory and Methods, **52**(9), 2946–2965. https://doi.org/10.1080/03610926.2021.1964529.

Yedidia, J. S., Freeman, W. T. and Weiss, Y. (2001), *Bethe free energy, Kikuchi approximations, and belief propagation algorithms*, Advances in Neural Information Processing Systems, **13**(24). https://api.semanticscholar.org/CorpusID:13980420.

Yedidia, J. S., Freeman, W. and Weiss, Y. (2000), *Generalized Belief Propagation*, Advances in Neural Information Processing Systems, **13**. https://proceedings.neurips.cc/paper_files/paper/2000/file/61b1fb3f59e28c67f3925f3c79be81a1-Paper.pdf.

Yeo, J., Gulsin, G., Brady, E., Dattani, A., Bilak, J., Marsh, A., Sian, M., Athithan, L., Parke, K., Wormleighton, J. and Graham-Brown, M. (2022), *Association of ambulatory blood pressure with coronary microvascular and cardiac dysfunction in asymptomatic type 2 diabetes*, Cardiovascular Diabetology, **21**(1), 1–13. https://doi.org/10.1186/s12933-022-01528-2.

Yi, J. and Tang, N. (2022), *Variational Bayesian Inference in High-Dimensional Linear Mixed Models*, Mathematics, **10**(3). Article no. 463. https://doi.org/10.3390/math10030463.

Yu, H. and Yu, L. (2023), *Flexible Bayesian quantile regression for nonlinear mixed effects models based on the generalized asymmetric Laplace distribution*, Journal of Statistical Computation and Simulation, 1–26. https://doi.org/10.1080/00949655.2023.2204437.

Yu, K., Lu, Z. and Stander, J. (2003), *Quantile Regression: Applications and Current Research Areas*, Journal of the Royal Statistical Society Series D: The Statistician, **52**(3), 331–350. https://doi.org/10.1111/1467-9884.00363.

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

Yu, K., Van Kerm, P. and Zhang, J. (2005), *Bayesian Quantile Regression: An Application to the Wage Distribution in 1990s Britain*, Sankhyā: The Indian Journal of Statistics, **67**(2), 359–377. https://www.jstor.org/stable/25053437.

Zhan, Z., Bastide-Van Gemert, S. L., Wiersum, M., Heineman, K. R., Hadders-Algra, M. and Heuvel, E. V. D. (2021), *A comparison of statistical methods for age-specific reference values of discrete scales*, Communications in Statistics - Simulation and Computation, 1–18. https://doi.org/10.1080/03610918.2021.1977824.

Zhou, B., Carrillo-Larco, R., Danaei, G., Riley, L., Paciorek, C., Stevens, G., Gregg, E., Bennett, J., Solomon, B., Singleton, R. and Sophiea, M. (2021), *Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants*, The Lancet, **398**(10304), 957–980. https://doi.org/10.1016/S0140-6736(21)01330-1.

Zou, H. (2006), *The Adaptive Lasso and Its Oracle Properties*, Journal of the American Statistical Association, **101**(476), 1418–1429. https://doi.org/10.1198/016214506000000735.

Zou, H. and Hastie, T. (2005), *Regularization and Variable Selection Via the Elastic Net*, Journal of the Royal Statistical Society Series B: Statistical Methodology, **67**(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00527.x.