View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Washington University St. Louis: Open Scholarship

Washington University in St. Louis
Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations | Arts & Sciences

Spring 5-17-2019

Variational Inference for Quantile Rgression

Bufei Guo
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

 Part of the Statistics and Probability Commons

Recommended Citation
Guo, Bufei, "Variational Inference for Quantile Rgression" (2019). *Arts & Sciences Electronic Theses and Dissertations*. 1743.
https://openscholarship.wustl.edu/art_sci_etds/1743

This Thesis is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics

Variational Inference for Quantile Rgression

by

Bufei Guo

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Master of Arts

May 2019
St. Louis, Missouri

Table of Contents

List of Figures

Acknowledgments

Dissertation Examination Committee:
Nan Lin
Jose Figueroa-Lopez

Foremost, I would like to thank my advisor Prof. Nan Lin for all the support of my A.M. thesis study and research, for his patience, motivation, enthusiasm, and immense knowledge. Without his guidance, the completion of this thesis would have been unreachable. I would like to thank Prof. Jose Figueroa-Lopez for his help and insightful suggestions on my thesis.

I would also like to thank the Department of Mathematics and Statistics for the generous support, without which I could not obtain my master's degree. I am full of gratitude to all the faculties in the department, from whom I learned so much knowledge and who kindly helped me.

Last but not least, I would like to thank my family and friends for their unconditional love and support.

Bufei Guo

Washington University in St. Louis

May 2019

Dedicated to My Parents.

Abstract of the Thesis

Variational Inference for Quantile Rgression

by

Bufei Guo

A.M. in Statistics

Washington University in St. Louis, 2019.

Professor Nan Lin, Chair

Quantile regression (QR) (Koenker and Bassett, 1978), is an alternative to classic linear regression with extensive applications in many fields. This thesis studies Bayesian quantile regression (Yu and Moyeed, 2001) using variational inference, which is one of the alternative methods to the Markov chain Monte Carlo (MCMC) in approximating intractable posterior distributions. The lasso regularization is shown to be effective in improving the accuracy of quantile regression (Li and Zhu, 2008). This thesis developed variational inference for quantile regression and regularized quantile regression with the lasso penalty. Simulation results show that variational inference is a computationally more efficient alternative to the MCMC, while providing a comparable accuracy.

1. Introduction

Regression is a technique used to explain the relationship between explanatory variable $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ and a response variable $\mathbf{y}$. Least square estimation (LSE) is one of the widely used methods, that targets the conditional expectation $\mathbb{E}(\mathbf{y}|X = [\mathbf{x}_1, \ldots, \mathbf{x}_n])$. When heterogeneity is present in the random error, it provides an inadequate description of the distribution of response variable $\mathbf{y}$ as only the average relationship between X and $\mathbf{y}$ is considered. Quantile regression(QR) was first introduced by Koenker and Bassett (1978), which provides an alternative to least square estimator, especially for the linear model with non-Gaussian errors. QR is able to provide a more complete description of the relationship between the explanatory variables and the response by modeling the conditional distribution $\mathbf{y}|X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$at different quantiles. Oftentimes, QR gives comparable estimation accuracy as the least square method under Gaussian errors and provides a more robust alternative when the "outliers" in the model are difficult to identify. QR has a broad application in many fields like survival analysis (Koenker and Geling, 2001), financial economics (Bassett and Chen, 2001) and environmental modeling (Pandey and Nguyen, 1999).

1.1 Bayesian quantile regression

Consider $X = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})'$ is the explanatory variable and $\mathbf{y} = (y_1, \ldots, y_n)'$ is the response variable. The pth $(0 < p < 1)$ conditional quantile of y_i given $\mathbf{x}_i$ is defined as

$$Q_p(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}_p,$$

where $\boldsymbol{\beta}_p \in \mathbf{R}^k$ is the vector of coefficients. The pth $(0 < p < 1)$ quantile regression estimator of $\boldsymbol{\beta}$ is the solution to the quantile regression minimization problem given by

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i'\boldsymbol{\beta}), \tag{1.1}$$

where $\rho_p(\cdot)$ is an asymmetrix loss function,

$$\rho_p(u) = u(p - \mathbf{I}(u < 0)), \tag{1.2}$$

and $\mathbf{I}(\cdot)$ is the indicator function. Equivalently, we can write (1.2) as

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

Figure 1.1 shows the loss functions at three different quantiles, namely $p = 0.25$, 0.50 and 0.75.

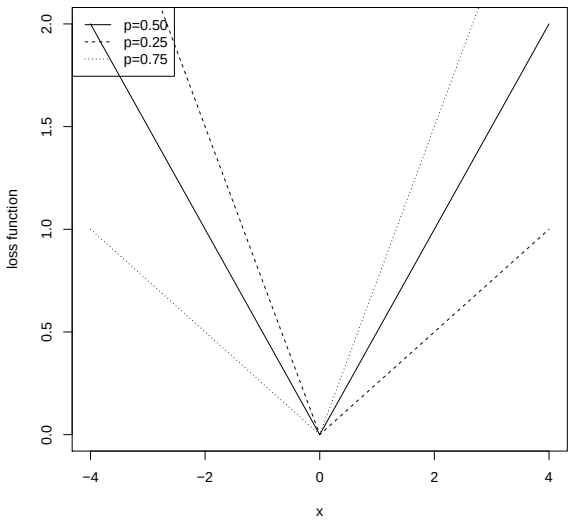

Figure 1.1. Loss function in (1.2) at different quantiles

Regularization, e.g. lasso (Tibshirani, 1996), is also adapted in QR to prevent over-fitting and improve prediction when the explanatory variables are high-dimensional (Li et al., 2010). This problem can be modified as an optimization over the quantile functions $f_i = \mathbf{x}_i\boldsymbol{\beta},\ \ i = 1, \ldots, n$ for

$$\sum_{i=1}^{n} \rho_p(y_i - f_i) + \lambda||f||_q,$$

where $f = (f_1, \ldots, f_n)$ and $||\cdot||_q$ is the qth norm (Abeywardana and Ramos, 2015). For example, when $q = 1$, this is the lasso penalty.

In Bayesian quantile regression, the coefficient $\boldsymbol{\beta}_p$ is sampled from its posterior distribution using the random walk Metropolis-Hastings algorithm (Yu and Moyeed, 2001) or Gibbs samplers (Tsionas, 2003). Li et al. (2010) studied Bayesian regularized quantile regression with the group lasso and elastic net penalty. Yu and Moyeed (2001) proposed that QR can be incorporated into Bayesian inference framework by assuming the error

terms follow the asymmetric Laplace distribution (ALD). Based on the ALD distributed error assumption, the likelihood function of $\boldsymbol{\beta}$ can be constructed, then its posterior distribution can be derived using Bayes' theorem.

Asymmetric Laplace distribution

The probability density function (pdf) of an asymmetric Laplace distribution (ALD) is defined as

$$f(x;\mu,\sigma,p) = \frac{p(1-p)}{\sigma}\exp\left(-\frac{(x-\mu)}{\sigma}[p-\boldsymbol{I}(x\leq\mu)]\right), \quad x\in(-\infty,\infty), \tag{1.3}$$

where $0<p<1$ is the skewness parameter, $\sigma>0$ is the scale parameter, and $\mu\in\mathbb{R}$ is the location parameter. The distribution ALD $(x;\mu,\sigma,p)$ has mean $\mathbb{E}(x)=\frac{1-2p}{p(1-p)}$ and variance $\mathrm{Var}(x)=\frac{1-2p+2p^2}{p^2(1-p)^2}$. The corresponding CDF and quantile function are, respectively,

$$F(x;\mu,\sigma,p) = \begin{cases} p\exp\left(\frac{1-p}{\sigma}(x-mu)\right), & x\leq\mu, \\ 1-(1-p)\exp\left(-\frac{p}{\sigma}(x-\mu)\right), & x>\mu, \end{cases} \tag{1.4}$$

and

$$F^{-1}(x;\mu,\sigma,p) = \begin{cases} \mu\exp\left(\frac{\sigma}{1-p}\log(\frac{x}{p})\right), & 0\leq x<p, \\ \mu-\frac{\sigma}{p}\log\left(-\frac{1-x}{1-p}\right), & p<x\leq 1. \end{cases} \tag{1.5}$$

As the pth quantile of the ALD distribution equals to the location parameter μ, i.e. $F^{-1}(x;\mu,\sigma,p)|_{x=p}=\mu$, the ALD is used as the error distribution in quantile regression models (Yu and Moyeed, 2001).

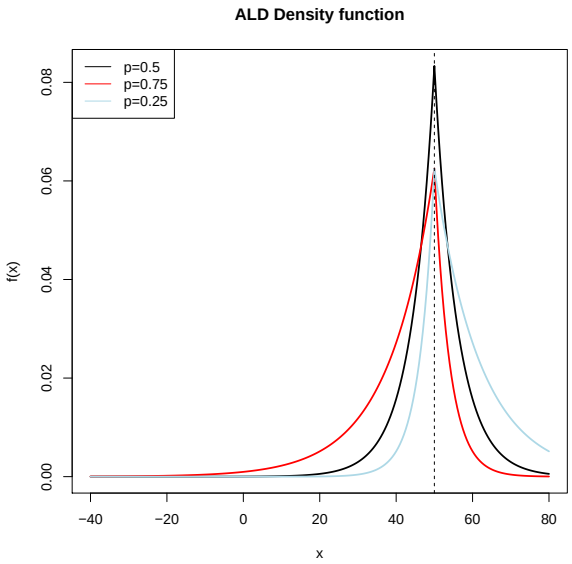

Figure 1.2. Density of ALD with $\mu = 50$, $\sigma = 1$, and $p = (0.25,\ 0.5,\ 0.75)$

The ALD can also be represented as a the mixture of an exponential and a normal distribution (Reed and Yu, 2009). If a variable ϵ follows the ALD in (1.3), we can represent ϵ as a location-scale mixture of normal distributions given by

$$\epsilon = \theta z + \tau\sqrt{z}u, \tag{1.6}$$

where

$$\theta = \frac{1-2p}{p(1-p)} \text{ and } \tau^2 = \frac{2}{p(1-p)}.$$

1.1.1 Quantile regression under the asymmetric Laplace distributed error

Consider the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_p + \epsilon_i, \quad i = 1, \ldots, n \tag{1.7}$$

where $y_i \in \mathbb{R}$ is the response variable, $\mathbf{x}_i \in \mathbb{R}^k$ is the explanatory variable, $\boldsymbol{\beta}_p \in \mathbb{R}^k$ is the regression parameter for the pth quantile and $\epsilon_i \sim \text{ALD}(0, \sigma, p)$ is the error term. By representing ϵ_i as in (1.6), Equation (1.7) can be rewritten as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_p + \theta z_i + \tau\sqrt{\sigma z_i}u_i, \quad i = 1, \ldots n, \tag{1.8}$$

where $u_i \sim \text{N}(0, 1)$ and z_i follows the exponential distribution with rate σ, e.g., $\exp(\sigma)$. Both θ and τ are constants with

$$\theta = \frac{1-2p}{p(1-p)} \text{ and } \tau^2 = \frac{2}{p(1-p)}.$$

From (1.8), y_i also follows an asymmetric Laplace distribution with location parameter $\mathbf{x}_i\boldsymbol{\beta_p}$, scale parameter σ and asymmetry parameter p, e.g., $\text{ALD}(\mathbf{x}_i'\boldsymbol{\beta}_p, \sigma, p)$. The conditional distribution of y_i given z_i is a normal distribution with mean $\mathbf{x}_i\boldsymbol{\beta}_p - \theta z_i$ and variance $\tau^2 z_i$. The conditional density of $\mathbf{y} = (y_1, \ldots, y_n)'$ given $\mathbf{z} = (z_1, \ldots, z_n)'$ and $\boldsymbol{\beta_p}$ is

$$f(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z}) \propto (\Pi_{i=1}^n z_i^{-\frac{1}{2}}) \exp\left(-\Pi_{i=1}^n \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_p - \theta z_i)^2}{2\tau^2 z_i}\right). \tag{1.9}$$

Connection with Gibbs sampling

The conditional distribution $p(z_j|\mathbf{z}_{-j}, \boldsymbol{\beta}_p, \mathbf{y})$ is known as the *full conditional* in MCMC (Casella and George, 1992). The approximating density function in quantile regression can be derived from the full conditional distribution of variables in Gibbs sampling. We will first review quantile regression using Gibbs sampling which iteratively samples from the full conditional distributions.

Gibbs sampling for quantile regression

We first consider the model in (1.8). First we consider the quantile regression with the scale parameter σ fixed at $\sigma = 1$. Let the prior distribution be

$$\boldsymbol{\beta}_p \sim N(\mu_{p0}, \Sigma_{p0}), \quad z_i \sim \exp(1). \tag{1.10}$$

The full conditional density of $\boldsymbol{\beta}_p$ can be shown as (Kozumi and Kobayashi, 2011)

$$\boldsymbol{\beta_p}|\mathbf{y}, \mathbf{z} \sim N(\mu_p, \Sigma_p), \tag{1.11}$$

where

$$\Sigma_p^{-1} = \sum_i^n \frac{\mathbf{x}_i \mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}, \tag{1.12}$$

$$\boldsymbol{\mu}_p = \Sigma_p \left(\sum_{i=1}^n \frac{\mathbf{x}_i (y_i - \theta z_i)}{\tau^2 z_i} + \Sigma_{p0}^{-1} \boldsymbol{\mu}_{p0} \right). \tag{1.13}$$

The full conditional density of $z_i, i = 1, ..., n$ is a generalized inverse Gaussian distribution(Kozumi and Kobayashi, 2011)

$$\mathbf{z}_i|\mathbf{y}, \boldsymbol{\beta_p} \sim GIG(\frac{1}{2}, a_i, b_i), \tag{1.14}$$

where

$$a_i = 2 + \frac{\theta^2}{\tau^2} \text{ and } \quad b_i = \frac{(y_i - x_i' \boldsymbol{\beta}_p)^2}{\tau^2}. \tag{1.15}$$

The pdf of a generalized inverse Gaussian distribution is given by

$$f(p, a, b) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp(-\frac{1}{2}(ax + b/x)), a > 0, b > 0, p \in \mathbb{R}, \tag{1.16}$$

where $K_p(\cdot)$is the Bessel function of the third type. Next we extend our discussion to the more general case by treating σ as an unknown variable and a prior is assigned to σ. The prior distribution for $\boldsymbol{\beta}_p$ is the same as (1.11)

$$\boldsymbol{\beta}_p \sim \mathrm{N}(\mu_{p0}, \Sigma_{p0}), \tag{1.17}$$

and a prior was given to σ

$$\sigma \sim \text{IG}(m_0, n_0), \tag{1.18}$$

where IG$(\cdot)$ is the inverse gamma distribution. Similar to the previous situation, the posterior for $\boldsymbol{\beta}_p$ and $\mathbf{z}$ is the same as (1.11) and (1.14) respectively. The full conditional density for σ is then given by

$$\sigma|\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z} \sim IG(m_p, n_p), \tag{1.19}$$

where

$$m_p = 3n + m_0$$
$$n_p = n_0 + \sum_{i=1}^{n}(z_i + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_p - \theta z_i)^2}{2\tau^2 z_i})$$

Then the algorithm of Gibbs sampler iterates between the full conditional distribution of $\boldsymbol{\beta}_p$ given $\mathbf{y}$, $\mathbf{z}$, σ, the full conditional distribution of z_i given $\mathbf{y}$, $\boldsymbol{\beta}_p$, σ, and the full conditional distribution of σ given $\mathbf{y}$, $\boldsymbol{\beta}_p$, $\mathbf{z}$. But this process can be time consuming when the parameter space is high dimensional and the data set is large.

Gibbs Sampling for quantile regression with the lasso penalty

The lasso regularization on quantile regression is brought up in Li and Zhu (2008), where the L_1-norm penalty (lasso) is added to the minimization problem

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_p(y_i - \mathbf{x}_i'\boldsymbol{\beta}) + \lambda \sum_{i=1}^{k} |\beta_i|.$$

Li et al. (2010) proposed an equivalent Bayesian formulation to the problem by putting a Laplace prior with mean zero and scale $\frac{\sigma}{\lambda_j}, j = 1, \ldots, m$ on $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta}_p|\sigma, \boldsymbol{\lambda}) = \prod_{j=1}^{k} \frac{\lambda_j}{\sigma} \exp(-\frac{\lambda_j}{\sigma}|\beta_j|), \tag{1.20}$$

which leads to the posterior distribution

$$p(\boldsymbol{\beta}_p|\mathbf{y},\sigma,\mathbf{z}) \propto \exp(-\frac{1}{\sigma}\sum_{i=1}^{n}\rho_p(y_i-\mathbf{x}_i'\boldsymbol{\beta}_p)-\frac{\lambda_j}{\sigma}\sum_{j=1}^{k}|\beta_j|), \tag{1.21}$$

where λ_j's are the regularization parameter for corresponding regression coefficients β_j. We put an inverse gamma prior on σ and a gamma prior on $\eta_j = \frac{\lambda_j}{\sigma}, \quad j = 1,\ldots,m$. Let $\mathbf{s} = (s_1,\ldots,s_k)$ and the prior of $\boldsymbol{\beta}_p$ can be further written as

$$\begin{aligned} p(\boldsymbol{\beta}_p|\sigma,\boldsymbol{\lambda}) &= \prod_{j=1}^{k}\frac{\lambda_j}{\sigma}\exp\left(-\frac{\lambda_j}{\sigma}|\beta_j|\right) \\ &= \prod_{j=1}^{k}\int_0^\infty \frac{1}{\sqrt{a\pi s_j}}\exp\left(-\frac{\beta_j^2}{2s_j}\right)\frac{\eta_j^2}{2}\exp\left(-\frac{\eta_j^2 s_j}{2}\right)ds_j. \end{aligned} \tag{1.22}$$

Then the full conditional distribution of β_j is $N(\tilde{\mu}_j,\tilde{\omega}_j^2)$, with

$$\tilde{\mu}_j = \tilde{\omega}_j^2\frac{1}{\sigma\tau^2}\sum_{i=1}^{n}y_{i,j}x_{i,j}z_i^{-1}, \tag{1.23}$$

and

$$\tilde{\omega}_j^{-2} = \frac{1}{\sigma\tau^2}\sum_{i=1}^{n}x_{i,j}^2 z_i^{-1} + s_j^{-1}, \tag{1.24}$$

where

$$y_{i,j} = y_i - \theta z_i - \sum_{l=1,l\neq j}^{k}x_{i,l}\beta_l. \tag{1.25}$$

The full conditional distribution of z_i follows the same distribution as the previous case in (1.14), which is $GIG(\frac{1}{2},\tilde{a}_i,\tilde{b}_i)$, with

$$\tilde{a}_i = \frac{\theta^2}{\sigma\tau^2}+\frac{2}{\sigma}, \tag{1.26}$$

and

$$\tilde{b}_j = \frac{(y_i-\mathbf{x}_i'\boldsymbol{\beta}_p)^2}{\sigma\tau^2}. \tag{1.27}$$

The full conditional distribution for s_j is given by

$$\begin{aligned} p(s_j|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}_p, \mathbf{s}_{-j}, \tau, \eta^2) &\propto p(\beta_j|s_j)p(s_j|\eta^2) \\ &\propto s_k^{-1/2} \exp\{-\frac{1}{2}(\eta^2 s_j + \beta_j^2 s_j^{-1})\}, \end{aligned} \tag{1.28}$$

which is $GIG(\frac{1}{2}, \eta_j^2, \beta_j^2)$. The full conditional distribution for σ is also an inverse gamma distribution $IG(\tilde{m}, \tilde{n})$, with

$$\tilde{m} = 3n + m_0, \tag{1.29}$$

and

$$\tilde{n} = n_0 + \sum_{i=1}^{n}(z_i + \frac{(y_i - \mathbf{x}_i'\boldsymbol{\beta}_p - \theta z_i)^2}{2\tau^2 z_i}). \tag{1.30}$$

The full conditional distribution of η^2 follows a gamma distribution with shape parameter $c+1$ and rate parameter $\frac{1}{2}\sum_{j=1}^{k} s_k + d$, which is given by

$$\begin{aligned} p(\eta^2|\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}_p, \mathbf{s}, \sigma) &\propto p(\mathbf{s}|\eta^2)p(\eta^2) \\ &\propto (\eta^2)^{k+c-1} \exp\left\{-\eta^2\left(\sum_{j=1}^{k}\frac{s_k}{2} + d\right)\right\}, \end{aligned} \tag{1.31}$$

where c and d are constants given in the joint prior distribution of τ and η^2

$$\tau, \eta^2 \sim \tau^{\psi-1}\exp(-\xi\tau)(\eta^2)^{c-1}\exp(-d\eta^2). \tag{1.32}$$

1.1.2 Variational inference

Variational inference is one of the popular methods to approximate intractable or difficult-to-compute posterior distributions $p(\mathbf{y}|\cdot)$ with an approximate posterior distribution $q(\mathbf{y})$. Compared with MCMC such as Gibbs sampling, variational inference tends to be faster while achieves comparable prediction especially when dealing with large-scale data sets (Blei et al., 2016).

The idea of variational inference is to approximate the conditional density of latent variables given observed variables using optimization. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the set of observed variables and $\mathbf{z} = (z_1, \ldots, z_m)$ be the set of latent variables. The joint density of $\mathbf{x}$ and $\mathbf{z}$ is $p(\mathbf{x}, \mathbf{z})$. In the case that the conditional distribution $p(\mathbf{z}|\mathbf{x})$ is not directly tractable, variational inference provides an alternative approach by approximating the conditional distribution $p(\mathbf{z}|\mathbf{x})$ using a tractable distribution $q^*(\mathbf{z}) \in \Theta$, where Θ is the family of densities over the latent variables. All density functions $q(\mathbf{z}) \in \Theta$ are candidate approximations to $p(\mathbf{z}|\mathbf{x})$. By solving the optimization problem, one can try to find the member in Θ that is the closest to $p(\mathbf{z}|\mathbf{x})$ in the Kullback-Leibler (KL) distance (Blei et al., 2016)

$$q^*(\mathbf{z}) = \underset{q(\mathbf{z} \in Q)}{\operatorname{argmin}} \mathrm{KL}\big(q(\mathbf{z}||p(\mathbf{z}|\mathbf{x}))\big), \tag{1.33}$$

where $\mathrm{KL}(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$ is the Kullback-Leiler distance between the posterior distribution $p(\mathbf{z}|\mathbf{x})$ and the candidate distribution $q(\mathbf{z})$ in the family Θ. It is defined as

$$\mathrm{KL}\big(q(\mathbf{z}||p(\mathbf{z}|\mathbf{x}))\big) = \mathbb{E}(\log q(\mathbf{z})) - \mathbb{E}(\log p(\mathbf{z}|\mathbf{x})), \tag{1.34}$$

which is always non-negative (van Erven and Harremos, 2014). Wang and Blei (2018) gave the asymptotic properties of variational inference by proving that the posterior density given by variational inference converges to the KL minimizer of a normal distribution centered at the truth. Zhang and Gao (2017) proved that the upper bound of the convergence rate, at which the variational posterior $q^*(\mathbf{z})$ converges to the true posterior $p(\mathbf{z}|\mathbf{x})$ is given by

$$\epsilon_n^2 + \frac{1}{n} \inf_{q(\mathbf{z}) \in \Theta} p_0^{(n)} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})), \tag{1.35}$$

where ϵ_n^2 is the rate of convergence of the posterior distribution $p(\mathbf{z}|\mathbf{x})$. The second term is the variational approximation error with respect to Θ under $p_0^{(n)}$, where $p_0^{(n)}$ is the process that generates all the $\mathbf{x}_i$'s. If $q(\mathbf{z})$ equals to the exact posterior distribution $p(\mathbf{z}|\mathbf{x})$, the second term will be zero. The convergence rate will be the convergence rate of the posterior distribution given by MCMC. In general, the second term in (1.35)

$$\frac{1}{n} \inf_{q(\mathbf{z}) \in \Theta} p_0^{(n)} KL(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x}))$$

is dominated by the first term ϵ_n^2 in (1.35). Variational inference does not require the sampling process required in MCMC and Gibbs sampling. Hence, it provides computational advantages without violating the asymptotic property of estimators in large-scale data set situations.

2. Variational inference for quantile regression

2.1 Algorithm of variational Bayes

Numerical implementation of the variational inference, the CAVI (coordinate ascent variational inference) algorithm in Blei et al. (2016), is closely related to Gibbs sampling. In each iteration, CAVI optimizes every parameter sequentially, while keep others fixed. Finally, a local optimum is reached. Consider the model with parameter vector (latent variable) $\boldsymbol{\theta}$ and observed variable $\mathbf{y}$. Bayesian inference is based on the posterior density function

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})}. \tag{2.1}$$

Let $q(\cdot)$ be an arbitrary density function over the density family Θ. The logarithm of the marginal likelihood function satisfies

$$\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y}) \int q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z}) / q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y}) / q(\boldsymbol{\beta}_p, \mathbf{z})} \right\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&= \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})} \right\} d\boldsymbol{\beta}_p d\mathbf{z} + \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})} \right\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&\geq \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})} \right\} d\boldsymbol{\beta}_p d\mathbf{z},
\end{aligned} \tag{2.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}_p, \mathbf{z})$ and $q(\cdot) \in \Theta$ is the candidate distribution used to approximate $p(\cdot)$. The above inequality holds because the second integral in (2.2)

$$\int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{q(\boldsymbol{\beta}_p, \mathbf{z})}{p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})} \right\} d\boldsymbol{\beta}_p d\mathbf{z}, \tag{2.3}$$

is the KL distance between $q(\boldsymbol{\beta}_p, \mathbf{z})$ and $p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})$, which is always non-negative by definition (Kullback and Leibler, 1951). The equality holds if and only if $q(\boldsymbol{\beta}_p, \mathbf{z}) = p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})$. Under this special case, the estimation of variational inference will coincide with the estimation given by Gibbs sampling. Recall from Equation (1.33), that the goal of variational inference is to find the distribution $q(\cdot)$ that is closest to the conditional distribution $p(\boldsymbol{\theta}|\mathbf{y})$ in KL distance. According to (2.2), minimizing the KL distance in (2.3) between $q(\boldsymbol{\beta}_p, \mathbf{z})$ and $p(\boldsymbol{\beta}_p, \mathbf{z}|\mathbf{y})$ is equivalent to maximizing the lower bound

$$L = \int q(\boldsymbol{\beta}_p, \mathbf{z}) \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})} \right\} d\boldsymbol{\beta}_p d\mathbf{z}. \tag{2.4}$$

In variational inference, the assumption of the complexity of the density family Θ determines the complexity of optimization problem. In the *mean-filed variational family* (a.k.a. *naive mean approach*) (Blei et al., 2016), where the latent variables $\boldsymbol{\theta}$ are assumed to be mutually independent and governed by distinct factors in the variational density $q(\theta_i)$, is used to approximate the conditional distribution $p(\theta_i|\mathbf{y}, \boldsymbol{\theta}_{-i})$. e.g.

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{n} q(\theta_i),$$

where each latent variable θ_i is governed by its own variational factor. One can also use other approximations such as *generalized mean field* (Blei et al., 2016), in which the

parameters of interest are divided into groups and the parameters inside each group are allowed to be dependent. e.g.

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{n} q(\theta_{i_1}, \ldots, \theta_{i_m}).$$

In this thesis, we adopt the *mean-field variational family* approach, by assuming independence between latent variables $\boldsymbol{\beta}_p$ and $\mathbf{z}$. Then the log-likelihood function can be written as

$$\begin{aligned}
\int q(\boldsymbol{\beta}_p, \mathbf{z}) \log\{\frac{p(\mathbf{y}, \boldsymbol{\beta}_p, \mathbf{z})}{q(\boldsymbol{\beta}_p, \mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} &= \int q(\boldsymbol{\beta}_p) q(\mathbf{z}) \log\{\frac{p(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z}) p(\boldsymbol{\beta}_p) p(\mathbf{z})}{q(\boldsymbol{\beta}_p) q(\mathbf{z})}\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&= \int q(\boldsymbol{\beta}_p) q(\mathbf{z}) \log\{p(\mathbf{y}|\boldsymbol{\beta}_p, \mathbf{z})\} d\boldsymbol{\beta}_p d\mathbf{z} \\
&+ \int q(\boldsymbol{\beta}_p) \log\{\frac{p(\boldsymbol{\beta}_p)}{q(\boldsymbol{\beta}_p)}\} d\boldsymbol{\beta}_p \\
&+ \int q(\mathbf{z}) \log\{\frac{p(\mathbf{z})}{q(\mathbf{z})}\} d\mathbf{z},
\end{aligned} \tag{2.5}$$

where $q(\cdot)$ is the candidate density from the density family Θ. Consider the jth variable z_j in the latent variable $\mathbf{z}$. The conditional density of z_j conditioning on all other latent variables and observed variables is

$$p(z_j | \mathbf{z}_{-j}, \boldsymbol{\beta}_p, \mathbf{y}),$$

where $\mathbf{z}_{-j} = (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_n)$. Then one can fix all other variational factors in $\mathbf{z}_{-j}$, and maximize the lower bound of this conditional distribution with respect to the density of z_j. The optimal $q_j^*(z_j) \in \Theta_j$ is proportional to the exponential of the expected log conditional density (Blei et al., 2016)

$$q^*(z_j) \propto \exp(\mathbb{E}_{-j}(\log p(z_j | \mathbf{y}, \mathbf{z}_{-j}))). \tag{2.6}$$

The latent variables are updated successively using (2.6). The iteration stops when the difference between two sequential lower bound is negligible. i.e., smaller than a prespec-

ified tolerance level.

Algorithm: (CAVI)

Step1 Initialize $q(\boldsymbol{\theta})$

Step2 Update $q(z_j)^*, \quad j = 1, \ldots, n$ and $q(\boldsymbol{\beta}_p)$ by

$q^*(z_j) \propto \exp(\mathbb{E}_{-j}(\log p(z_j|\mathbf{y}, \mathbf{z}_{-j})))$,

$\vdots$

$q(\boldsymbol{\beta_p}) \propto \exp(\mathbb{E}_{\mathbf{z}}(\log p(\boldsymbol{\beta_p}|\mathbf{y}, \mathbf{z})))$

Step3 Update the lower bound L, repeat step 2 and 3 until the change in L is negligible.

2.2 Variational inference for quantile regression without regularization

We start from the simplest case, where the scale parameter is fixed at $\sigma = 1$. Then only $\boldsymbol{\beta}_p$ and $\mathbf{z}$ need to be updated in the iterations of variational inference.

The optimal approximation density function $q^*(\boldsymbol{\beta}_p)$ is given by

$$q(\boldsymbol{\beta_p}) \propto \exp(\mathbb{E}_{\mathbf{z}}(\log p(\boldsymbol{\beta_p}|\mathbf{y}, \mathbf{z})))$$

with

$$
\begin{aligned}
\log p(\boldsymbol{\beta}_p|\mathbf{y},\mathbf{z}) = & -\frac{1}{2}\left(\log(2\pi) + \log\left(\det\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}\right)^{-1}\right)\right) \\
& - \frac{1}{2}(\boldsymbol{\beta}_p - \boldsymbol{\mu}_{p0})'\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}\right)(\boldsymbol{\beta}_p - \boldsymbol{\mu}_{p0}) \\
& = -\frac{1}{2}\log\left(\det\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}\right)^{-1}\right) \\
& - \frac{1}{2}\boldsymbol{\beta}_p'\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}\right)\boldsymbol{\beta}_p \\
& + \boldsymbol{\beta}_p'\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i(y_i - \theta z_i)}{\tau^2 z_i} + \Sigma_{p0}^{-1}\boldsymbol{\mu_{p0}}\right) \\
& - \frac{1}{2}\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i(y_i - \theta z_i)}{\tau^2 z_i} + \Sigma_{p0}^{-1}\boldsymbol{\mu_{p0}}\right)'\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i'}{\tau^2 z_i} + \Sigma_{p0}^{-1}\right)^{-1}\left(\sum_{i=1}^{n}\frac{\mathbf{x}_i(y_i - \theta z_i)}{\tau^2 z_i} + \Sigma_{p0}^{-1}\boldsymbol{\mu}_{p0}\right) \\
& + const.
\end{aligned}
\tag{2.7}
$$

The expectation of $\log p(\boldsymbol{\beta_p}|\mathbf{y},\mathbf{z})$ is with respect to z_i, where z_i follows the generalized inverse Gaussian distribution $\mathrm{GIG}(\frac{1}{2}, a_{q_i}, b_{q_i})$, with

$$
\mathbb{E}(z_i) = \frac{\sqrt{b_{q_i}}K_{3/2}(\sqrt{a_{q_i}b_{q_i}})}{\sqrt{a_{q_i}}K_{1/2}(\sqrt{a_{q_i}b_{q_i}})},
$$

$$
\mathbb{E}(\frac{1}{z_i}) = \frac{\sqrt{a_{q_i}}K_{3/2}(\sqrt{a_{q_i}b_{q_i}})}{\sqrt{b_{q_i}}K_{1/2}(\sqrt{a_{q_i}b_{q_i}})} - \frac{1}{b_{q_i}},
$$

and

$$
\mathbb{E}(\ln z_i) = \ln\frac{\sqrt{b_{q_i}}}{\sqrt{a_{q_i}}} + \frac{\partial}{\partial p}\ln K_p(\sqrt{a_{q_i}b_{q_i}}),
$$

where $\mathrm{K}_p(\cdot)$ is the Bessel function with order p. The approximation of the expectations could be used for simplicity in some situations (Abeywardana and Ramos, 2015), with

$$
\mathbb{E}(z_i) = \sqrt{\frac{b_{q_i}}{a_{q_i}}}, \tag{2.8}
$$

and

$$\mathbb{E}(\frac{1}{z_i}) = \sqrt{\frac{a_{q_i}}{b_{q_i}}} - \frac{1}{b_{q_i}}. \tag{2.9}$$

In general the exact values of these expectations are preferred, as the approximated value might cause convergence problems. e.g., the value of the lower bound sometimes diverges if the approximated values are used. The expectation of the logarithm of the full conditional density $\log p(\boldsymbol{\beta}_p|\mathbf{y}, \mathbf{z})$ is

$$\begin{aligned}\mathbb{E}_{\mathbf{z}}(\log p(\boldsymbol{\beta_p}|\mathbf{y}, \mathbf{z})) = \boldsymbol{\beta}'_p(\sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\tau^2}\mathbb{E}(\frac{1}{z_i}) - \sum_{i=1}^{n} \frac{\theta}{\tau^2}\mathbf{x}_i + \Sigma_{p0}^{-1}\boldsymbol{\mu}_{p0}) \\ - \frac{1}{2}\boldsymbol{\beta}'_p(\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}'_i}{\tau^2}\mathbb{E}(\frac{1}{z_i}) + \Sigma_{p0}^{-1})\boldsymbol{\beta}_p \\ + const.\end{aligned} \tag{2.10}$$

Taking exponential of the expectation $\mathbb{E}_{\mathbf{z}}(\log p(\boldsymbol{\beta_p}|\mathbf{y}, \mathbf{z}))$, we then see that the density function of $q(\boldsymbol{\beta}_p)$ is for the multivariate normal distribution $\mathrm{N}(\mu_q, \Sigma_q)$, with

$$\Sigma_q = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}'_i}{\tau^2}\mathbb{E}\left(\frac{1}{z_i}\right) + \Sigma_{p0}^{-1}\right)^{-1}, \tag{2.11}$$

and

$$\mu_q = \Sigma_q\left(\sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\tau^2}\mathbb{E}\left(\frac{1}{z_i}\right) - \sum_{i=1}^{n} \frac{\theta}{\tau^2}\mathbf{x}_i + \Sigma_{p0}^{-1}\boldsymbol{\mu}_{p0}\right). \tag{2.12}$$

The pdf of $q(z_i)$'s are calculated in a similar manner as

$$q(z_i) \propto \exp(\mathbb{E}_{\boldsymbol{\beta}_p}(\log \mathbb{P}(z_i|\mathbf{y}, \boldsymbol{\beta}_p))), \tag{2.13}$$

where

$$\log p(z_i|\mathbf{y}, \boldsymbol{\beta}_p) = -\frac{1}{2}\log(z_i) - \frac{1}{2}(a_i z_i + b_i/z_i) + const. \tag{2.14}$$

The expectation of the log conditional density $\log p(z_i|\mathbf{y}, \boldsymbol{\beta}_p)$ with respect to $\boldsymbol{\beta}_p$ follows the GIG distribution

$$q(z_i) \sim GIG(\frac{1}{2}, a_{q_i}, b_{q_i}), \tag{2.15}$$

with

$$a_{q_i} = 2 + \frac{\theta^2}{\tau^2}, \tag{2.16}$$

and

$$b_{q_i} = \frac{y_i^2 - 2y_i\mathbf{x}_i'\boldsymbol{\mu}_q + \mathbf{x}_i'(\boldsymbol{\mu}_q\boldsymbol{\mu}_q' + \Sigma_q)\mathbf{x}_i}{\tau^2}. \tag{2.17}$$

The μ_q and Σ_q in (2.17) are the mean and variance of $q(\boldsymbol{\beta_p})$, which are given in (2.11) and (2.12). From (2.4), the lower bound is given by

$$\mathbb{E}(\log p(\mathbf{y}|\boldsymbol{\theta})) + \mathbb{E}(\log(p(\boldsymbol{\theta}))) - \mathbb{E}(\log(q(\boldsymbol{\theta}))), \tag{2.18}$$

with $\boldsymbol{\theta} = (\mathbf{z}, \boldsymbol{\beta}_p)$. Then the lower bound l is

$$\begin{aligned} l &= \int q(\mathbf{z})q(\boldsymbol{\beta}_p) \log\left(p(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}_p)\right) d\mathbf{z}d\boldsymbol{\beta}_p + \int q(\mathbf{z}) \log\left(p(\mathbf{z})/q(\mathbf{z})\right) d\mathbf{z} \\ &+ \int q(\boldsymbol{\beta}_p) \log(p(\boldsymbol{\beta}_p)/q(\boldsymbol{\beta}_p))d\boldsymbol{\beta}_p \\ &= \mathbb{E}_{q(\mathbf{z}),q(\boldsymbol{\beta}_p)} \log\left((\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}_p)\right) + \mathbb{E}_{q(\mathbf{z})} \log(p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})} log(q(\mathbf{z})) \\ &+ \mathbb{E}_{q(\boldsymbol{\beta}_p)} \log(p(\boldsymbol{\beta}_p)) - \mathbb{E}_{q(\boldsymbol{\beta}_p)} \log(q(\boldsymbol{\beta}_p)). \end{aligned} \tag{2.19}$$

The variational inference algorithm when $\sigma = 1$ is

Algorithm 1:

Step1 Initialize mean $\boldsymbol{\mu}_q$ and covariance matrix Σ_q.

Step2 Repeat Steps 3-5 if the absolute change in lower bound $l \geq t$, where t is the tolerance given, e.g. $t = 10^{-5}$.

Step3 Update parameters in $q(\mathbf{z})$. $q(z_i) \sim GIG(\frac{1}{2}, a_{q_i}, b_{q_i})$, where

$$a_{q_i} = 2 + \frac{\theta^2}{\tau^2},$$

$$b_{q_i} = \frac{y_i^2 - 2y_i \mathbf{x}_i' \boldsymbol{\mu}_q + \mathbf{x}_i'(\boldsymbol{\mu}_q \boldsymbol{\mu}_q' + \Sigma_q)\mathbf{x}_i}{\tau^2}.$$

Step4 Update parameters in $q(\boldsymbol{\beta}_p)$. $q(\boldsymbol{\beta}_p) \sim N(\mu_q, \Sigma_q)$, where

$$\Sigma_q = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i'}{\tau^2} \mathbb{E}\left(\frac{1}{z_i}\right) + \Sigma_{p0}^{-1} \right)^{-1}$$

$$\mu_q = \Sigma_q \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\tau^2} \mathbb{E}\left(\frac{1}{z_i}\right) - \sum_{i=1}^{n} \frac{\theta}{\tau^2} \mathbf{x}_i + \Sigma_{p0}^{-1} \boldsymbol{\mu}_{p0} \right).$$

Step5 Update lower bound l

$$l = \mathbb{E}_{q(\mathbf{z}), q(\boldsymbol{\beta}_p)}(\mathbf{y}|\mathbf{z}, \boldsymbol{\beta}_p) + \mathbb{E}_{q(\mathbf{z})} \log(p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})} log(q(\mathbf{z}))$$

$$+\mathbb{E}_{q(\boldsymbol{\beta}_p)} \log(p(\boldsymbol{\beta}_p)) - \mathbb{E}_{q(\boldsymbol{\beta}_p)} \log(q(\boldsymbol{\beta}_p))$$

When the scale parameter σ is taken into account, similar as the case in Gibbs sampling, a prior distribution of σ is assumed and we update the value of $\boldsymbol{\beta}_p$, $\mathbf{z}$ and σ successively in the iteration of variational inference. The approximation distribution of $\boldsymbol{\beta}_p$, $\mathbf{z}$ and σ are given by

$$q(\boldsymbol{\beta}_p) \sim N(\boldsymbol{\mu}_q, \Sigma_q), \tag{2.20}$$

$$q(z_i) \sim GIG(a_{q_i}, b_{q_i}), \tag{2.21}$$

and

$$q(\sigma) \sim IG(m_q, n_q), \tag{2.22}$$

with

$$\Sigma_q = \left(\sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i'}{\tau^2} \mathbb{E}\left(\frac{1}{z_i}\right) \mathbb{E}\left(\frac{1}{\sigma}\right) + \Sigma_{p0}^{-1} \right)^{-1}, \tag{2.23}$$

$$\mu_q = \Sigma_q \left(\mathbf{Q} \mathbb{E}\left(\frac{1}{\sigma}\right) + \Sigma_{p0}^{-1} \boldsymbol{\mu}_{p0} \right), \tag{2.24}$$

$$a_{q_i} = \left(2 + \frac{\theta^2}{\tau^2}\right) \mathbb{E}\left(\frac{1}{\sigma}\right), \tag{2.25}$$

$$b_{q_i} = \frac{\mathbf{M}}{\tau^2} \mathbb{E}\left(\frac{1}{\sigma}\right), \tag{2.26}$$

$$m_q = 3n + m_0, \tag{2.27}$$

and

$$n_q = n_0 + \sum_{i=1}^{n} \mathbb{E}(z_i) + \mathbf{N}. \tag{2.28}$$

The $\mathbf{Q}$ in Equation (2.24) is given by

$$\mathbf{Q} = (\sum_{i=1}^{n} \frac{\mathbf{x}_i y_i}{\tau^2} \mathbb{E}(\frac{1}{z_i}) - \sum_{i=1}^{n} \frac{\theta}{\tau^2} \mathbf{x}_i). \tag{2.29}$$

The $\mathbf{M}$ is Equation (2.26) is given by

$$\mathbf{M} = y_i^2 - 2 y_i \mathbf{x}_i' \boldsymbol{\mu}_q + \mathbf{x}_i' (\boldsymbol{\mu}_q \boldsymbol{\mu}_q' + \Sigma_q) \mathbf{x}_i. \tag{2.30}$$

The $\mathbf{N}$ in Equation (2.28) is given by

$$\mathbf{N} = \sum_{i=1}^{n} \frac{M}{2\tau^2} \mathbb{E}\left(\frac{1}{z_i}\right) - \frac{\theta(y_i - \mathbf{x}_i' \boldsymbol{\mu}_p)}{\tau^2} + \frac{\theta^2}{2\tau^2} \mathbb{E}(z_i). \tag{2.31}$$

Algorithm 2:

Step1 Initialize mean $\boldsymbol{\mu}_q$ and covariance matrix Σ_q.

Step2 while absolute change in lower bound $l \geq t$, t is the tolerance given, e.g. $t = 10^{-5}$.

Step3 Update parameters in $q(\mathbf{z})$, using $q(z_i) \sim GIG(\frac{1}{2}, a_{q_i}, b_{q_i})$.

Step4 Update parameters in $q(\boldsymbol{\beta}_p)$, using $q(\boldsymbol{\beta}_p) \sim N(\mu_q, \Sigma_q)$.

Step5 Update parameters in $q(\sigma)$, using $q(\sigma) \sim IG(m_q, n_q)$.

Step6 Update the lower bound l

2.3 Variational inference for quantile regression with the lasso penalty

The approximation density function $q(\cdot)$ is calculated using the same method given in Section (2.2)

$$q(\theta_i) \propto \exp(\mathbb{E}_{\boldsymbol{\theta}_{-i}}(\log p(\theta_i|\boldsymbol{\theta}_{-i}))), \tag{2.32}$$

where $\boldsymbol{\theta}_{-i}$ is the vector of variables without the ith variable θ_i. The density function of $q(\beta_j)$ follows the normal distribution

$$N(\mu_{q_j}, \omega_{q_j}), \tag{2.33}$$

with

$$\omega_{q_j}^{-2} = \frac{\mathbb{E}(1/\sigma)}{\tau^2} \sum_{i=1}^{n} x_{i,j}^2 \mathbb{E}(1/z_i) + \mathbb{E}(1/s_j), \tag{2.34}$$

and

$$\mu_{q_j} = \omega_{q_j}^2 \left(\frac{\mathbb{E}(1/\sigma)}{\tau^2} \sum_{i=1}^{n} x_{i,j} \left(y_i \mathbb{E}\left(\frac{1}{z_i}\right) - \theta - \sum_{l=1, l\neq j}^{k} x_{i,l} \mu_{q_j} \mathbb{E}\left(\frac{1}{z_i}\right) \right) \right), \tag{2.35}$$

where $\mathbb{E}(1/\sigma) = \frac{m_q}{n_q}$ with m_q and n_q given in (2.44) and (2.45). The density function of $q(z_i)$ follows the generalized inverse Gaussian

$$GIG(\frac{1}{2}, \tilde{a}_{q_i}, \tilde{b}_{q_i}), \tag{2.36}$$

with

$$a_{q_i} = \left(\frac{\theta^2}{\tau^2} + 2\right) \mathbb{E}\left(\frac{1}{\sigma}\right), \tag{2.37}$$

and

$$b_{q_i} = \frac{(y_i^2 - 2y_i \mathbf{x}_i' \boldsymbol{\mu}_q + \mathbf{x}_i' \mathbb{E}(\boldsymbol{\beta}_p \boldsymbol{\beta}_p') \mathbf{x}_i) \mathbb{E}(\frac{1}{\sigma})}{\tau^2}. \tag{2.38}$$

The density functions of $q(s_j)$ and $q(\eta^2)$ follow the generalized inverse Gaussian

$$GIG(\frac{1}{2}, \eta_{q_j}^2, \beta_{q_j}^2), \tag{2.39}$$

and the Gamma distribution

$$\text{Gamma}\left(k + c, \sum_{j=1}^{k} \frac{\mathbb{E}(s_j)}{2} + d\right), \tag{2.40}$$

respectively, with

$$\eta_{q_j}^2 = \mathbb{E}(\eta^2) = \frac{k + c}{\sum_{j=1}^{k} \frac{\mathbb{E}(s_j)}{2} + d}, \tag{2.41}$$

and

$$\beta_{q_j}^2 = \mathbb{E}(\beta_j^2) = \mu_{q_j}^2 + \omega_{q_j}^2. \tag{2.42}$$

The density function of $q(\sigma)$ follows the inverse Gamma distribution

$$IG(m_q, n_q), \tag{2.43}$$

with

$$m_q = 3n + m_0, \tag{2.44}$$

and

$$n_q = n_0 + \sum_{i=1}^{n} \left\{ \left(1 + \frac{\theta^2}{2\tau^2}\right) \mathbb{E}(z_i) + \frac{\mathbb{E}(y_i - \mathbf{x_i} \boldsymbol{\beta}_p)^2}{2\tau^2} \mathbb{E}\left(\frac{1}{z_i}\right) - \frac{\mathbb{E}(y_i - \mathbf{x}_i \boldsymbol{\beta})}{\tau^2} \theta \right\}. \tag{2.45}$$

Algorithm 3:

Step1 Initialize parameters in the prior distribution, including m_0, n_0, c and d.

Step2 Initialize mean $\boldsymbol{\mu}_q$ and covariance matrix Σ_q.

Step3 Update parameters in $q(\sigma)$, using $q(\sigma) \sim IG(m_q, n_q)$.

Step4 Update parameters in $q(s_j)$, using $q(s_j) \sim GIG(\frac{1}{2}, \eta_{q_j}^2, \beta_{q_j}^2)$.

Step5 Update parameters in $q(\eta^2)$, using $q(\eta^2) \sim Gamma(k + c, \sum_{j=1}^{k} \frac{\mathbb{E}(s_j)}{2} + d)$.

Step6 Update parameters in $q(z)$, using $q(z_i) \sim GIG(\frac{1}{2}, \tilde{a}_{q_i}, \tilde{b}_{q_i})$.

Step7 Update parameters in $q(\boldsymbol{\beta}_p)$, using $q(\beta_j) \sim N(\mu_{q_j}, \omega_{q_j})$.

Step8 Update the lower bound l, if the difference between two consecutive l is bigger than the tolerance level specified, repeat Steps $3 \sim 8$.

---

3. Simulation studies

We compare the variational inference with the Gibbs sampling method in terms of accuracy and speed using simulated data. CPU time is used to measure the speed of different algorithms and predictive mean squared error (MSE) is used to measure the accuracy. Assuming independent and identically (i.i.d) distributed errors, we conduct the simulation using the following models.

1. Sparse case with Gaussian noise: $\beta_1 = (3, 1.5, 0, 0, 2, 0, 0, 0), \epsilon_i \sim N(0, 0.6^2)$.

2. Dense case with Gaussian noise:$\beta_2 = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85), \epsilon_i \sim N(0, 0.6^2)$.

3. Very sparse case with Gaussian noise: $\beta_3 = (2, 4, \underbrace{0, \ldots, 0}_{10}), \epsilon_i \sim N(0, 0.6^2)$.

4. High-dimensional: $\beta_4 = (\underbrace{2, \ldots, 2}_{40}, \underbrace{0, \ldots, 0}_{40}, \underbrace{3, \ldots, 3}_{40}), \ \epsilon_i \sim N(0, 0.6^2)$

For first three models, we set the sample size n equal to 1000. And the sample size for the last model is 50. In model 1$\sim$3 we run the regression using variational inference without the lasso penalty (**Algorithm 2**). And variational inference with the lasso penalty is applied in the high-dimensional case. Gibbs sampling are applied in all four cases for comparison. Fig. 3.1 and Fig. 3.2 show the CPU time and predictive MSE using variational inference and Gibbs sampling under different quantiles, respectively.

The Gibbs sampling on qunatile regression is conducted using the `bayesQR` function in `R`

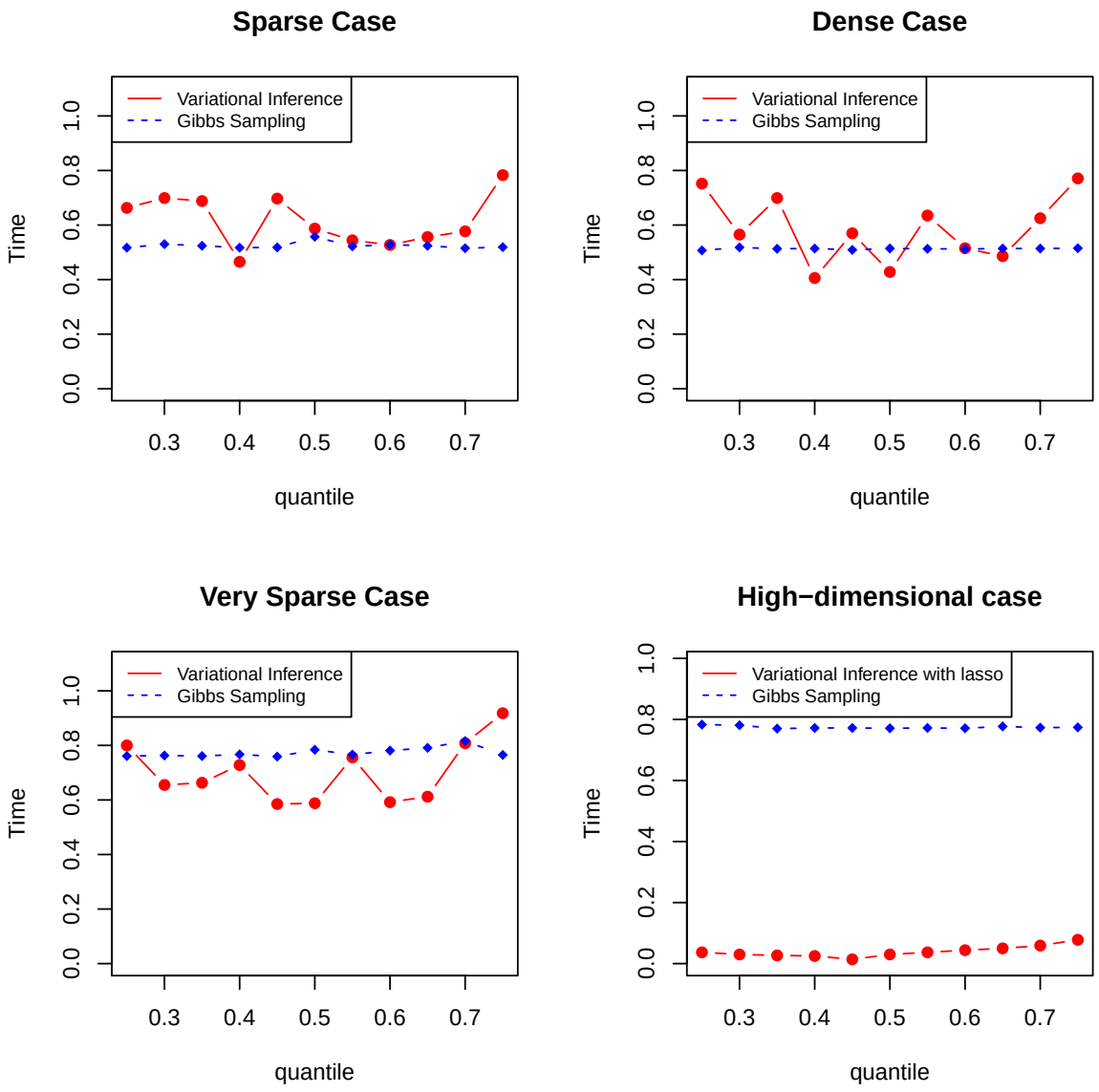

Figure 3.1. CPU time at different quantiles

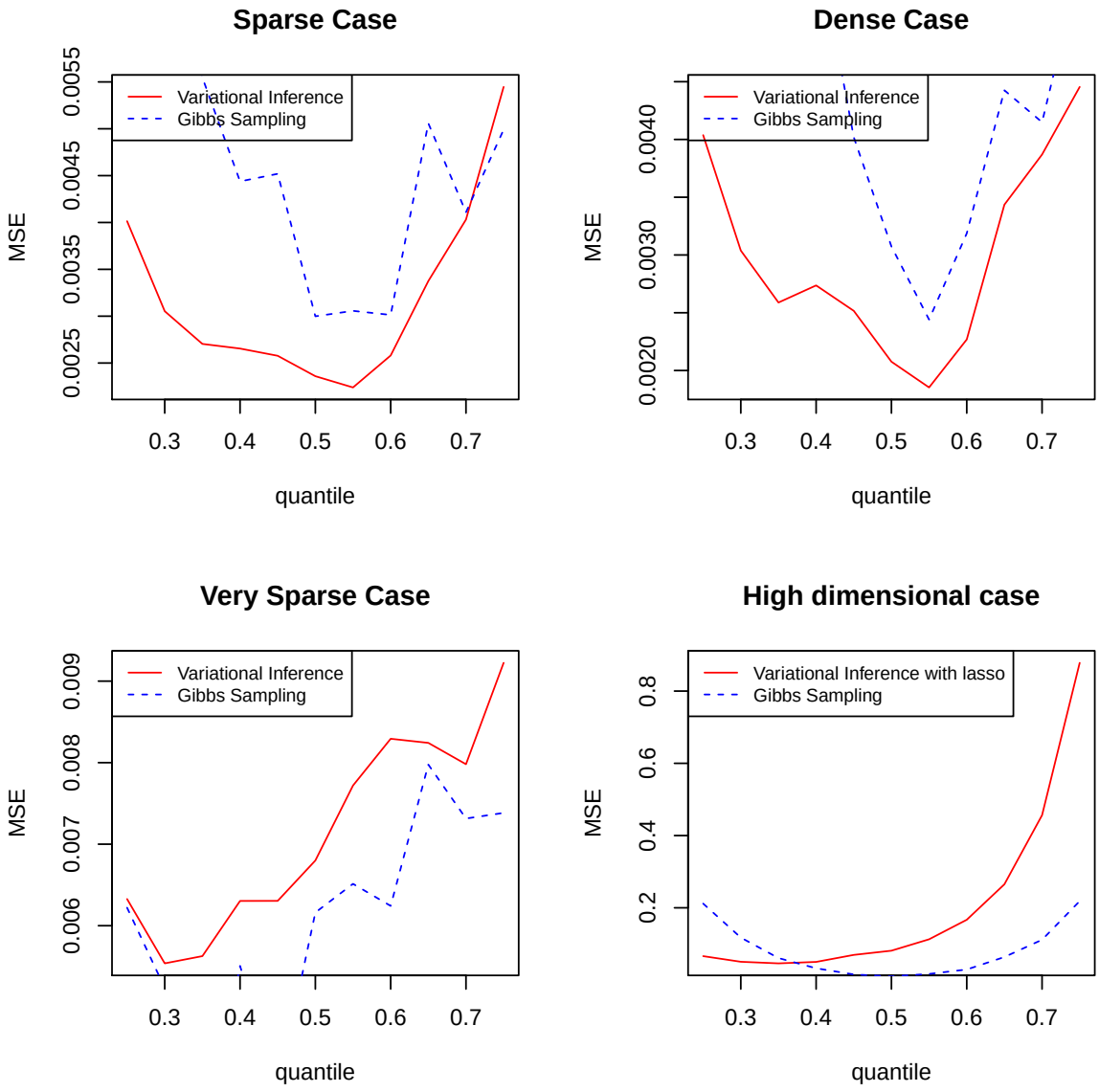

Figure 3.2. Predictive MSE at different quantiles

package `bayesQR` (Benoit and den Poel, 2017). In the sparse and dense case, variational inference has a comparable speed compared with Gibbs sampling, but variational inference maintains a lower MSE. Variational inference tends to need more time for extreme quantiles. In the very sparse case and the case when predictor is more than sample size, variational inference spends less time than quantile regression but has a slightly larger MSE. Variational inference could be used as a faster alternative to Gibbs sampling when the dimension of the predictor is high, while provides a comparable accuracy in terms of MSE. We also compare the number of iterations needed to converge using quantile regression with the lasso penalty and Gibbs sampling under quantile $p = \{0.25, 0.5, 0.75\}$. The results are shown in Fig. 3.3 and Fig. 3.4. It shows that the two methods take almost the same number of iterations to converge. However, turning points of variational inference usually occur before that in Gibbs sampling, which indicates that the declining of MSE in variational inference is usually faster in the first few iterations.

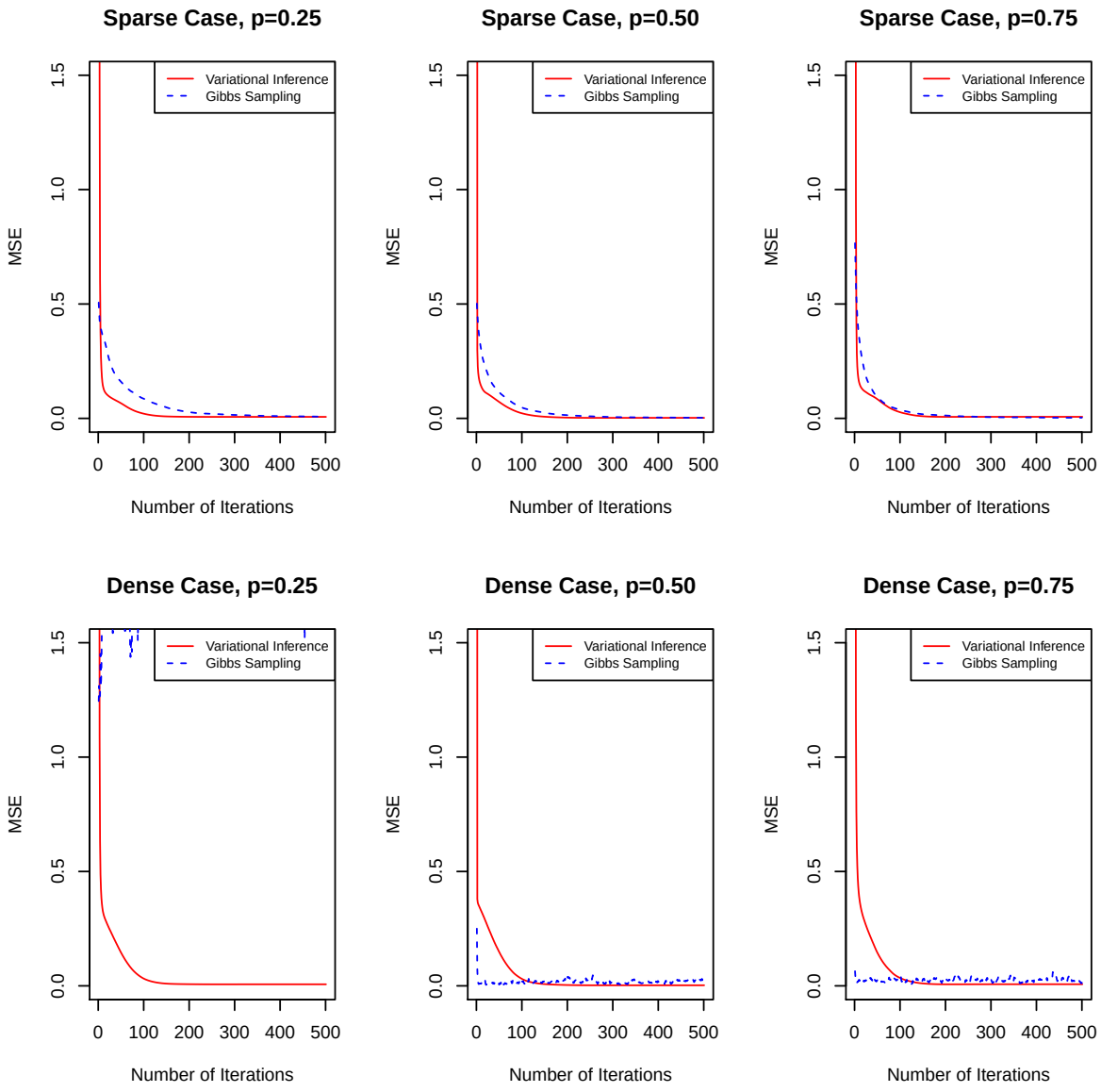

Figure 3.3. Iteration trajectories of variational inference and Gibbs sampling

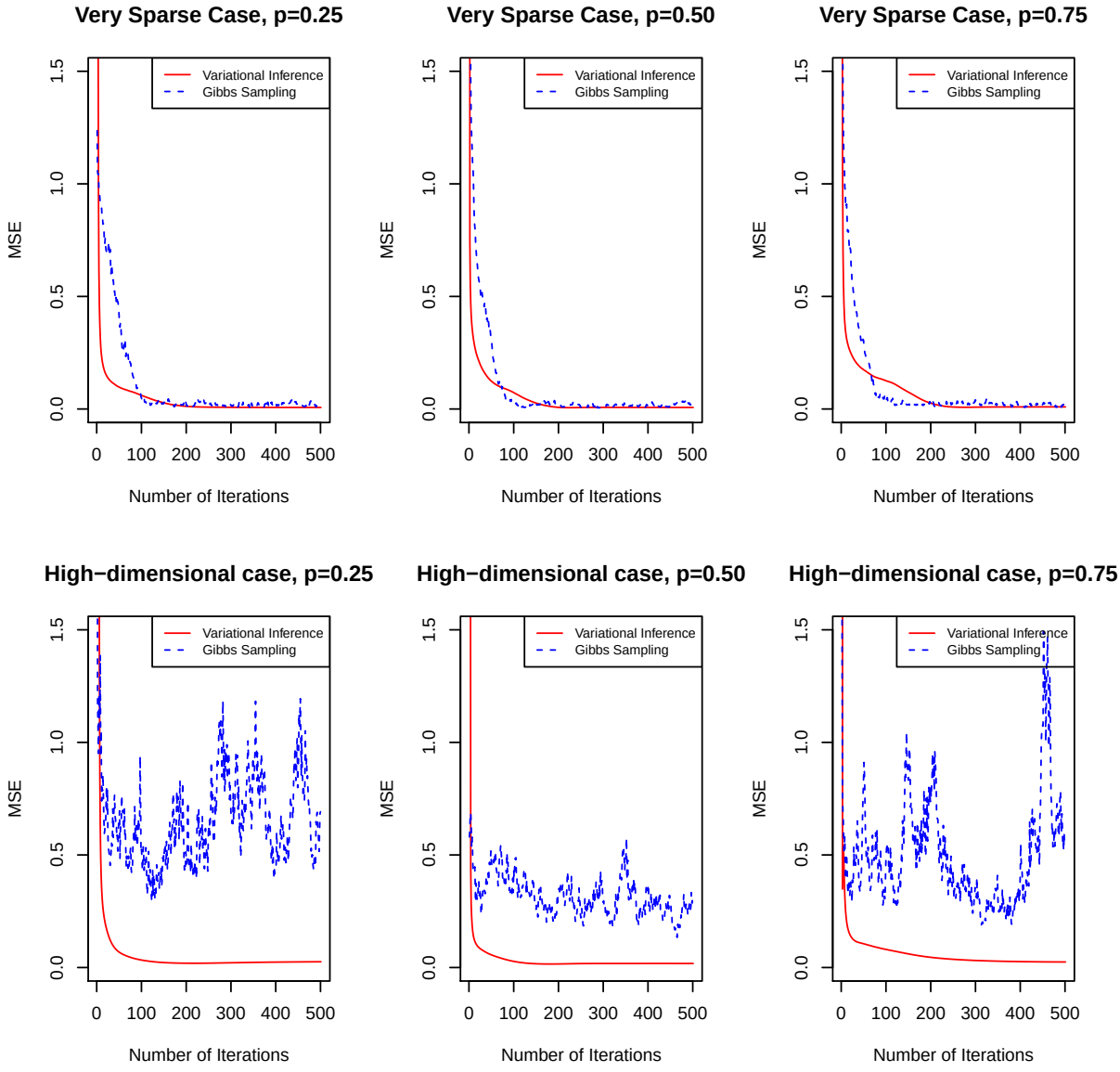

Figure 3.4. Iteration trajectories of variational inference and Gibbs sampling

4. Conclusions

This thesis derive the variational inference algorithm for quantile regression with and without the lasso regularization. Simulated studies show that, in comparison with Gibbs sampling, variational inference has a faster MSE declining within few iterations. Usually variational inference could maintain a comparable accuracy with Gibbs sampling. In very sparse data sets and the case when predictor is more than sample size, variational inference could usually perform better without sacrificing significant accuracy.

REFERENCES

S. Abeywardana and F. Ramos. Variational inference for nonparametric bayesian quantile regression. pages 1686–1692, 2015.

G. Bassett and H.-L. Chen. Portfolio style: Return-based attribution using quantile regression. *Empirical Economics*, 26:293–305, 2001.

D. Benoit and D. V. den Poel. bayesqr: A Bayesian approach to quantile regression. *Journal of Statistical Software*, 76(7):1–32, 2017.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv e-prints*, 2016.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

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

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

H. Kozumi and G. Kobayashi. Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11):1565–1578, 2011.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951.

Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556, 09 2010.

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

G. R. Pandey and V.-T.-V. Nguyen. A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225:92–101, 1999.

C. Reed and K. Yu. A partially collapsed Gibbs sampler for Bayesian quantile regression. 01 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

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

T. van Erven and P. Harremos. Renyi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, pages 1–15, 2018.

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

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *arXiv e-prints*, 2017.