Doctoral Thesis

Some New Developments for Quantile Regression

Author:
Xi Liu

Supervisor:
Prof. Keming Yu

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

College of Engineering Design and Physical Sciences
Department of Mathematics

April 2018

Declaration of Authorship

I, Xi Liu, declare that this thesis titled, 'Some New Developments for Quantile Regression' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ____

Date: ____

Abstract

Quantile regression (QR) (Koenker and Bassett, 1978), as a comprehensive extension to standard mean regression, has been steadily promoted from both theoretical and applied aspects. Bayesian quantile regression (BQR), which deals with unknown parameter estimation and model uncertainty, is a newly proposed tool of QR. This thesis aims to make some novel contributions to the following three issues related to QR. First, whereas QR for continuous responses has received much attention in literatures, QR for discrete responses has received far less attention. Second, conventional QR methods often show that QR curves crossing lead to invalid distributions for the response. In particular, given a set of covariates, it may turn out, for example, that the predicted 95th percentile of the response is smaller than the 90th percentile for some values of the covariates. Third, mean-based clustering methods are widely developed, but need improvements to deal with clustering extreme-type, heavy tailed-type or outliers problems.

This thesis focuses on methods developed over these three challenges: modelling quantile regression with discrete responses, ensuring non-crossing quantile curves for any given sample and modelling tails for collinear data with outliers. The main contributions are listed as below:

- The first challenge is studied in Chapter 2, in which a general method for Bayesian inference of regression models beyond the mean with discrete responses is developed. In particular, this method is developed for both Bayesian quantile regression and Bayesian expectile regression. This method provides a direct Bayesian approach to these regression models with a simple and intuitive interpretation of the regression results. The posterior distribution under this approach is shown to not only be coherent to the response variable, irrespective of its true distribution, but also proper in relation to improper priors for unknown model parameters.

- Chapter 3 investigates a new kernel-weighted likelihood smoothing quantile regression method. The likelihood is based on a normal scale-mixture representation of an asymmetric Laplace distribution (ALD). This approach benefits of the same good design adaptation just as the local quantile regression (Spokoiny et al., 2014) does and ensures non-crossing quantile curves for any given sample.

- In Chapter 4, we introduce an asymmetric Laplace distribution to model the response variable using profile regression, a Bayesian non-parametric model for clustering responses and covariates simultaneously. This development allows us to model more accurately for clusters which are asymmetric and predict more accurately for extreme values of the response variable and/or outliers.

In addition to the three major aforementioned challenges, this thesis also addresses other important issues such as smoothing extreme quantile curves and avoiding insensitive to heteroscedastic errors as well as outliers in the response variable. The performances of all the three developments are evaluated via both simulation studies and real data analysis.

Acknowledgments

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Keming Yu, not only for his tremendous academic support, but also for his patience, motivation, and continuous encouragement throughout my PhD. His guidance helped me throughout my research and writing of papers. I have been extremely lucky to have a supervisor whom has shown so much consideration towards my work, and whom responded to my questions and queries so promptly.

Besides my supervisor, I would like to thank the staff of the Statistics Group at Brunel University London for their useful suggestions and enthusiastic support. I would specially like to thank my colleagues Alina Peluso, Linda Huang and Hadeel Kalktawi, my co-supervisor Dr. Silvia Liverani, for their brilliant comments and the stimulating discussions. Without their precious support, it would not be possible to conduct this research.

Also, I am grateful towards my parents and my family, who gave unconditional support and continuous encouragement for my PhD. A special thanks goes to my beloved husband Dr. Rui Li, who played the role of a mentor and soul mate for my PhD. I really appreciate the efforts my family made on my behalf.

Finally, I would like to express my thanks to the Department of Mathematics, which always conducts many valuable seminars and provides financial support to us for international academic conferences. This kind of experience has helped broaden my insight and obtained a more profound understanding of my research.

Abbreviations and Acronyms

Although most of the abbreviations are explained when they are used for the first time in the text, we also list them here.

AIC	Akaike information criterion
ALD	Asymmetric Laplace distribution
AND	Asymmetric normal distribution
BMI	Body mass index
BQR	Bayesian quantile regression
c.d.f.	Cumulative distribution function
DALD	Discrete asymmetric Laplace distribution
DAND	Discrete asymmetric normal distribution
DP	Dirichlet process
DPMM	Dirichlet process mixture model
IG	Inverse Gamma distribution
LD	Laplace distribution
LoS	Length of Stay
MAE	Mean absolute error
MCMC	Markov chain Monte Carlo
MH	Metropolis Hasting
MLE	Maximum likelihood estimation
MQF	Multilevel quantile function
OLS	Ordinary least squares
p.d.f.	Probability density function
p.m.f.	Probability mass function
QR	Quantile regression
RMSE	Root mean square error

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Quantile regression (QR), proposed by Koenker and Bassett (1978), provides a broad approach to explore the relationships among different variables; see Koenker (2005), Yu et al. (2003) and Cade and Noon (2003) for an overview. In contrast to mean-based regression, quantile regression relies on the entire conditional distribution of a given predictor variable. Simultaneously, quantile regression can fully depict the influence of explanatory variables on the whole distribution of explanatory variables. Furthermore, QR overcomes the limitation of only revealing the influence of the response variable on the mean of explanatory variables, which is a problem that mean-based regression faces. Hence, QR is more informative. A comparative review of statistical methods for body mass index (BMI) has been discussed by Yu et al. (2016), where BMI is treated as the response variable. If one fits quantile regression to BMI with age as one of the covariates, then Figure 1.1 displays the typical age coefficient and its 95% confidence bands against the BMI quantile τ. QR provides a more comprehensive way to illustrate the effect of age on BMI than mean-based regression. Currently, QR is a very important technique and has been profoundly recognised as a comprehensive extension to standard mean regression (Koenker, 2005).

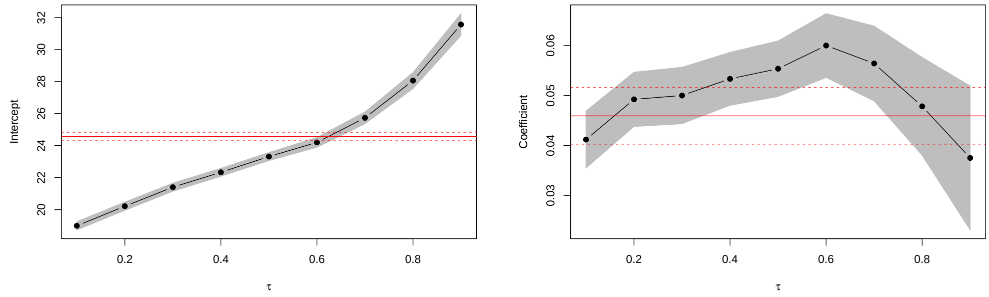


Figure 1.1: Age coefficient plotted against the BMI quantile τ with 95% confidence bands (The data (8151 observations) was obtained from UK data service: UK Data Archive Study Number 7402 - Health Survey for England, 2011: Teaching Dataset)

To further highlight the importance of QR and demonstrate its application, we provide a more elaborate way of visualising this by considering an example that superimposes several estimated conditional quantile functions on the US Health Examination Surveys (elaborated in Chapter 3). This dataset comes from the second and third health examination surveys of the USA (National Centre for US Health Examination Surveys, 1970; 1973). Taken together these provide data on the anthropometry of children between the ages of 6 years and 18 years, with about 400 to 600 children of each sex seen in each year of age (Cole, 1998). Here, along with Yu and Jones (1998), the weights and ages of 4011 US girls are analysed. In the resulting Figure 1.2, the median regression line is represented by a solid blue line, and the least squares line as a dashed red line. The other quantile regression lines appear in grey. It can be observed that the conditional median (the 0.50 quantile) and mean curves are different, and therefore the standard mean regression estimate is insufficient to estimate the relationship between weights and ages. In contrast QR models are flexible models and insensitive to heteroscedastic errors and outliers in the response variable, which are also adapted in many real-world applications (Yu et al., 2003; Koenker, 2005).

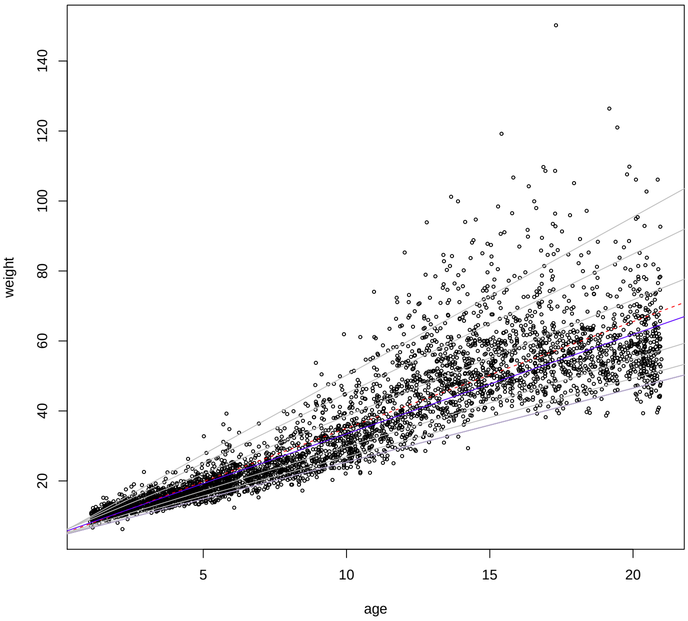


FIGURE 1.2: The relationship between weights and ages of 4011 US girls (Cole, 1998). QR curve estimates from the highest to the lowest quantiles are plotted for $\tau \in \{95\%, 90\%, 75\%, 50\%, 25\%, 10\%, 5\%\}$. The fitted standard mean regression curve is illustrated by the dashed red line.

1.1 Quantile regression

Given a sample of observations $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)$, the τth quantile regression equation can be denoted as:

$$Q_y(\tau|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}(\tau),$$

where $y_i (i = 1, ..., n)$ is the outcome of interest, $\boldsymbol{x}_i$ is a p-dimensional vector denoting the ith row of the $n \times p$ design matrix $\boldsymbol{X}$, and the unknown quantity $\boldsymbol{\beta}(\tau)$ is a vector of coefficients. Let the conditional distribution function of y be $F_y(y|\boldsymbol{x})$, and the inverse function be $Q_y(\tau|\boldsymbol{x}) = \inf\{y : F_y(y|\boldsymbol{x}) \geq \tau\}$.

Recall the classic least squares method. When y is a linear function of $\boldsymbol{x}$, the regression coefficient $\boldsymbol{\beta}$ can be optimized by

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2, \tag{1.1}$$

where $\boldsymbol{x}_i^T\boldsymbol{\beta}$ denotes the conditional mean of y. The least squares method is to estimate the parameters by minimising the residual sum of squares, which reflects the average effect of the variable $\boldsymbol{x}$ on the response variable y.

Similarly, to establish the linear function of the parameters for the τth quantile $Q_y(\tau|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}(\tau)$, quantile regression can also be used to give a solution to a simple optimization problem, such that the regression coefficient $\boldsymbol{\beta}(\tau)$ can be optimized by

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}(\tau)\right), \tag{1.2}$$

where τ is the quantile of interest, $Q_y(\tau|\boldsymbol{x})$ is the sample condition τth-quantile, and $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), ..., \beta_p(\tau))$ is the regression coefficient vector dependent on τ. $\rho_\tau(\cdot)$ is an asymmetric loss function that satisfies

$$\rho_\tau(u) = \begin{cases} \tau u & u \geq 0 \\ (\tau - 1)u & u < 0. \end{cases} \tag{1.3}$$

Equivalently, Eq.(1.3) is sometimes expressed as:

$$\rho_\tau(u) = \frac{|u| + (2\tau - 1)u}{2}.$$

Figure 1.3 shows the check function at three different quantiles, namely 0.50, 0.75 and 0.95.

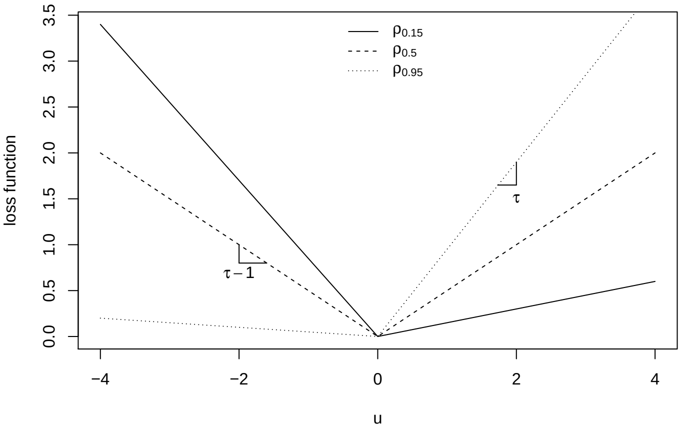


Figure 1.3: Check function in Eq.(1.3) at different τs.

Let $\tau \in (0, 1)$, define the objective function $m(\tau) = \sum_{i=1}^{n} \rho_\tau \left(y_i - Q_y(\tau|\boldsymbol{x})\right)$. Note that the objective function is a weighted sum of absolute deviations, so that the estimated vector of parameters is insensitive to the observations at the outer or edge, which is very effective for estimating the global distribution. According to Koenker and Bassett (1978), the parameter estimation of the model can be obtained by optimizing the following:

$$\hat{\boldsymbol{\beta}}(\tau) = \underset{\beta}{\operatorname{argmin}}\, m(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - Q_y(\tau|\boldsymbol{x}_i) \right), \tag{1.4}$$

Considering the following standard linear model: $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, Eq.(1.4) can be further written as:

$$\operatorname{argmin} \left(\sum_{y_i \geq \boldsymbol{x}_i^T \boldsymbol{\beta}} \tau(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) - \sum_{y_i < \boldsymbol{x}_i^T \boldsymbol{\beta}} (1-\tau)(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) \right). \tag{1.5}$$

When $\tau = 0.5$, Eq.(1.5) reduces to $\sum_{i=1}^{n} (0.5) \left| y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}(0.5) \right|$, then $\hat{y}(0.5) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}(0.5)$ is called the median regression equation, and $\hat{\boldsymbol{\beta}}(0.5)$ is called the median regression coefficient estimator.

1.1.1 Developments in quantile regression

Koenker and Bassett proposed the theory of QR in 1978. Following this, they introduced quantile regression linear hypothesis test as well as heteroscedasticity robustness test in 1982. Koenker and Machado (1999) constructed the Wald test and the likelihood ratio test to diagnose the significance of regression. These diagnostic tests proved the applicability of QR methods. Moreover, the methodologies of QR have also been improved constantly for decades. Bassett and Koenker (1986) investigated the strong consistency of the regression quantiles. Buchinsky (1995) discussed the asymptotic covariance matrix of the estimation of a quantile regression model and the truncated quantile regression by Monte Carlo simulation and elaborated the estimation procedures, including the design matrix bootstrap, error bootstrap, order statistics, homogeneous variance kernel and heterogeneous variance kernel. Koenker and Xiao (2002) solved the issue of having to address specific types of inferences when quantile regression is implemented. Kim and Muller (2000) gave a study of the asymptotic nature of two-step quantile regression. Tasche (2001) studied the unbiased properties of the minimum quantile regression. Kim and White (2003) studied the consistency and asymptotic normality of estimators for nonlinear quantile regression and proposed the hypothesis test and statistical inference of the quantile regression model. Koenker and Xiao (2006) imported the autoregressive model into quantile regression framework and proposed the quantile auto-regression (QAR) model, whereby the coefficients are dependent on quantiles.

Subsequent to the theory of quantile regression being developed, new and effective algorithms for modelling quantile regression are also constantly developing. The most widely used are: (1) Simplex method (Koenker and Bassett, 1978; Koenker and d'Orey, 1993; Portnoy et al., 1997). Randomly pick a vertex and then search along the boundary of the feasible solution to the polygon until the best point is found. The characteristics of this algorithm make it suitable for cases with small sample sizes and few variables; (2) Interior point method (Koenker and Park, 1996). It achieves optimization by going through the middle of the solid defined by the problem rather than around its surface, which further effectively solves the large-scale computing problems.

Advancements made in the aforementioned methods and technologies in modelling, computing and related software packages (such as R, Splus, SAS and other programs), have ensured the availability of quantile regression which is now widely applied in statistics and numerously in other areas (Koenker and Hallock, 2001, Yu et al., 2003, Briollais and Durrieu, 2014 and among others). For example, Machado and Mata (2005) proposed a method to decompose the changes found in the distribution of wages over a period of time into several factors contributing to those changes, which is based on the estimation of marginal wage distributions consistent with a conditional distribution estimated by quantile regression. Xu et al. (2016) developed a novel quantile auto-regression neural network model, which is able to evaluate value-at-risk in practice and achieves high prediction accuracy. Yu et al. (2016) provided a key resource and statistical library for researchers in public health and medicine to deal with obesity and body mass index data analysis, and especially elaborated on both classical and modern QR methods to improve the understanding of the complex system of intercorrelated influences on BMI.

1.1.2 Nonparametric estimation of quantile regression

Since Koenker and Bassett (1978) proposed linear quantile regression under the parametric framework, these simple linear models have been refined to account for nonparametric effects via additive models (Koenker and Hallock, 2001; Yu and Lu, 2004; Fenske et al., 2011). Further methods include local quantile regression (Yu and Jones, 1998; Spokoiny et al., 2014), single-index quantile regression (Wu et al., 2010), semiparametric quantile regression (Noh et al., 2015), nonparametric quantile regression (Horowitz and Lee, 2007; Chernozhukov et al., 2008; Li and Racine, 2008; Li et al., 2013) and quantile regression for time series (Chen et al., 2009; Xiao and Koenker, 2009).

Nonparametric quantile regression not only relaxes the assumption of linearity in the regression parameters, it also avoids the need to specify a precise functional form for the relationship between the response and regressors. In general, two types of methods are conventionally used to solve nonparametric regression models: one is global approximation and the other is local approximation. Spline is the most commonly used method for estimating the global approximation, which uses spline functions to approximate the nonparametric model and transfer it to parametric estimation (De Boor et al., 1978). The estimation result depends on the number and position of the spline nodes. Amongst local approximation methods, local constant estimation and local linear estimation are widely used, where the estimation accuracy depends on the kernel (weight) function selection and bandwidth selection.

Consider the following nonparametric model:

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{1.6}$$

where $y = (y_1, \cdots, y_n)^T$ is the response variable, $X = (x_1, \cdots, x_n)^T$ is the dependent variable, and the errors $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^T$ are assumed to be equally distributed.

1.1.2.1 Spline smoothing quantile

Polynomial splines are piecewise polynomials with the polynomial pieces joining together smoothly at a set of interior knot points (Huang et al., 2004). Hendricks and Koenker (1992), Koenker et al. (1994), He et al. (1998) and He and Ng (1999) utilized a smooth spline method to estimate nonparametric conditional quantiles. Kukush et al. (2005) also explored the true conditional quantile function based on the local polynomial approximation, and constructed the consistency, asymptotic normality and asymptotic significance interval of the estimator.

A polynomial spline of degree $M \geq 0$ with knot sequence $\xi_l, l = 1, \cdots, K$ is a piecewise-polynomial of degree M, and globally has continuous $M - 1$ derivatives for $M \geq 1$. A piecewise constant function, linear spline, quadratic spline and cubic spline corresponds to $M = 0, 1, 2, 3$, respectively.

Suppose that $g(x_i)$ in Eq.(1.6) can be approximated by some spline function:

$$g(x_i) \approx \sum_{j}^{K+M+1} \theta_j h_j(x_i),$$

where $h_j(x), j = 1, \cdots, K + M + 1$ is a basis for the spline functions with a fixed degree and knot sequence. Take a cubic spline as an example, Eq.(1.6) can be further treated as a parametric model, where the position parameter can be estimated directly via quantile regression. The quantile regression estimation for a cubic polynomial spline function can be expressed as:

$$(\hat{b}_\tau^*, \hat{\Theta}) = \underset{b_\tau^*, \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - b_i^* - \sum_{j=2}^{K+4} \theta_j h_j(x_i) \right),$$

where $\hat{b}_\tau^*$ is the τth-quantile of $\varepsilon + \theta_1$, where ε is the error in Eq.(1.6) and $\hat{\Theta} = (\hat{\theta}_2, \cdots, \hat{\theta}_{K+4})^T$.

Therefore, the conditional quantile estimation of $y|X$ is $X^T \hat{\Theta} + \hat{b}_\tau^*$, with

$$X^T = \begin{bmatrix} h_2(x_1) & \dots & h_{K+4}(x_1) \\ \vdots & \ddots & \vdots \\ h_2(x_n) & \dots & h_{K+4}(x_n) \end{bmatrix}.$$

1.1.2.2 Kernel smoothing quantile

In addition to the spline function method, the local constant estimation (kernel estimation) and local linear estimation for nonparametric regression models are also widely used. For instance, Welsh (1996) and Yu and Jones (1998) considered the local polynomial estimation of quantile regression and established conditions under which these estimators achieve optimal rates of convergence. Yu and Lu (2004) considered nonparametric additive regression estimation by kernel weighted local linear fitting to cope with multivariate covariates. Wu et al. (2010) further proposed the minimised average loss estimation for single-index quantile regression to cope with high-dimensional nonparametric estimation problems involving multivariate covariates.

Recall Eq.(1.6), kernel quantile regression treats the conditional quantile of the explanatory variable as $g(x) = Q_\tau(Y|X)$, hence nonparametric regression quantiles obtained by inverting a kernel estimator of $g(x)$ can be expressed as follows:

$$\hat{g}(x) = \operatorname{argmin} \sum_{i=1}^{n} \rho_\tau(y_i - g(x))K_h(x_i - x),$$

where $h > 0$ is the bandwidth, and $K_h(x_i - x) = K(\frac{x_i - x}{h})$ denotes a kernel function that satisfies $K(u) > 0, \int K(u)du = 1$ and $\int K(u)udu = 0$.

Kernel estimation enjoys consistency and asymptotic normality. However, due to the boundary effect, that is, the velocity converges to the actual function at the boundary is slower than the convergence rate at the interior point, the local constant estimate is not the optimal estimate. As it was pointed out in Yu and Jones (1997), local linear estimation is thought to be superior to kernel regression. In general, local linear estimation removes a bias term from the kernel estimator, therefore it behaves better near the boundary of the explanatory variables and reduces the estimated error everywhere. The idea of the local linear fit is to approximate the unknown τth quantile $Q_\tau(x)$ by a linear function $Q_\tau(z) = Q_\tau(x) + Q'_\tau(x)(z - x) = a + b(z - x)$, for z in a neighbourhood of x. The local linear estimator of $g(x)$ can be further derived as follows:

$$(\hat{g}(x), \hat{g}'(x)) = (\hat{a}, \hat{b}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau(y_i - a - b(x_i - x))K_h(x_i - x).$$

1.2 Bayesian quantile regression

The application of traditional quantile regression has been widely acknowledged, yet there are many issues that remain and can be tackled to a certain degree using Bayesian inference. The key to estimating QR when using Bayesian inference is to let the error terms follow an asymmetric Laplace distribution (ALD). Based on this, the maximum likelihood function of the parameter can be constructed, and the transformation from the prior distribution of parameters to the posterior distribution can be further derived according to Bayes' theorem.

Bayesian inference is widely used in general linear or extended models, especially for complex objective functions, and the posterior probability distribution of parameters can be obtained by MCMC simulation. However, in the field of QR, there exists very few pieces of literature based on Bayesian inference, only Fatti et al. (1998) before the year 2000 has made a simple attempt to apply Bayesian quantile regression (BQR). Yu and Moyeed (2001) first proposed that quantile regression can be incorporated into the Bayesian inference framework by using ALD.

1.2.1 Asymmetric Laplace distribution

The density function of the Laplace distribution (LD) is

$$f(x) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\sqrt{2}\frac{|x-\mu|}{\sigma}\right),$$

where $-\infty < \mu < \infty$, $\sigma > 0$, μ and σ^2 are the mean and variance, respectively. This distribution is also known as type-1 LD, while the normal distribution can be treated as type-2 LD. The classical ordinary least squares (OLS) regression is based on the type-2 LD with the goal of minimising the sum of squared deviations. Compared to the normal distribution, type-1 LD presents features which are fat-tailed and leptokurtic.

ALD is proposed on the basis of LD. For a random variable X, if it follows an asymmetric Laplace distribution, its density function can be written as:

$$f(x;\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{\frac{x-\mu}{\sigma}\left[\tau - \mathrm{I}(x \leq \mu)\right]\right\}, \qquad x \in (-\infty, +\infty), \tag{1.7}$$

where $0 < \tau < 1$ is the skew parameter, $\sigma > 0$ is the scale parameter, and $-\infty < \mu < \infty$ is the location parameter. The corresponding distribution function and quantile function are given respectively by:

$$F(x;\mu,\sigma,\tau) = \begin{cases} \tau \exp\left\{\dfrac{1-\tau}{\sigma}(x-\mu)\right\} & x \leq \mu, \\ 1-(1-\tau)\exp\left\{-\dfrac{\tau}{\sigma}(x-\mu)\right\} & x > \mu, \end{cases} \tag{1.8}$$

and

$$F^{-1}(x;\mu,\sigma,\tau) = \begin{cases} \mu + \dfrac{\sigma}{1-\tau}\log\left(\dfrac{x}{\tau}\right) & 0 \leq x < \tau, \\ \mu - \dfrac{\sigma}{\tau}\log\left(\dfrac{1-x}{1-\tau}\right) & \tau < x \leq 1. \end{cases} \tag{1.9}$$

An important property of this quantile function is that the τth-quantile of random variable x is equal to the location parameter μ, $F^{-1}(x;\mu,\sigma,\tau)|_{x=\tau} = \mu$, with the basis that ALD is used as the error distribution of the quantile regression model.

Similar to a normal distribution, any ALD can be derived from a standard ALD. If $X \sim ALD(0,1,\tau)$, then $Y = \mu + \sigma X \sim ALD(\mu,\sigma,\tau)$. The normalised density function can be written as:

$$f(x;0,1,\tau) = \tau(1-\tau)\exp\left\{-\rho_\tau(x)\right\}, \quad x \in (-\infty,\infty), \tag{1.10}$$

where $0 < \tau < 1$ and $\rho_\tau(u)$ is as defined in Eq.(1.3). When $\tau = 0.5$, Eq.(1.10) reduces to the density function of a standard symmetric Laplace distribution. For all other values of τ, the density in Eq.(1.10) is asymmetric. The mean of X is $(1-2\tau)/\tau(1-\tau)$ and it is positive only when $\tau > 0.5$. The variance given by $(1-2\tau+2\tau^2)/\tau^2(1-\tau)^2$ increases quite rapidly as τ approaches 0 or 1. The skewness of standard asymmetric Laplace density functions varies with τ, i.e. when $\tau = 0.05$, ALD focuses more on the right tail; when $\tau = 0.5$, the distribution is symmetrical (see Figure 1.4).

Compared with normal distributions of exponential quadratic form, linear exponential typed ALD always presents features which are fat-tailed and leptokurtic. In addition, Yu and Zhang (2005) pointed out that ALD can be obtained by a linear combination of two independent exponential random variables, such that if ξ and η follow standard exponential distribution independently, then $\frac{\xi}{\tau} - \frac{\eta}{1-\tau} \sim ALD(0,1,\tau)$. Moreover, a mixture representation of ALD based on exponential and normal distributions was proposed in Reed and Yu (2009) and Kozumi and Kobayashi (2011), which stated that if a random variable ε follows the ALD with density in Eq.(1.10), then ε can be represented as a location-scale mixture of normals given by:

$$\varepsilon = \mu z + \delta\sqrt{z}e,$$

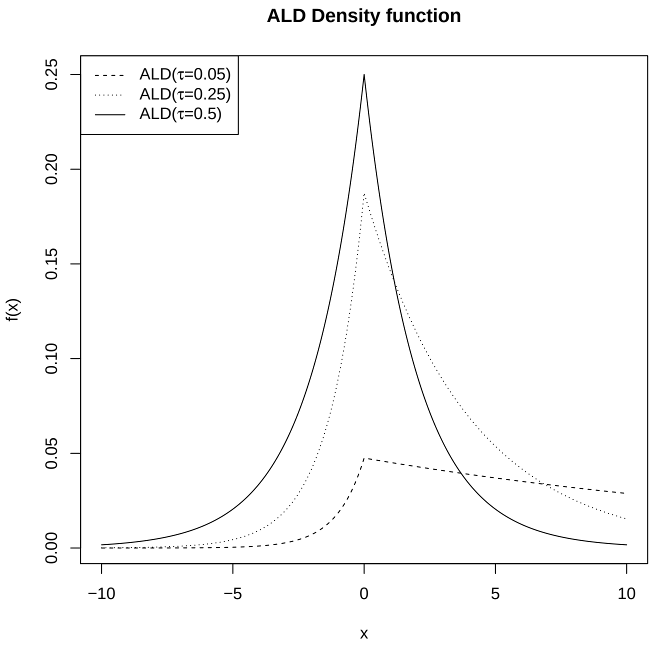


FIGURE 1.4: Density functions of asymmetric Laplace distribution at different τs.

where z is a standard exponential variable and e is a standard normal variable. $\mu = \frac{1-2\tau}{\tau(1-\tau)}$ and $\delta^2 = \frac{2}{\tau(1-\tau)}$.

1.2.2 Bayesian quantile regression based on asymmetric Laplace distribution

Yu and Moyeed (2001) and Yu and Stander (2007) set the error terms of QR to follow the ALD, and found that the maximum of the likelihood function is equivalent to the minimum of the loss function of QR. Therefore, the traditional quantile regression optimization method can be replaced by Bayesian inference based on the ALD-likelihood. Yu and Moyeed (2001) also verified the validity of the proposed method by MCMC simulation experiments and the analysis of two case studies. Yu and Moyeed (2001), with the help of ALD, have opened the door for parameter estimation of Bayesian quantile regression (BQR), which has received increasing attention.

Recall that quantile regression is a minimization of the loss function

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - Q_y\left(\tau | \boldsymbol{x}_i\right)\right), \tag{1.11}$$

where $Q_y\left(\tau|\boldsymbol{x}\right)$ is the τth quantile given $\boldsymbol{x}$ and $\rho_\tau(u) = u(\tau - \mathrm{I}(u < 0))$ is the loss function, with $\mathrm{I}(\cdot)$ is an indicator function.

Let $\varepsilon \sim ALD(0, \sigma, \tau)$, then we have $y \sim ALD\left(Q_y\left(\tau|\boldsymbol{x}\right), \sigma, \tau\right)$, where $0 < \tau < 1$ is the skew parameter, σ is the scale parameter. Then, the density function of y can be written as:

$$\begin{aligned} & f\left(y; Q_y\left(\tau|\boldsymbol{x}\right), \sigma, \tau\right) \\ & = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{\rho_\tau\left(y - Q_y\left(\tau|\boldsymbol{x}_i\right)\right)}{\sigma}\right\}. \end{aligned} \tag{1.12}$$

Therefore the likelihood function is derived as:

$$\begin{aligned} & L\left(y; Q_y\left(\tau|\boldsymbol{x}\right), \sigma, \tau\right) \\ & = \frac{\tau^n(1-\tau)^n}{\sigma^n} \exp\left\{-\frac{1}{\sigma}\sum_{i=1}^{n} \rho_\tau\left(y_i - Q_y\left(\tau|\boldsymbol{x}_i\right)\right)\right\}. \end{aligned} \tag{1.13}$$

According to Bayes' theorem, the posterior density is proportional to the product of the prior density and the sample likelihood function. Hence, the joint posterior density of parameters can be presented as:

$$\pi(\boldsymbol{\beta}, \sigma|y) \propto L\left(y; Q_y\left(\tau|\boldsymbol{x}\right), \sigma, \tau\right) f(\boldsymbol{\beta})\phi(\sigma), \tag{1.14}$$

where $f(\boldsymbol{\beta})$ and $\phi(\sigma)$ are the prior densities of the coefficients and the scale parameters.

The maximization of the likelihood function in Eq.(1.13) is equivalent to the minimization of the loss function in Eq.(1.11) for a given τ, so the parameter estimates for the quantile regression can be optimized by Eq.(1.13). Thus, by using ALD, Bayesian parametric estimation for quantile regression can be easily implemented.

By basing Bayesian inference on MCMC, the sampling distributions of quantile regression parameters and convergence of the sampling test can be obtained effectively. Since the likelihood function in Eq.(1.13) is continuous but not derivable, there is no analytic solution for the parameter derivation. In this case, the MCMC simulation (detailed in Section 1.2.4) can be effectively adopted to obtain the posterior distribution of the parameters.

Yu and Moyeed (2001) derived that if the prior distribution of the parameters is evenly distributed, even though the choice is improper, the resulting joint posterior distribution is still proper. The prior distribution in BQR can therefore be set as a non-informative prior, such as a uniform distribution or normal distribution with large variance. However, Alhamzawi et al. (2011) argued that the distributions of the parameters at the high or low quantiles may be different in practice, and the distributions of the parameters at various quantiles may therefore differ. Prior distributions for different quantiles should be set accordingly using the Power Prior method. Sriram et al. (2013) provided an asymptotic justification to the claim that the use of ALD is satisfactory even if it is not the true underlying distribution in Yu and Moyeed (2001), by establishing posterior consistency and deriving the rate of convergence under the ALD misspecification.

Yu et al. (2005) found that BQR has the following advantages in comparison to the conventional methods: (1) The Bayesian approach has less standard errors than the Frequentist method, although these two methods have similar point estimates, because the traditional method is based on the “asymptotic property”; (2) BQR is a full posterior distribution of the dependent variable, rather than a single value, and thus a more comprehensive understanding of the estimated parameters could be proposed; (3) The hypothesis test of Frequentists is based on the parameter distribution setting, whereas BQR is based on the hypothesis test of the HPD (Highest Posterior Density) of the parameter posterior distribution, which would provide higher efficiency.

Bayesian quantile regression for continuous responses has received increasing attention from both theoretical and empirical viewpoints. The first Bayesian linear quantile regression method by Yu and Moyeed (2001) is based on an ALD for likelihood and has been implemented in SAS [1] and R (Alhamzawi, 2012; Benoit et al., 2014). This method has been extended in many different contexts and applications. Yu and Stander (2007) developed Bayesian inference Tobit quantile regression. Geraci and Bottai (2007) extended the method to random effect quantile regression. Yuan and Yin (2010) extended the inference for longitudinal data. Li et al. (2010) discussed the prediction accuracy of Bayesian quantile regression via regularization. Reed and Yu (2009) and Kozumi and Kobayashi (2011) proposed a Gibbs sampling algorithm for the inference. Gerlach et al. (2011) applied the method for financial Value-at-Risk analysis. Lee and Neocleous (2010) combined the jittering approach and ALD inference for count data, and applied in

[1] http://support.sas.com/rnd/app/examples/stat/BayesQuantile/quantile.htm

the field of environmental epidemiology. Lum and Gelfand (2012) extended the method to spatial quantile regression, and among others. See a recent review by Yang et al. (2016). Alternatively, Reich et al. (2010) applied a Bayesian infinite mixture of Gaussian densities for quantile of interest. Yang and He (2012) used the empirical likelihood for Bayesian quantile regression. But all these methods use likelihood functions from continuous responses.

1.2.3 Semi and nonparametric estimation of Bayesian quantile regression

Bayesian quantile regression uses a parametric polynomial quantile regression function, which performs better than parameter-based quantile regression in parameter estimation and statistical inferences. However, there are drawbacks to this approach: it is susceptible to outliers; the degree of the polynomial needs to be determined in advance; there may be overfitting and misconvergence of MCMC for higher-degree polynomials, and other issues. To address this, semi and nonparametric estimation of Bayesian quantile regression has been further developed, and attention has been drawn to conduct further research focusing on the regression function and error settings.

Walker and Mallick (1999) and Kottas and Gelfand (2001) examined the semi-parametric estimation of Bayesian quantile regression for the median quantile. The median regression is parametrized, yet a nonparametric model is established for the error terms. The model can be estimated by using the Polya tree or Dirichlet process (DP). Tsionas (2003) proposed a Bayesian semi-parametric quantile regression model based on the scale mixture of normal distributions. Chamberlain and Imbens (2003) and Dunson and Taylor (2005) proposed a semi-parametric Bayesian inference for linear quantile models.

Kottas and Krnjajić (2009) argued that ALD-based Bayesian semi-parametric models are parametric and linear in quantile regression functions, although they are relatively flexible in the error distribution. They proposed an alternative Bayesian semi-parametric model so that the error terms are subject to Dirichlet Process (DP) mixture. Approaching this using the DP mixture allows the data to drive the shape of the error density and thus provides more reliable predictive inference than models based on parametric error distributions. Monte Carlo simulation results show that the Bayesian semi-parametric approach prevents misspecification of the model, therefore being more robust than when the parametric Bayesian model is based on ALD and also more reliable in statistical inferencing.

Bayesian semi-parametric regression models provide more flexibility than Bayesian parametric models, although these methods are simply based on linear quantile regression. On the other hand, the use of MCMC involved in statistical inference is relatively complex. In order to deal with the complexity of this method, Koenker et al. (1994) and Koenker and Mizera (2004) proposed a nonparametric quantile regression method by using the total variation regularization for univariate and bivariate smoothing. Yu and Lu (2004), Horowitz and Lee (2005) and Cai and Xu (2008) also proposed a nonparametric approach to quantile regression based on local polynomial fitting. Taddy and Kottas (2010) proposed a nonparametric approach based on the Dirichlet process mixture model (DPMM). This was developed on the basis of the parametric estimation methods proposed by Yu and Moyeed (2001) and Tsionas (2003), using a mixed-scale asymmetric Laplace distribution, therefore providing flexibility and enabling information on the error distribution to be captured (such as skewness, leptokurtosis, and other features).

Thompson et al. (2010) proposed Bayesian nonparametric quantile regression based on natural cubic splines to improve the Bayesian parametric quantile regression of Yu and Moyeed (2001). This method allows quantile regression curves to be fit with more flexibility. In order to set up a more flexible quantile regression, Yue and Rue (2011) proposed an additive mixed quantile regression model that allows the distribution of the dependent variable to be non-linear with the independent variables. The function also permits the addition of other conditions (random effects, time trends, seasonal changes, etc) to be taken into consideration, and to adopt MCMC for statistical inference.

1.2.4 Bayesian sampling algorithm

When the function of the posterior distribution is of a familiar form (such as the Gaussian, Gamma, or Beta distribution, etc.), the posterior distribution of relevant parameters can be easily simulated. However, generally speaking, the posterior distribution of parameters is unknown. In this case, Bayesian sampling algorithms can be used to simulate the unfamiliar (nonstandard) posterior distribution.

The commonly used Bayesian sampling algorithms are as follows: (1) Metropolis-Hastings (M-H) algorithm. Metropolis et al. (1953) first proposed the Metropolis algorithm, which has been extended by Hastings (1970) to propose the use of M-H algorithm in the MCMC process. Green (1995) proposed the reversible jump M-H to sample the parameters in different dimensional spaces. A particular form of M-H is the random walk Metropolis (RWM) algorithm, which effectively determines the accepting rate that reflects the representative sample. Another is the independent sampler M-H algorithm, which is based on the Laplace approximation, which was described in Erkanli (1994); (2) Gibbs sampling (see details in Casella and George, 1992; Smith and Roberts, 1993); (3) Reject sampling and important sampling (see details in Robert, 2004 and Givens and Hoeting, 2012).

1.2.4.1 Metropolis algorithm

Metropolis et al. (1953) first proposed the Metropolis algorithm. Suppose that we need to sample from the target probability density function $p(\theta)$, while θ satisfies $-\infty < \theta < \infty$. The Metropolis algorithm generates a sequence according to the Markov chain:

$$\theta^{(1)} \to \theta^{(2)} \to \cdots \to \theta^{(t)} \to,$$

where $\theta^{(t)}$ denotes the state of the Markov chain at t.

Let $\mathbf{Q}$ be a transition matrix for the Markov chain, a proposal distribution $q\left(\theta|\theta^{(t-1)}\right)$ denotes the transfer probability from state $\theta^{(t-1)}$ to state $\theta^{(t)}$. However, it may not satisfy the balance condition:

$$p(\theta^{(t)})q\left(\theta^{(t-1)}|\theta^{(t)}\right) \neq p(\theta^{(t-1)})q\left(\theta^{(t)}|\theta^{(t-1)}\right).$$

Therefore, one may introduce an acceptance probability α:

$$\alpha\left(\theta^{(t)}|\theta^{(t-1)}\right) = p(\theta)q\left(\theta^{(t-1)}|\theta^{(t)}\right).$$

Then it satisfies that

$$p(\theta^{(t)})q\left(\theta^{(t-1)}|\theta^{(t)}\right)\alpha\left(\theta^{(t-1)}|\theta^{(t)}\right) = p(\theta^{(t-1)})q\left(\theta^{(t)}|\theta^{(t-1)}\right)\alpha\left(\theta^{(t)}|\theta^{(t-1)}\right).$$

Now the Markov chain satisfies the detailed balance condition and $q\left(\theta^{(t-1)}|\theta^{(t)}\right)\alpha\left(\theta^{(t-1)}|\theta^{(t)}\right)$ denotes the transfer probability. Then the MCMC algorithm flow is as follows:

Algorithm 1.1

```
Step 1: Initialize time $t = 1$.
Step 2: Set $u$ and initialize state $\theta^{(t)} = u$.
Step 3: Repeat :
    3.1. Let $t = t + 1$
    3.2. Generate $\theta^{(*)}$ from a proposal distribution $q\left(\theta^{(t)}|\theta^{(t-1)}\right)$
    3.3. Calculate the acceptance probability: $\alpha = p\left(\theta^{(*)}\right) q\left(\theta^{(t-1)}|\theta^{(*)}\right)$
    3.4. Generate a random variable $a$ from an uniform distribution: $a \sim U[0,1]$
    3.5. If $a \leq \alpha$, accept $\theta^{(t)} = \theta^{(*)}$; otherwise $\theta^{(t)} = \theta^{(t-1)}$
Step 4 : Till $t = T$.
```

1.2.4.2 Metropolis-Hastings sampling

The acceptance probability of the Metropolis algorithm in 1.2.4.1 could be very small, which requires a large number of iterations to converge to the stationary distribution $p(\theta)$. Hastings (1970) extended the Metropolis algorithm to propose the use of Metropolis Hastings (M-H) algorithm in the MCMC process. M-H sampling has since become an important sampling method in the Monte-Carlo Markov chain.

The M-H algorithm takes a random value $\theta^{(1)}$ in the parameter space as a starting point. A new candidate state $\theta^{(*)}$ is then generated by using a proposal distribution $q\left(\theta|\theta^{(t-1)}\right)$ and the new value is subsequently accepted or rejected according to a certain probability. In the M-H sampling algorithm, the probability is:

$$\alpha = \min\left(1, \frac{p\left(\theta^{(*)}\right) q\left(\theta^{(t-1)}|\theta^{(*)}\right)}{p\left(\theta^{(t-1)}\right) q\left(\theta^{(*)}|\theta^{(t-1)}\right)}\right).$$

This process continues until the sampling process converges. After convergence, the sample $\theta^{(t)}$ is the sample in the target distribution $p(\theta)$.

Based on the above analysis, we can summarize the following M-H sampling algorithm flow:

Algorithm 1.2

```
Step 1: Initialize time $t = 1$.
Step 2: Set $u$ and initialize state $\theta^{(t)} = u$.
Step 3: Repeat :
    3.1. Let $t = t + 1$
    3.2. Generate $\theta^{(*)}$ from a proposal distribution $q\left(\theta^{(t)}|\theta^{(t-1)}\right)$
    3.3. Calculate the acceptance probability: $\alpha = \min\left(1, \frac{p\left(\theta^{(*)}\right)q\left(\theta^{(t-1)}|\theta^{(*)}\right)}{p\left(\theta^{(t-1)}\right)q\left(\theta^{(*)}|\theta^{(t-1)}\right)}\right)$
    3.4. Generate a random variable $a$ from an uniform distribution: $a \sim U[0,1]$
    3.5. If $a \leq \alpha$, accept $\theta^{(t)} = \theta^{(*)}$; otherwise $\theta^{(t)} = \theta^{(t-1)}$
Step 4 : Till $t = T$.
```

1.2.4.3 Gibbs sampling

Gibbs sampling is the most commonly used MCMC algorithm and also a special case of the Metropolis-Hastings algorithm, when the acceptance probability = 1. For high-dimensional data sampling, Stuart Geman and Donald Geman proposed the Gibbs sampling algorithm in 1984 (Geman and Geman, 1984). The detailed stationary condition at this time can be expressed as

$$p(x_1, y_1)p(y_2|x_1) = p(x_1, y_2)p(y_1|x_1).$$

Then the transition matrix $\mathbf{Q}$ can be represented as $p(y|x_1)$, so in the n-dimensional space the transition matrix can be defined for the probability distribution $p(x_1, x_2, \ldots, x_n)$ as follows:

$$\mathbf{Q}\left((x_i, x_{-i}) \rightarrow (\hat{x}_i, x_{-i})\right) = p(\hat{x}_i|x_{-i}).$$

If the current state is $x_1, x_2, \ldots, x_n$, the transition can only be made along the axes. The transition probability is defined by the conditional probability $p(x_i|x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ when a transition is made along the x-axis, while other transition probabilities are set equal to 0. Specifically, the Gibbs sampling algorithm flow follows:

Algorithm 1.3

```
Step 1: Initialize $x_i : i = 1, 2, \ldots, n$.
Step 2: Let $t = 0, 1, 2, \ldots,$
    2.1. $x_1^{t+1} \backsim p(x_1|x_2^t, x_3^t, \ldots, x_n^t)$
    2.2. $x_2^{t+1} \backsim p(x_2|x_1^{t+1}, x_3^t, \ldots, x_n^t)$
    2.3. $\ldots$
    2.4. $x_j^{t+1} \backsim p(x_j|x_1^{t+1}, \ldots, x_{j-1}^{t+1}, x_{j+1}^t, x_n^t)$
    2.5. $\ldots$
    2.6. $x_n^{t+1} \backsim p(x_n|x_1^{t+1}, x_2^{t+1}, \ldots, x_{n-1}^{t+1})$
```

1.3 Challenges in quantile regression methodologies

Quantile regression (QR) (Koenker and Bassett, 1978), as a comprehensive extension to standard mean regression, has been steadily promoted from both theoretical and practical aspects. However, there exists many areas in QR methods that require improvement, from which we select three recognized and intriguing challenges of QR methodologies to investigate and develop. There are modelling quantile regression with discrete responses, ensuring non-crossing quantile for any given sample, and clustering collinear data with outliers based on a quantile representation. The emerging challenges also include quantile regression in big data, which is beyond the scope of this thesis.

1.3.1 Quantile regression modelling with discrete responses

Amongst numerous areas of application, discrete observations with integer values (e.g., -2, -1, 0, 1, 2, 3, etc.) on a response are easily collected. In particular, current big data largely consists of data with discrete observations such as the number of online transactions, the number of days in hospitals, the number of votes and so on. Classic regression models for discrete responses include Poisson regression models and their variants. However, variants of the Poisson regression model and the model itself are often criticized for their inability to deal with large numbers of over-scattered zero-strain variables. The difficulty of the problem lies in the distance between the corresponding variables. Previously, the dependent variable would be conditioned using the conditional density of the Lebesgue measure, but this is no longer suitable for the discrete dependent variables. Moreover, discrete responses are generally skewed, hence the mean-based regression analysis would be insufficient for a complete analysis.

Machado and Santos Silva (2005) proposed a simple smoothing approach which allows quantile regression to be applied to count data. We can imagine that in each of the possible dependent variable values, the discrete distribution of the dependent variable is replaced by a piecewise linear interpolation. Thus, the entire discrete distribution can be represented by the constant density of the segments. Converting a set of discrete integer value data into a continuous smoothing density can be achieved by vibration. For integers $Y = \{y_i, i = 1, ..., n\}$, consider

$$\tilde{y}_i = y_i + u_i,$$

where u_i is independent and identically distributed (i.i.d.) on an uniform distribution $U[0,1]$. It is possible to show that

$$Q_{\tilde{Y}}(\tau|x) = Q_Y(\tau|x) + \frac{\tau - \sum_{y=0}^{Q_Y(\tau|x)-1} Pr(Y = y|x)}{Pr(Y = Q_Y(\tau|x)|x)}.$$

Hence, continuity is achieved by interpolating each jump in the conditional quantile function of the counts using an integrated kernel (See Machado and Santos Silva (2005) for more details).

Since Machado and Santos Silva (2005) proposed this conventionally used jittering approach, quantile regression modelling with discrete responses has been constantly developing. Bottai et al. (2010) described the use of logistic quantile regression and modelled the probability of binary outcomes with the widespread use of logistic and probit regression. Quantile regression implemented in a Bayesian setting for binary data has been proposed by Hewson and Yu (2008), which transformed the responses to pseudonormal variables. Quantile regression for count data may also be achieved via density regression as showed in Canale and Dunson (2011), but this approach may result in a global estimation of regression coefficients.

Using a Bayesian approach, Benoit and Van den Poel (2012) and Rahman (2016) combined continuous latent variables and ALD-based likelihood for Bayesian inference of ordinal regression, including binary responses. Smith et al. (2015) proposed a multilevel quantile function for modelling quantile functions of discrete responses via combining continuous latent variables with a Pareto tail.

Consider the binary response model with the frequently used form:

$$Y_i^* = \boldsymbol{X}_i^T \boldsymbol{\beta} + u_i$$
$$Y_i = 1 \ \text{ if } \ y_i^* \geq 0, \quad Y_i = 0 \ \text{ otherwise},$$

where Y_i^* is a continuous latent variable that allows the dependent binary variable Y_i to be determined. Benoit and Van den Poel (2012) placed an ALD on the latent variable Y^*. Given the data Y and the quantile of interest τ, then joint posterior density of β and Y^* is given by:

$$\pi(\boldsymbol{\beta}, Y^*|Y, \tau) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \{I(Y_i^* \geq 0)I(y_i = 1) + I(Y_i^* < 0)I(y_i = 0)\} \times \text{ALD}\left(Y_i^*; \boldsymbol{X}_i^T \boldsymbol{\beta}, 1, \tau\right).$$

This posterior is of unknown form, thus Metropolis–Hastings was used for extracting this posterior distribution (see Benoit and Van den Poel (2012) for more details). However, note that using this approach would pose difficultly in interpretation due to the fact that latent values were involved, and this method is also an indirect Bayesian approach to binary responses.

Regardless of whether Bayesian inference has been applied, quantile regression has so far been extended to discrete responses. However, none of these proposed methods provides a direct Bayesian quantile regression for discrete responses.

1.3.2 Quantile regression with non-crossing curves

As Koenker and Bassett (1978) mentioned, since quantile regression curves are estimated individually, conventional quantile regression methods often show that different quantile regression curves intersect, which leads to an invalid distribution for the response variable. In particular, it may turn out that for a given set of covariates, for example, that the predicted 95th percentile of the response is smaller than the 90th percentile, which is impossible. For example, Dette and Volgushev (2008) presented that spline estimates (Koenker et al., 1994) of quantile curves $(0.1, 0.2, \cdots, 0.9)$ for the bone mineral density data (Friedman et al., 2001) are intersected (Figure 1.5 (a), (b)).

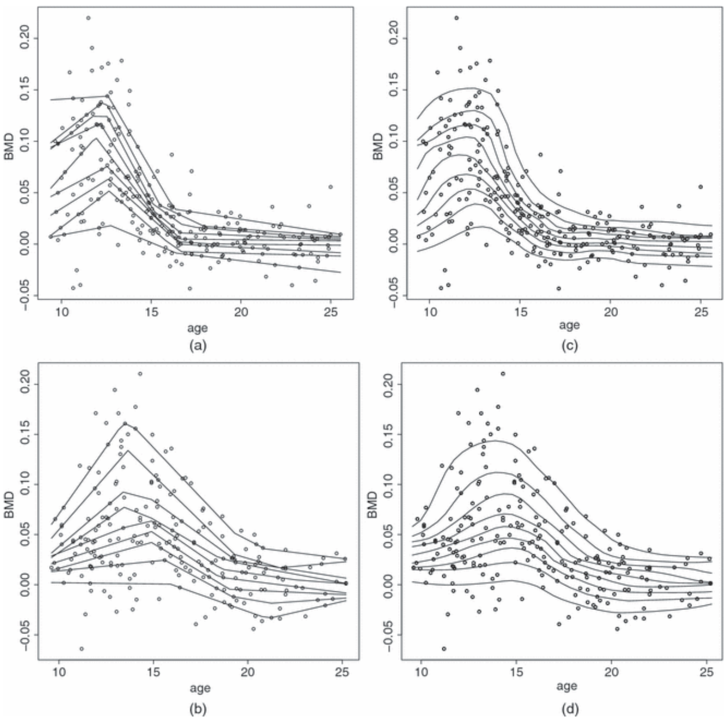


FIGURE 1.5: (a), (b) Spline (Friedman et al., 2001) and (c), (d) non-crossing estimates of quantile curves (Dette and Volgushev, 2008) for the bone mineral density data: (a), (c) females; (b), (d) males.

This problem is well known, however no simple and general solution currently exists. Koenker (1984) proposed a parallel qubit method to avoid the problem of crossing, but the hypothesis that hypoid planes are parallel is obviously too strict, therefore not representative of the actual situation. He (1997) assumes that the model is a heteroscedastic regression model so that covariates can influence the distribution of response variables by changing the position and scale of the distribution, which in many cases is not true. Wu and Liu (2009) proposed to estimate each quantile regression curve separately in order to ensure that the curve to be estimated does not intersect with the previous one. However, the estimation results are affected by the estimation order. Hall et al. (1999) and Dette and Volgushev (2008) (Figure 1.5 (c), (d)) achieved the non-crossing by estimating the conditional distribution function and then solving the method of each quantile. However, such methods can only be used to estimate conditional quantiles, instead of explicitly expressing the effects of covariates on the response variable, for example when the model is a parametric model. Bondell et al. (2010) proposed a method for estimating different quantile functions simultaneously, which estimates the non-crossed quantile functions of any sample. They also extended this method for nonparametric quantile curve estimation.

1.3.3 Clustering techniques based on quantile representation

The combination of quantile regression and widely used clustering methods are also of great interest. Clustering models identify subgroups from the data with similar characteristics. They can be useful for uncovering hidden patterns in the data, summarising data, data discovery, and learn about reoccurring patterns or underlying rules. In general, the major fundamental clustering methods can be classified into the following categories:

Table 1.1: Overview of commonly used clustering methods.

Method	Example	General Characteristics
Partitioning Methods	k-means (MacQueen et al., 1967) k-medoids (Kaufman and Rousseeuw, 2009)	- Finds mutually exclusive spherical clusters - Distance-based - May use mean or medoid (etc.) to determine cluster centres - Effective for small to medium sized data sets
Hierarchical Method	Single-link (Gower and Ross, 1969) Complete-link (Späth, 1980)	- Clustering is a hierarchical decomposition (i.e., multiple levels) - Cannot correct erroneous merges or splits - May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	DBSCAN (Ester et al., 1996) OPTICS (Ankerst et al., 1999) DENCLUE (Hinneburg et al., 1998)	- Can find arbitrarily shaped clusters - Clusters are dense regions of objects in space that are separated by low-density regions - Cluster density: each point must have a minimum number of points within its "neighbourhood" - May filter out outliers
Grid-based methods	STING (Wang et al., 1997) CLIQUE (Agrawal et al., 1998)	- Uses a multi-resolution grid data structure - Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

However, some of these traditional clustering methods require the number of clustering or mixed components to be set in advance (Parente and Silva, 2015; Reich et al., 2010), and the poor choices of these numbers will directly lead to overfitting or underfitting. Moreover, these clustering models usually model the average points of the subgroups, which can result in overlooking outliers. Modelling the response variable with an asymmetric Laplace distribution increases the accuracy of modelling clusters which are asymmetric and makes predictions for extreme predictions for extreme values/outliers of the response variable more accurate.

1.4 Thesis outline

Chapter 2 focuses on Bayesian regression beyond the mean for discrete responses. In this chapter, we propose Bayesian inference quantile regression for discrete responses by introducing a discrete version of ALD-based likelihood function. This approach not only keeps the 'local property' of quantile regression, but also enjoys the coherency and finite posterior moments of the posterior distribution. Following this, we then introduce Bayesian expectile regression for discrete responses, which proceeds by forming the likelihood function based on a discrete asymmetric normal distribution (DAND). The performance of the method is evaluated via two simulation studies and the analysis of data from two real case studies. (This chapter is a revised manuscript)

In Chapter 3, a new kernel-weighted likelihood smoothing quantile regression method is proposed from a Bayesian perspective. The likelihood is based on a normal scale-mixture representation of an asymmetric Laplace distribution (ALD). This approach enables flexibility in its design, just as the local quantile regression (Spokoiny et al., 2014) does, particularly for smoothing extreme quantile curves, and ensures non-crossing quantile curves for any given sample. An analysis for a real world application is promising. (This chapter is an accepted manuscript)

Chapter 4 illustrates how the complex relationships between the predictors can be deconstructed and analysed within a Bayesian framework. In particular, we propose a statistical approach to distinguish and interpret the complex relationship between several predictors and a response variable in the presence of 1) high correlation between the predictors and 2) the interest is in the extremes of the distribution of the response variable. The mixture modelling approach is demonstrated on both simulated and real data. (This chapter is an under review manuscript)

Finally, Chapter 5 summarizes the thesis and provides recommendations for future researches in the QR area.

Each chapter of this thesis is presented in the form of an article, thus enabling the reader to clearly understand the aims, techniques, main findings and conclusions of each chapter.

1.5 Real data

This section provides brief descriptions of the real data sets that will be used in this thesis to illustrate the applications of the proposed methods throughout the thesis.

1.5.1 Length of stay (LoS) data

Data extracted from the Worcester Heart Attack Study (Hosmer et al., 2008) is used to test the behaviour of the proposed methods in Chapter 2. Data used were collected during 13 one-year periods beginning in 1975 and extending through 2001 on all MI patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. Specifically, a subsample consisting of 500 observations on four independent variables (age, gender, BMI (Body Mass Index) and hr (Initial Heart Rate)) plus an outcome variable (LoS), taken an approximately 23 percent random sample from the cohort years 1997, 1999, and 2001, are used. The data is available in R-package *smoothHR*.

1.5.2 Lidar data

In Chapter 3, we analysed the popular Lidar data available in the R-package *SemiPar* to test the performance of the proposed kernel-weighted likelihood smoothing quantile regression method. This data has 221 observations from a light detection and ranging (LIDAR) experiment, and was originally reported by Sigrist (1994) and analysed by many authors (Ruppert et al., 2003, Royston and Sauerbrei, 2008, Spokoiny et al., 2014, among others). This data contains an explanatory variable ‘range’, which is the distance travelled before the light is reflected back to its source; and a response variable ‘logratio’, which is the logarithm of the ratio of received light from two laser sources.

1.5.3 US girls weight data

The US girls weight data from US Health Examination Surveys (Cole, 1998), is used to test the behaviour of the proposed methods in Chapter 3. This data describes the relationship between the weight and age of 4011 individuals and was previously analysed using various Bayesian and non-Bayesian approaches (Cole and Green (1992); Yu and Jones (1998); Royston and Sauerbrei (2008), among others).

1.5.4 The English longitudinal study of ageing (ELSA) analysis

To test the performance of the proposed method in Chapter 4, we collected data from the nurse visit conducted at Wave 2 of ELSA (2004-2005). A total of 7,666 people took part in this visit where biological data were collected for the first time. The data are available for download from the UK Data Service at `http://dx.doi.org/10.5255/UKDA-SN-5050-9`. Specifically, a subsample consisting of 2,859 observations on 11 independent variables plus an outcome variable are used. The outcome variable is the blood glucose levels.

Chapter 2

Bayesian Regression beyond the Mean for Discrete Responses

For decades regression models beyond the mean for continuous responses have attracted great attention in the literature. These models typically include quantile regression and expectile regression. But there is little research on these regression models for discrete responses, particularly from a Bayesian perspective. By forming the likelihood function based on suitable discrete probability mass functions, this chapter introduces a general method for Bayesian inference of these regression models with discrete responses. In contrast to latent process based Bayesian inference of binary quantile regression in the literature, this method provides a direct Bayesian approach of these discrete regression models with natural and easy interpretation of the regression coefficients. In this chapter, Bayesian quantile regression for discrete responses is first developed. Then this method is extended to Bayesian expectile regression for discrete responses. The posterior distribution under this approach is shown not only coherent irrespective of the true distribution of the response but also proper with regarding to improper priors for the unknown model parameters. The performance of the method is evaluated via extensive Monte Carlo simulation studies and one real data analysis.

2.1 Introduction

Regression models for dealing with responses following a non-normal distribution have been drawing significant attention in the literature. For example, quantile regression and expectile regression have been widely developed in the literature and increasingly applied to a greater variety of scientific questions. See, Efron (1991), Koenker (2005), Waltrup et al. (2015), Ehm et al. (2016), Delbaen et al. (2016), Ziegel (2016) among others.

Typically, quantile regression estimates various conditional quantiles of a response or dependent random variable, including the median (0.5th quantile). Putting different quantile regressions together provides a more complete description of the underlying conditional distribution of the response than a simple mean regression. This is particularly useful when the conditional distribution is asymmetric or heterogeneous or fat-tailed or truncated. Quantile regression has been widely used in statistics and numerous application areas (Koenker and Hallock, 2001, Yu et al., 2003, Briollais and Durrieu, 2014 and among others), including environment modelling (Anderson, 2008, Cannon, 2011), economics analysis (Coad and Rao, 2008, Fitzenberger et al., 2013), survival analysis (Atella et al., 2008, Peng and Huang, 2010, Portnoy, 2003), medicine (Cole and Green, 1992, Wei et al., 2006, Bottai et al., 2010), finance and insurance (Tsai, 2012, Taylor and Yu, 2016a, Sriram et al., 2016) and ultra-high dimensional data analysis (Wu and Yin, 2015, Zhang et al., 2016), among others.

Amongst these numerous application areas, discrete observations such as integer values (e.g., -2, -1, 0, 1, 2, 3, etc.) on a response are easily collected. In particular, many big data nowadays contain discrete observations such as number of online transaction, number of days of hospital stay, number of votes and so on. Classic regression models for discrete responses include logistic, Poisson and negative Binomial regression. Discrete responses are generally skewed, so the mean-based regression analysis would not be sufficient for a complete analysis. However, quantile regression for discrete responses receives far less attention than for continuous responses in the literature. A semi-parametric jittering approach for quantile regression with count has been introduced (Machado and Santos Silva 2005) but some degree of smoothness has to be artificially imposed on the approach. Quantile regression for count data may be achieved via density regression as showed in Canale and Dunson (2011) but this approach may result in a global estimation of regression coefficients. Benoit and Van den Poel (2012) and Rahman (2016) combined continuous latent variables and asymmetric Laplace distribution (ALD) likelihood for Bayesian inference of binary quantile regression, and also discussed variable selection of binary quantile regression (Benoit et al., 2013). Smith et al. (2015) proposed a multilevel quantile function for modelling quantile functions of discrete responses via combining continuous latent variables with a Pareto tail. But these approaches via latent variables would be hard to interpret. None of these used methods is a direct Bayesian quantile regression for discrete responses.

Similarly, there is little research on expectile regression for discrete responses, let alone from a Bayesian perspective (Kneib, 2013).

In this chapter we propose Bayesian inference quantile regression for discrete responses via introducing a discrete version of ALD-based likelihood function. This approach not only keeps the 'local property' of quantile regression, but also enjoys the coherency and finite posterior moments of the posterior distribution. Along this line, we then introduce Bayesian expectile regression for discrete responses, which proceed by forming the likelihood function based on a discrete asymmetric normal distribution (DAND). Section 2.2 introduces a discrete asymmetric Laplace distribution (DALD) and discusses its natural link with quantile regression for discrete responses. Section 2.3 and 2.4 detail this Bayesian approach for quantile regression and expectile regression with discrete responses, respectively. Section 2.5 illustrates the numerical performance and applications of the proposed method. Section 2.6 concludes with a brief discussion.

2.2 Discrete Asymmetric Laplace Distribution

Let Y be a real-valued random variable with its τth $(0 < \tau < 1)$ quantile μ $(-\infty < \mu < \infty)$, then it is well-known that μ could be found by minimizing the expected loss of Y with respect to the loss function (or check function) $\rho_\tau(y) = y(\tau - I(y < 0))$, or $\min_\mu E_{F_0(Y)}\rho_\tau(Y - \mu)$, where $F_0(Y)$ denotes the distribution function of Y, which is usually unknown in practice.

When Y is a continuous random variable, the inference based on the loss function $\rho_\tau(y - \mu)$ was linked to a maximum likelihood inference based on an $\text{ALD}(\mu, \tau)$ with local parameter μ and shape parameter τ:

$$f(y;\mu,\tau) = \tau(1-\tau)\exp\left\{-\rho_\tau\left(y-\mu\right)\right\}. \tag{2.1}$$

Now if Y is a discrete random variable, let Y take integer values in $\mathbb{Z}$. We first derive a discrete version of ALD or a DALD and then show that the τth quantile μ can also be estimated via this DALD.

To this end, note that the corresponding cumulative distribution function (c.d.f.) of an ALD in Eq.(2.1) can be written as:

$$F(y;\mu,\tau) = \begin{cases} 1-(1-\tau)\exp\left\{-\tau(y-\mu)\right\}, & y \geq \mu, \\ \tau\exp\left\{(1-\tau)(y-\mu)\right\}, & y < \mu. \end{cases} \tag{2.2}$$

Let $S(y;\mu,\tau)$ be the survival function of this ALD, which is given by:

$$S(y;\mu,\tau) = 1 - F(y;\mu,\tau) = \begin{cases} (1-\tau)\exp\left\{-\tau(y-\mu)\right\}, & y \geq \mu, \\ 1-\tau\exp\left\{(1-\tau)(y-\mu)\right\}, & y < \mu, \end{cases} \tag{2.3}$$

then, according to Roy (2003), the probability mass function (p.m.f.) of a DALD can be defined as:

$$\phi(y;\mu,\tau) = \begin{cases} S(y;\mu,\tau) - S(y+1;\mu,\tau), & y \in \mathbb{Z}, \\ 0, & otherwise, \end{cases} \tag{2.4}$$

with $S(y;\mu,\tau)$ in Eq.(2.3). It follows:

$$\begin{aligned} \phi(y;\mu,\tau) = \rho_\tau(-\text{sgn}(y-\mu))\left[\exp\{-\rho_\tau(\text{sgn}(y-\mu))\}-1\right]\exp\left\{-\rho_\tau(y-\mu)\right\} \\ y = \cdots, -1, 0, 1, \cdots, \end{aligned} \tag{2.5}$$

and the loss function (or check function) is

$$\rho_\tau\left(u\right) = \frac{|u| + (2\tau-1)u}{2}.$$

Remark 2.1. *One could also incorporate scale parameter σ in Eq.(2.5) to obtain*

$$\phi(y;\mu,\tau) = \rho_\tau(-sgn(y-\mu)) \left[\exp\left\{-\rho_\tau\left(sgn\left(\frac{y-\mu}{\sigma}\right)\right)\right\} - 1\right] \exp\left\{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)\right\}, \quad y = \cdots, -1, 0, 1, \cdots .$$

According to Yang et al. (2016), any fixed σ can be utilised to obtain asymptotically valid posterior inference and make the results asymptotically invariant. Here, we simply fix σ as 1.

Given a sample $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$ of the discrete response Y whose distribution $F_0(y)$ may be unknown, consider the DALD-based likelihood function for μ:

$$L(\boldsymbol{Y}|\mu) = \prod_{i=1}^{n} \left[\rho_\tau(-\text{sgn}(Y_i-\mu)) \left[\exp^{-\rho_\tau(\text{sgn}(Y_i-\mu))} - 1\right] \exp\left\{-\rho_\tau(Y_i-\mu)\right\}\right]. \tag{2.6}$$

Then we have

$$\begin{aligned} &\underset{\mu}{\text{argmax}}\, L(\boldsymbol{Y}|\mu) \\ &= \underset{\mu}{\text{argmax}} \log L(\boldsymbol{Y}|\mu) \\ &= \underset{\mu}{\text{argmax}} \left\{-\sum_{i=1}^{n} \rho_\tau(Y_i-\mu)\right\} \\ &= \underset{\mu}{\text{argmin}} \sum_{i=1}^{n} \rho_\tau\left(Y_i-\mu\right). \end{aligned}$$

This means that the estimation of the τth quantile μ of a discrete random variable Y with respect to the loss function $\rho_\tau(\cdot)$ is equivalent to maximization of the likelihood function Eq.(2.6) based on the DALD. According to Bissiri et al. (2016), a Bayesian inference of μ can be developed. That is, if $\pi(\mu)$ represents prior beliefs about the τth quantile μ, and $\boldsymbol{Y}$ are observed data from the unknown distribution $F_0(Y)$ of the discrete random variable Y, then a posterior $\pi(\mu|\boldsymbol{Y})$ which is a valid and coherent update of $\pi(\mu)$ can be obtained via the DALD-based likelihood function Eq.(2.6) and is given by:

$$\pi(\mu|\boldsymbol{Y}) \propto \pi(\mu)\, L(\boldsymbol{Y}|\mu). \tag{2.7}$$

Coherence here means if ν denotes a probability measure on the space of μ, then ν is named coherent if

$$\int \int \rho_\tau(Y-\mu)dF_0(Y)\nu(d\mu) \leq \int \int \rho_\tau(Y-\mu)dF_0(Y)\nu_1(d\mu),$$

for all other probability measure ν_1 on the space of μ in terms of expected loss of Y given by $E_{F_0(Y)}\rho_\tau(Y-\mu)$. This Coherence property aims to ensure the consistency of posterior from the proposed inference even if the 'working likelihood' in Eq.(2.3)-(A.1) is misspecified.

2.3 Bayesian Quantile Regression with Discrete Responses

Generalized linear models (GLMs) extend the linear modelling capability to scenarios that involve non-normal distributions $f(y;\mu)$ or heteroscedasticity, with $f(y;\mu)$ specified by the values of $\mu = E[Y|\boldsymbol{X} = \boldsymbol{x}]$ conditional on $\boldsymbol{x}$, including to involve a known link function g, $g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$. Specifically, GLMs also applies to the so-called 'exponential' family of models, which typically include Poisson regression with log-link function.

When we are interested in the conditional quantile $Q_Y(\tau|\boldsymbol{x})$ of a discrete response, according to Yu and Moyeed (2001), we could still cast the problem in the framework of the generalized linear model, no matter what the original distribution of the data is, by assuming that (i) $f(y;\mu)$ follows a DALD in the form of Eq.(2.5) or Eq.(A.1) and (ii) $g(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}(\tau) = Q_Y(\tau|\boldsymbol{x})$ for any $0 < \tau < 1$.

When covariate information such as a covariate vector $\boldsymbol{X}$ is available, quantile regression denoted by $Q_Y(\tau|\boldsymbol{X})$ for μ is introduced. Consider a linear regression model for $Q_Y(\tau|\boldsymbol{X})$: $Q_Y(\tau|\boldsymbol{X}) = \boldsymbol{X}^T\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the regression parameter vector, although the quantile of a discrete random variable may not be unique.

Given observations $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$ of the discrete response Y, one of the aims in regression analysis is the inference of $\boldsymbol{\beta}$. Let $\pi(\boldsymbol{\beta})$ be the prior distribution of $\boldsymbol{\beta}$, then the posterior distribution of $\boldsymbol{\beta}$, $\pi(\boldsymbol{\beta}|\boldsymbol{Y})$ is given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{Y}) \propto \pi(\boldsymbol{\beta})\, L(\boldsymbol{Y}|\boldsymbol{\beta}), \tag{2.8}$$

where the likelihood function $L(\boldsymbol{Y}|\boldsymbol{\beta})$ is given by:

$$L(\boldsymbol{Y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\rho_\tau(-\text{sgn}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})) \left[\exp^{-\rho_\tau(\text{sgn}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}))} - 1\right] \exp\left\{-\rho_\tau(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})\right\}\right].$$

The numerical computation of the posterior distribution can be carried out by the Metropolis-Hastings algorithm. That is, we first generate a candidate $\boldsymbol{\beta}^*$ according to a random walk, then accept or reject $\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}$ according to the acceptance probability $p(\boldsymbol{\beta}^*|\boldsymbol{\beta}) = \min\left(1, \frac{L(\boldsymbol{y}|\boldsymbol{\beta}^*)}{L(\boldsymbol{y}|\boldsymbol{\beta})}\right)$.

Besides the coherent property discussed in Section 2.2 for posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{Y})$, it is important to verify the existence of the posterior distribution when the prior of $\boldsymbol{\beta}$ is improper, i.e,

$$0 < E\left\{\pi(\boldsymbol{\beta}|\boldsymbol{Y})\right\} < \infty,$$

or, equivalently,

$$0 < E\left\{\pi(\boldsymbol{\beta})\, L(\boldsymbol{Y}|\boldsymbol{\beta})\right\} < \infty.$$

Moreover, it is preferable to check that the existence of posterior moments of the regression parameters is entirely unaffected by improper priors and quantile index τ (Fernández and Steel, 1998 and among others), i.e,

$$E\left[\left(\prod_{j=0}^{m} |\beta_j|^{r_j}\right) \middle| \boldsymbol{Y}\right] < \infty, \tag{2.9}$$

where r_j denotes the order of the moments of β_j.

To this end, we have the following conclusion:

Theorem 2.1. *Assume the posterior is given by Eq.(2.8) and $\pi(\boldsymbol{\beta}) \propto 1$, then all posterior moments of $\boldsymbol{\beta}$ in Eq.(2.9) exist.*

The proof of Theorem 2.1 is available in the Appendix A.

2.4 Bayesian Expectile Regression with Discrete Responses

Instead of defining the τ-th quantile of a response Y by $\operatorname{argmin}_{\mu} E\left(\rho_\tau(Y-\mu)\right)$, Newey and Powell (1987) defined the θ-th expectile of Y by

$$Expectile_\theta(Y) = \underset{\mu}{\operatorname{argmin}}\, E\left(\rho_\theta^{(E)}(Y-\mu)\right), \tag{2.10}$$

in terms of an asymmetric quadratic loss function

$$\rho_\theta^{(E)}(u) = u^2\left|\theta - I(u<0)\right|,$$

where $\theta \in (0,1)$ determines the degree of asymmetry of the loss function. Note that θ is typically not equal to τ, although there is a one-to-one relationship between τ-th quantile and θ-th expectile (Yao and Tong, 1996).

Corresponding to $\rho_\theta^{(E)}(u)$ and considering the case of continuous y, we can define an asymmetric normal distribution (AND) whose density function is given by

$$f^{(E)}\left(y;\mu,\theta\right) = k\begin{cases} \exp\left\{-\theta\left(y-\mu\right)^2\right\}, & y \geq \mu, \\ \exp\left\{\left(\theta-1\right)\left(y-\mu\right)^2\right\}, & y < \mu, \end{cases} \tag{2.11}$$

where $k = \frac{2}{\sqrt{\pi}}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\theta}+\sqrt{1-\theta}}$, μ and θ are the location parameter and shape parameter, respectively.

The corresponding c.d.f. of the AND can be written as:

$$F^{(E)}\left(y;\mu,\theta\right) = \begin{cases} k\sqrt{\frac{\pi}{\theta}}\Phi\left(\sqrt{2\theta}\left(y-\mu\right)\right) + \frac{k}{2}\left(\sqrt{\frac{\pi}{1-\theta}} - \sqrt{\frac{\pi}{\theta}}\right), & y > \mu, \\ k\sqrt{\frac{\pi}{1-\theta}}\Phi\left(\sqrt{2(1-\theta)}\left(y-\mu\right)\right), & y \leq \mu, \end{cases} \tag{2.12}$$

where $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution.

Therefore, based on the survival function $S^{(E)}(y;\mu,\tau) = 1 - F^{(E)}(y;\mu,\tau)$, we can derive the p.m.f. of the DAND by following the same procedure as in Eq.(2.4):

$$\begin{aligned}\phi^{(E)}(y;\mu,\tau) = k\sqrt{\frac{\pi}{\rho_\theta(\operatorname{sgn}(y-\mu))}}\Big[&\Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(y-\mu))}\,(y+1-\mu)\right)\\ &-\Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(y-\mu))}\,(y-\mu)\right)\Big],\quad y=\cdots,-1,0,1,\cdots.\end{aligned} \tag{2.13}$$

Now if Y is a discrete random variable with unknown distribution function $F_0(y)$, then given a sample $\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)$ of Y, the θ-th expectile of Y is estimated by the minimization of the loss function $\rho_\theta^{(E)}$ or $\operatorname{argmin}_\mu \sum_{i=1}^n \rho_\theta^{(E)}(Y_i - \mu)$. Consider the DAND-based likelihood function:

$$\begin{aligned}L^{(E)}(\boldsymbol{Y}|\mu) = \prod_{i=1}^n \Bigg[k\sqrt{\frac{\pi}{\rho_\theta(\operatorname{sgn}(Y_i-\mu))}}\Big[&\Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}\,(Y_i+1-\mu)\right)\\ &-\Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}\,(Y_i-\mu)\right)\Big]\Bigg].\end{aligned} \tag{2.14}$$

We can see that the expectile μ can also be estimated equivalently by the maximization of the likelihood function $L^{(E)}(\boldsymbol{Y}|\mu)$ in Eq.(2.14). In fact,

$$\begin{aligned}&\operatorname*{argmax}_\mu L^{(E)}(\boldsymbol{Y}|\mu)\\ &= \operatorname*{argmax}_\mu \prod_{i=1}^n \left[\Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}\,(Y_i+1-\mu)\right) - \Phi\left(\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}\,(Y_i-\mu)\right)\right]\\ &= \operatorname*{argmax}_\mu \prod_{i=1}^n \int_{\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}(Y_i-\mu)}^{\sqrt{2\rho_\theta(\operatorname{sgn}(Y_i-\mu))}(Y_i+1-\mu)} \varphi(u)du\end{aligned}$$

(According to Lagrange mean value theorem $\int_a^b \phi(u)du = \phi(\xi)(b-a)$,)

$$\begin{aligned}&= \operatorname*{argmax}_\mu \left[\exp\left\{-\rho_\theta(\operatorname{sgn}(Y_i-\mu))\sum_i^n (Y_i-\mu)^2\right\}\right]\\ &= \operatorname*{argmax}_\mu \left[-\rho_\theta(\operatorname{sgn}(Y_i-\mu))\sum_i^n (Y_i-\mu)^2\right]\\ &= \operatorname*{argmin}_\mu \sum_{i=1}^n \rho_\theta^{(E)}(Y_i-\mu),\end{aligned}$$

where $\varphi(\cdot)$ denotes the p.d.f. of the standard normal distribution.

Again, according to Bissiri et al. (2016), a Bayesian inference of the expectile μ can be developed. That is, a coherent posterior $\pi(\mu|\boldsymbol{Y})$ for the update of $\pi(\mu)$ exists and is given by $\pi(\mu|\boldsymbol{Y}) \propto \pi(\mu)\, L^{(E)}(\boldsymbol{Y}|\mu)$ with the likelihood function $L^{(E)}\,(\boldsymbol{Y}|\mu)$ in Eq.(2.14). Along the same discussion as in Section 2.3, we can prove that the posterior distribution under this Bayesian inference is proper with regarding to improper priors for regression parameter $\boldsymbol{\beta}$ in the expectile regression model $\mu = \boldsymbol{X}^T\boldsymbol{\beta}$, if covariate information $\boldsymbol{X}$ is available. The corresponding proofs are available in the Appendix A.

2.5 Numerical Analysis

In this section, we implement the proposed method to illustrate the Bayesian quantile regression for discrete responses via extensive Monte Carlo simulation studies and one real data analysis, including comparisons of the fitted model to a latent process based approach, named MQF (multilevel quantile function) (Smith et al. 2015). In all numerical analyses, we discard the first 10000 of 20000 runs in every case of MCMC outputs and then collect a sample of 10000 values from the posterior of each of the elements of $\boldsymbol{\beta}$. All numerical experiments are carried out on one Intel Core i5-3470 CPU (3.20GMHz) processor and 8 GB RAM.

2.5.1 Multilevel quantile function (MQF)

From a Bayesian point of view, Smith et al. (2015) proposed a multilevel quantile function (MQF) for modelling quantile functions of discrete responses via combining continuous latent variables with a Pareto tail.

The quantile distribution Q with a Pareto tail mentioned in Smith et al. (2015) is of the form

$$Q^*(\tau|\boldsymbol{X}) = \begin{cases} Q(\tau_L|\boldsymbol{X}) - \frac{\sigma_L}{\xi_L(\boldsymbol{X})}\left[(\tau/\tau_L)^{-\xi_L(\boldsymbol{X})} - 1\right], & \tau < \tau_L \\ Q(\tau|\boldsymbol{X}), & \tau_L \leq \tau \leq \tau_U \\ Q(\tau_U|\boldsymbol{X}) + \frac{\sigma_U}{\xi_U(X)}\left[(\frac{1-\tau}{1-\tau_U})^{-\xi_U(X)} - 1\right], & \tau > \tau_U \end{cases}$$

where $Q(\tau|\boldsymbol{X}) = \sum_{j=1}^{p} \boldsymbol{X}_j\beta_j(\tau)$; the scale parameters τ_L, τ_U are the density of the Pareto distribution evaluated at the thresholds $Q(\tau_L|\boldsymbol{X}), Q(\tau_U|\boldsymbol{X})$.

They further expanded this quantile function methodology to permit a discrete response g_i via interval-censored values of a continuous latent process. Specifically, they modelled a continuous value $G_i \in [g_i, g_i{+}1]$ and found the values U_{1i} and U_{2i} such that $Q(U_{1i}|\boldsymbol{X}) = g_i$ and $Q(U_{2i}|\boldsymbol{X}) = g_i + 1$. Here we conduct the numerical analysis with comparison to the MQF approach.

2.5.2 Simulated Example 1

Consider a simple regression model for which the sample $Y_i (i = 1, 2, \cdots, n)$ are counts and follow a Poisson distribution with parameter 3 and a Binomial distribution with parameters 20 and 1/5, respectively. 500 simulations for each case of $\tau \in \{$0.05, 0.25, 0.50, 0.75, 0.95$\}$ and $n \in \{200, 1000\}$ are performed.

In this example, the quantile regression $Q_\tau(Y) = \beta(\tau)$ is a constant depending on τ only. Table 2.1 compares the posterior means with the true values of $\beta(\tau)$ for each case under 500 simulations. Moreover, the expectile regression $Expectile_\theta(Y) = \beta(\theta)$ is also a constant depending on the θ-th expectile. Table 2.2 compares the posterior means with the true values of $\beta(\theta)$ obtained via an empirical estimation in Eq.(2.10) for different cases. Figures 2.1-2.2 show that good convergence diagnostics can be obtained on the trace for various parameters of both Bayesian quantile regression and Bayesian expectile regression with discrete responses. It is encouraging to see that the results obtained by the proposed Bayesian inference are reasonably accurate.

Table 2.1: Posterior mean and posterior standard deviations (S.D.) of $\beta(\tau)$ from simulated example 1.

	$n = 200$		$n = 1000$		
τ	Mean	S.D.	Mean	S.D.	True value
Case 1 : $Y \sim \text{Pois}(3)$					
0.05	1.191	0.119	1.037	0.024	1
0.25	2.103	0.072	2.009	0.006	2
0.50	3.097	0.069	3.007	0.006	3
0.75	4.316	0.157	4.149	0.043	4
0.95	6.438	0.321	6.228	0.116	6
Case 2 : $Y \sim \text{Binom}(20, 1/5)$					
0.05	1.255	0.110	1.028	0.007	1
0.25	3.139	0.078	3.011	0.009	3
0.50	4.175	0.109	4.030	0.011	4
0.75	5.453	0.182	5.441	0.066	5
0.95	7.430	0.310	7.166	0.115	7

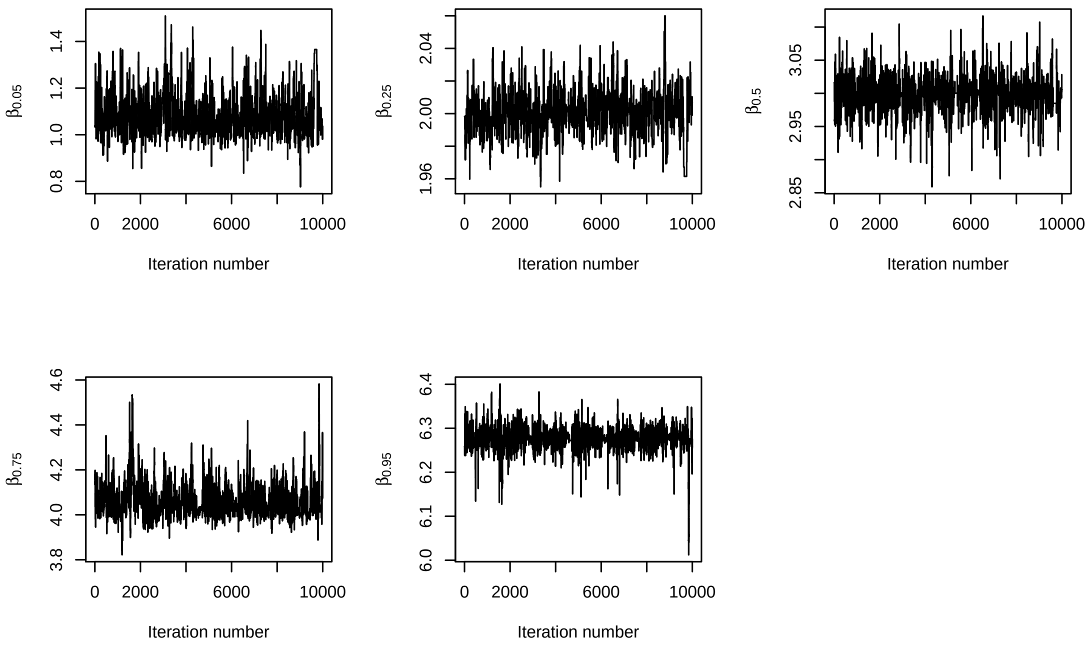


Figure 2.1: Convergence diagnostics on the trace for $Q_\tau(Y) = \beta(\tau)$ from simulated example 1 (n = 1000).

Table 2.2: Posterior mean and posterior standard deviations (S.D.) of $\beta(\theta)$ from simulated example 1.

	$n = 200$		$n = 1000$		
θ	Mean	S.D.	Mean	S.D.	True value
Case 1 : $Y \sim$ Pois(3)					
0.05	1.266	0.103	1.242	0.049	1.24
0.25	2.268	0.072	2.270	0.034	2.27
0.50	2.943	0.070	2.972	0.033	3
0.75	3.662	0.077	3.717	0.033	3.80
0.95	5.029	0.136	5.016	0.061	5.15
Case 2 : $Y \sim$ Binom$(20, 1/5)$					
0.05	2.321	0.109	2.072	0.049	2.11
0.25	3.360	0.071	3.256	0.032	3.23
0.50	4.086	0.070	4.064	0.032	4
0.75	4.825	0.077	4.869	0.034	4.80
0.95	6.051	0.128	6.294	0.056	6.15

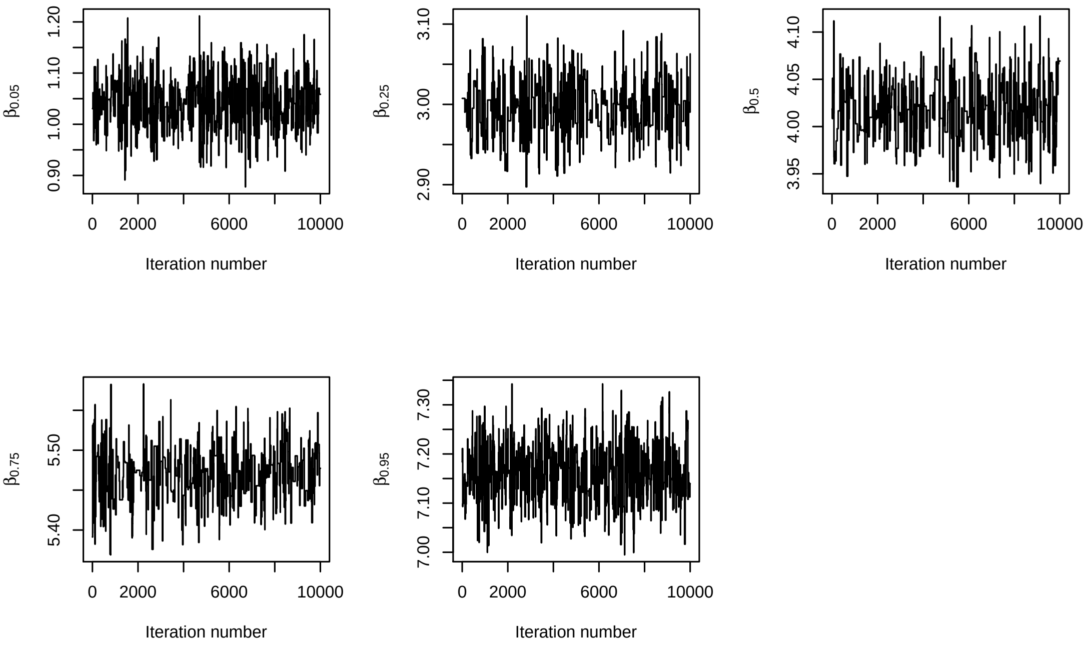


FIGURE 2.2: Convergence diagnostics on the trace for $Expectile_\theta(Y) = \beta(\theta)$ from simulated example 1 (n = 1000).

2.5.3 Simulated Example 2

We consider a discrete quantile linear regression:

$$Y_i = \beta_0 + \sum_{k}^{p} \beta_k X_{ik} + \varepsilon_i, \;\; i = 1, \cdots, n; \;\; k = 1, \cdots, p \tag{2.15}$$

where n and p denote the number of observations and independent variables, respectively. $\beta_k, k = 1, ..., p$ are the regression parameters. Let the random item ε_i follow a Poisson distribution with parameter 3. In this particular simulated example, a discrete distribution has to be artificially imposed to X_{ik} in order to obtain discrete responses.

500 simulations for each case of $\tau \in \{0.25, 0.50, 0.75\}$ and $n \in \{300, 1500\}$ are performed. A more thorough comparison of the fitted model to a latent process based approach MQF (Smith et al. 2015) is also provided. Furthermore, we illustrate the goodness of fit with posterior predictive power check via a partition of data into training data and test data and using the root mean square error (RMSE) and mean absolute error (MAE) of the predicted values with respect to the true outcome. That is, the first sample $n_1 \in \{200, 1000\}$ are used as training data for model fitting via parameter estimation and the

remaining sample $n_2 \in \{100, 500\}$ are left as test data for the out-of-sample evaluation. Then, we compare the predictive power of the proposed method and the MQF approach. These measures of goodness of fit are given by

$$\text{RMSE} = \sqrt{\frac{1}{n_2}\sum_{i=1}^{n_2}\left(Y_i - \mathcal{E}^{(M)}(i)\right)^2}$$

and

$$\text{MAE} = \frac{1}{n_2}\sum_{i=1}^{n_2}|Y_i - \mathcal{E}^{(M)}(i)|,$$

where $\mathcal{E}^{(M)}(i)$ denotes the mean for the posterior predictive distribution for Y_i.

2.5.3.1 Small p Case

Starting from a simple setting of Eq.(2.15), we set the number of independent variables $p = 2$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad i = 1, \cdots, n,$$

where, covariate $X_{1,i}$ is generated from a Geometric distribution with probability 1/4, and covariate $X_{2,i}$ is generated from a Poisson distribution with parameter 2. We generate the training data with $\beta_i = \{6, 2, -4\}, i = \{0, 1, 2\}$ and $\varepsilon_i \sim \text{Pois}(3)$. 500 simulations for each case of $\tau \in \{0.25, 0.50, 0.75\}$ and $n_1 \in \{200, 1000\}$ are performed.

Therefore, the corresponding discrete quantile function is of the form

$$Q_\tau(Y|X) = \beta_0(\tau) + \beta_1(\tau)X_1 + \beta_2(\tau)X_2.$$

Under the proposed Bayesian inference in Section 2.3, Figure 2.3 shows the comparison between the estimated and the true probability mass functions from this simulation with different τs. The boxplots in Figure 2.4 compare the posterior mean of the regression parameters $\beta_0(\tau), \beta_1(\tau)$ and $\beta_2(\tau)$, based on the proposed approach and MQF in Smith et al. (2015), under 500 simulations with $\varepsilon_i \sim \text{Pois}(3)$. Table 2.3 presents the computational efficiency by reporting the running time of the proposed method and MQF for discrete responses. MQF is implemented by function *qreg* in R package *BSquare*, with parameter $L = 4$ (the number of basis functions in MQF) and base set as 'Gaussian'.

Clearly, the proposed method is competitive overall, particularly smaller bias and more efficient when a large amount of data is to be processed.

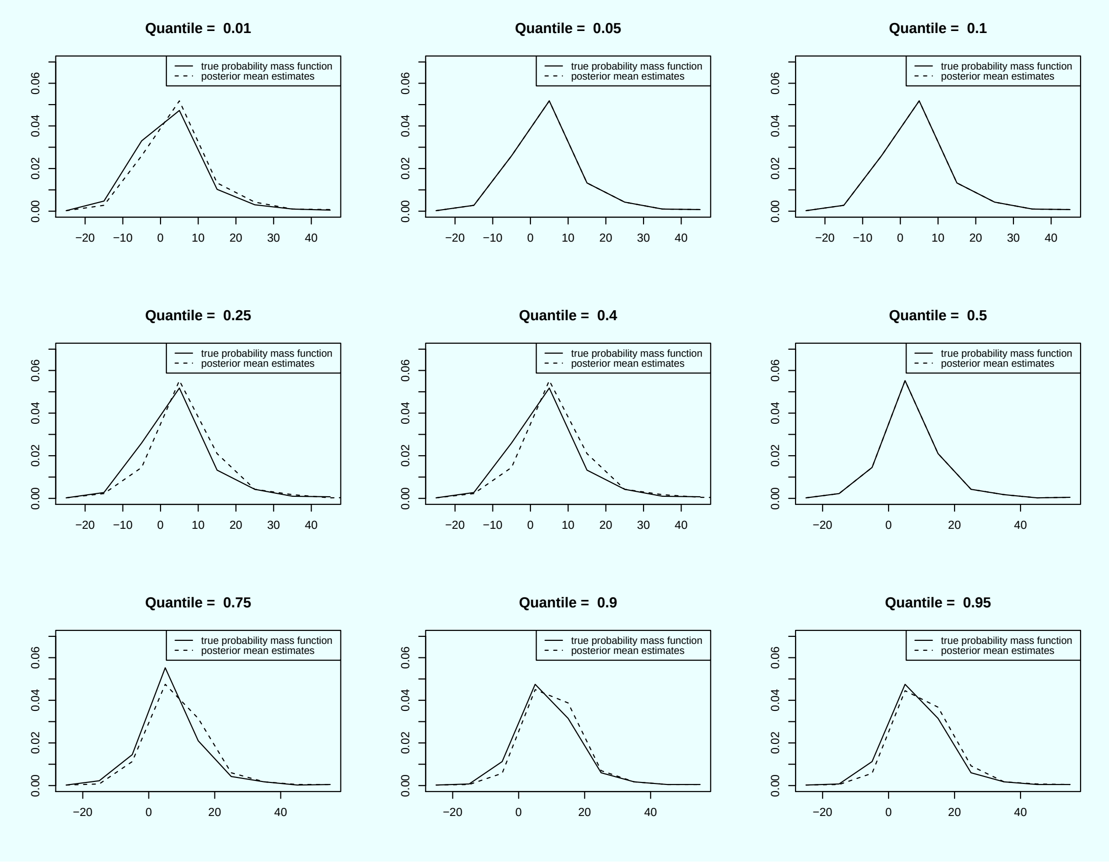


FIGURE 2.3: True probability mass functions (solid) with $\tau \in \{$ 0.01, 0.05, 0.1, 0.25, 0.40, 0.50, 0.75, 0.9, 0.95$\}$ and their posterior mean estimates (dashed) from simulated example 2.5.3.1 with $n_1 = 200$.

TABLE 2.3: Computation efficiency (in secs) from simulated example 2.5.3.1.

	$n_1 = 200$		$n_2 = 1000$	
τ	DALD	MQF	DALD	MQF
0.25	2.94	3.51	4.79	16.06
0.50	2.87	3.51	5.02	16.58
0.75	2.82	3.47	5.04	16.8

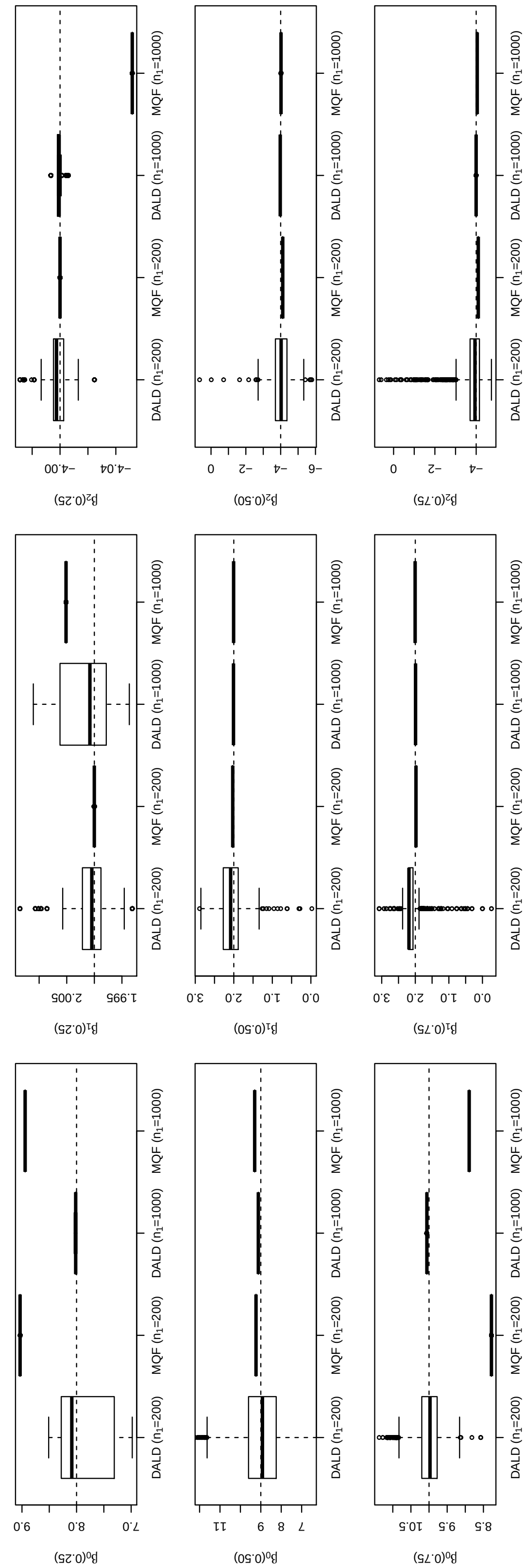


FIGURE 2.4: Boxplots of $\beta(\tau)$ at $\tau \in \{0.25, 0.5, 0.75\}$ from simulated example 2.5.3.1, where horizontal dashed lines denote the corresponding true values.

Moreover, although we have chosen improper flat priors in the above numerical experiments, one may use other priors for analysis in a relatively straightforward fashion. For example, along with Alhamzawi and Yu (2013), conditional conjugate prior distribution in the Normal-Gamma Inverse form for the unknown parameters $\boldsymbol{\beta}$ can be obtained. Given $\tau \in (0,1)$, for any $a > 0$, the prior mean and covariance matrix for $\boldsymbol{\beta}$ are given, respectively, by

$$E(\boldsymbol{\beta}) = \boldsymbol{\beta}_{a\tau}$$
$$Cov(\boldsymbol{\beta}) = 2g(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}^T)^{-1}$$

where $\boldsymbol{\beta}_a$ are anticipated values, and $g > 0$ is a known scaling factor. Various values of g have been used in the context of variable selection and estimation. Smith and Kohn (1996) performed variable selection using splines and suggested that the value of g is in the range $10 \leq g \leq 1000$. Following the discussions in Chen et al. (2011) and Alhamzawi and Yu (2013) among others, we set $g = 100$ in this chapter. Thus, given τ and $\boldsymbol{\beta}_{a\tau}$, the conditional prior distribution for $\boldsymbol{\beta}$ is readily available. Here we suggest a particular form of a conjugate Normal-Inverse Gamma family for $\boldsymbol{\beta}$ given by

$$\boldsymbol{\beta}|V, X \sim N(\boldsymbol{\beta}_a, 2g(\boldsymbol{X}\boldsymbol{V}\boldsymbol{X}^T)^{-1}),$$

where the prior information are set to be obtained by the semi-parametric jittering approach (Machado and Santos Silva, 2005), as presented in Table 2.4.

TABLE 2.4: The prior mean and covariance matrix for $\boldsymbol{\beta}$.

τ	$\boldsymbol{\beta}_a$	$Cov(\boldsymbol{\beta})$
0.25	[1.882, 0.489, -3.344]	$\begin{bmatrix} 28.696 & -27.642 & 224.511 \\ -27.642 & 27.209 & -221.267 \\ 224.512 & -221.267 & 1802.613 \end{bmatrix}$
0.50	[2.691, 0.349, -1.482]	$\begin{bmatrix} 0.107 & 0.007 & -0.114 \\ 0.007 & 0.003 & 0.003 \\ -0.114 & -0.013 & 0.136 \end{bmatrix}$
0.75	[2.320, 0.271, -0.674]	$\begin{bmatrix} 0.001 & 0.000 & -0.0010 \\ 0.000 & 0.000 & -0.001 \\ -0.001 & -0.000 & 0.002 \end{bmatrix}$

Under the proposed Bayesian inference in Section 2.2, Table 2.5 reports the posterior mean, standard deviation and 95% credible interval for the regression parameters $\beta_0(\tau), \beta_1(\tau)$ and $\beta_2(\tau)$, under 500 simulations with $\varepsilon_i \sim \text{Pois}(3)$, based on a conjugate Normal-Inverse Gamma prior for $\boldsymbol{\beta}$. It can be shown from both Figure 2.4 and Table 2.5 that under different prior settings, the regression coefficients obtained from the working likelihood analysis are consistent.

TABLE 2.5: Posterior mean, standard deviation and 95% credible interval of $\beta_k(\tau), k = 0, 1, 2$ from simulated example 2.5.3.1 based on a conjugate Normal-Inverse Gamma prior for $\boldsymbol{\beta}$ ($\varepsilon_i \sim \text{Pois}(3)$).

$\beta_k(\tau)$	$n = 200$			$n = 1000$			
	CI	Mean	S.D	CI	Mean	S.D	value
$\beta_0(.25)$	(7.894, 8.308)	8.075	0.111	(7.678, 8.131)	7.733	0.111	8
$\beta_0(.50)$	(8.881, 9.277)	9.061	0.098	(9.009, 9.051)	9.025	0.016	9
$\beta_0(.75)$	(9.697, 10.483)	10.086	0.186	(9.980, 10.116)	10.031	0.036	10
$\beta_1(.25)$	(1.976, 2.050)	2.007	0.020	(1.897, 2.016)	2.007	0.030	2
$\beta_1(.50)$	(1.926, 2.036)	1.996	0.025	(1.998, 2.003)	2.000	0.002	2
$\beta_1(.75)$	(1.926, 2.070)	1.999	0.034	(1.995, 2.018)	2.004	0.006	2
$\beta_2(.25)$	(-4.043, -3.867)	-3.977	0.042	(-4.006, -3.999)	-4.004	0.004	-4
$\beta_2(.50)$	(-4.032, -3.881)	-3.979	0.042	(-4.001, -3.983)	-3.998	0.003	-4
$\beta_2(.75)$	(-4.072, -3.679)	-3.930	0.101	(-4.014, -3.970)	-3.996	0.010	-4

2.5.3.2 Moderate p case

In order to further investigate the performance of the proposed method coping with relative complex settings, we consider different situations when p is moderate.

We generate training data from Eq.(2.15) with $n_1 = 200$ and $p = 20$ from different scenarios of $\boldsymbol{\beta}$, associated with $\varepsilon_i \sim \text{Pois}(3)$, while validation data with $n_2 = 100$. $X_{k,i}$ follows a Geometric distribution with probability $1/4$. 500 simulations for each case of $\tau \in \{0.25, 0.50, 0.75\}$ are performed. The corresponding discrete quantile function is of the form

$$Q_\tau(Y|X) = \beta_0(\tau) + \sum_{k=1}^{p} \beta_k(\tau) X_j.$$

Here, we consider dense, sparse and very sparse as three different scenarios of $\boldsymbol{\beta}$, presented in Table 2.6. We have chosen improper flat priors for simplicity.

TABLE 2.6: Three different scenarios of $\boldsymbol{\beta}$.

Scenarios of coefficients	$\boldsymbol{\beta}$
dense	$\beta_0 = 2; \beta_k = 2, k \in \{1, 2, \cdots, p\}$
sparse	$\beta_0 = 2; \beta_m = 2, 4, 5, 4, 1, m \in 1, 2, 5, 9, 14$ $\beta_n = 0, n \in \{k\} \backslash \{m\}, k \in \{1, 2, \cdots, p\}$
very sparse	$\beta_0 = 2; \beta_1 = 5; \beta_k = 0, k \in \{k = 2, 3, \cdots, p\}$

Table 2.7 compares the posterior mean, standard deviation and 95% credible interval for the regression parameters to those obtained by MQF in Smith et al. (2015). Table 2.7 reports a selective but representative result based on $\tau = 0.75$ with different $k \in \{0, 1, 2, 5, 9, 14\}$. It shows that as the number of covariates increases, the regression coefficients obtained from the working likelihood analysis are consistent. Table 2.8 presents the computational efficiency by reporting the running time of both DALD and MQF for discrete responses at $\tau \in \{0.25, 0.50, 0.75\}$. We ran 10,000 iterations of burn-in and 10,000 iterations after that, MQF is implemented by function *qreg* in R package *BSquare*, with parameter $L = 4$ (the number of basis functions in MQF) and base set as 'Gaussian'. Table 2.8 shows that DALD is less time-consuming than MQF. This advantage is more pronounced when the sample size is larger.

TABLE 2.7: Posterior mean, standard deviation and 95% credible interval of $\beta_k(\tau), k \in \{0, 1, 2, 5, 9, 14\}$ from simulated example 2.5.3.2 ($\tau = 0.75$).

$\boldsymbol{\beta}$	DALD			MQF			
	CI	Mean	S.D.	CI	Mean	S.D.	True value
Case 1 : Dense Scenario							
β_0	(4.832, 6.770)	5.737	0.500	(2.856, 5.856)	4.557	0.788	6
β_1	(1.919, 2.142)	2.032	0.058	(1.950, 2.074)	2.014	0.032	2
β_2	(1.918, 2.102)	2.012	0.046	(1.956, 2.053)	2.003	0.024	2
β_5	(1.909, 2.116)	2.016	0.052	(1.920, 2.038)	1.976	0.030	2
β_9	(1.908, 2.093)	2.003	0.048	(1.926, 2.061)	1.999	0.033	2
β_{14}	(1.920, 2.160)	2.025	0.061	(1.926, 2.061)	1.999	0.033	2
Case 2 : Sparse Scenario							
β_0	(4.044,7.218)	5.561	0.803	(4.624, 7.002)	5.728	0.633	6
β_1	(1.863,2.089)	1.964	0.063	(1.919, 2.048)	1.985	0.033	2
β_2	(3.865,4.087)	3.981	0.053	(3.977, 4.130)	4.051	0.039	4
β_5	(4.832, 5.086)	4.956	0.065	(4.975, 5.148)	5.060	0.043	5
β_9	(3.917,4.098)	4.006	0.050	(3.985, 4.128)	4.054	0.036	4
β_{14}	(0.955, 1.233)	1.088	0.069	(0.940, 1.070)	1.010	0.032	1
Case 3 : Very Sparse Scenario							
β_0	(4.506, 7.190)	5.946	0.722	(4.299, 6.419)	5.381	0.546	6
β_1	(4.793, 5.073)	4.953	0.069	(4.988, 5.131)	5.055	0.036	5
β_2	(-0.096,0.028)	-0.008	0.049	(-0.139, 0.039)	-0.048	0.045	0
β_5	(-0.098, 0.129)	-0.085	0.067	(-0.021, 0.146)	0.063	0.043	0
β_9	(-0.232, 0.059)	-0.011	0.070	(-0.022, 0.125)	0.054	0.038	0
β_{14}	(-0.084, 0.091)	0.027	0.041	(0.012, 0.144)	0.077	0.033	0

TABLE 2.8: Computation efficiency (in secs) from simulated example 2.5.3.2.

τ	DALD	MQF
Case 1 : Dense Scenario		
0.25	45.54	138.19
0.50	43.23	136.00
0.75	46.67	137.53
Case 2 : Sparse Scenario		
0.25	45.76	137.56
0.50	47.23	137.23
0.75	47.08	136.48
Case 3 : Very Sparse Scenario		
0.25	44.33	136.68
0.50	44.67	138.12
0.75	46.91	137.95

In order to evaluate the prediction accuracy for simulated examples 2.5.3.1 and 2.5.3.2, we separate the entire sample $n \in \{300, 1500\}$ into training data $n_1 \in \{200, 1000\}$ and validation data $n_2 \in \{100, 500\}$. We compute the average root mean square error (RMSE) and the average mean absolute error (MAE) for the prediction of Y_i for the validation data. The superiority of the proposed method is demonstrated in Table 2.9, which summarises the simulation results for three representative values of τ: 0.25, 0.50 and 0.75. The value of the two prediction indices for DALD is always less than those for MQF, which demonstrates that the proposed method outperforms MQF under both simple settings and complex settings.

Table 2.9: Average value of the prediction indices from simulated example 2. ($\varepsilon_i \sim$ Pois(3) for $p = 2$ and very sparse scenario for $p = 20$)

		$n = 300,\ p = 2$		$n = 1500,\ p = 2$		$n = 300,\ p = 20$	
τ	Indices	DALD	MQF	DALD	MQF	DALD	MQF
0.25	RMSE	1.943	2.173	1.883	2.133	1.885	2.316
	MAE	1.533	1.686	1.491	1.797	1.542	2.029
0.50	RMSE	1.946	1.948	1.876	1.887	1.491	1.970
	MAE	1.522	1.542	1.523	1.549	1.112	1.503
0.75	RMSE	1.680	1.947	1.878	2.110	1.951	2.331
	MAE	1.317	2.384	1.485	1.583	1.259	1.878

2.5.4 Simulated example 3

Alternatively, we simulate data from a Poisson distribution

$$Y_i|_{X_i=x_i} \sim \text{Pois}(\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}))$$

where X_{i1} is generated from a beta distribution with parameters 5/3 and 5/3, and X_{i2} is a dummy variable that equals 1 with probability 0.2 and 0 otherwise. We generate the simulated data from $\beta_i \in \{1, 1, 0\}, i \in \{0, 1, 2\}$ with sample size $n = 300$. 500 simulations for each case of $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$ are performed.

Table 2.10 compares the posterior mean and standard deviation for the regression parameters $\beta_0(\tau), \beta_1(\tau)$ and $\beta_2(\tau)$ to those obtained by MQF in Smith et al. (2015). Figure 2.5 displays three boxplots of the posterior mean of the regression parameters under 500 simulations, where the horizontal dashed lines denote the true values. The outperformance of the proposed method can be shown from both Table 2.10 and Figure 2.5. Moreover, the regression coefficients obtained from the working likelihood analysis are always accurate.

Table 2.10: Posterior mean and posterior standard deviations (S.D.) of $\beta(\tau)$ from simulated example 3.

	DALD		MQF		
τ	Mean	S.D.	Mean	S.D.	True value
$\beta_0(.10)$	-0.031	0.379	-0.013	0.577	-0.265
$\beta_0(.25)$	0.549	0.210	0.824	0.515	0.497
$\beta_0(.50)$	0.904	0.110	2.193	0.616	0.952
$\beta_0(.75)$	1.375	0.113	3.757	0.612	1.289
$\beta_0(.90)$	1.734	0.135	5.396	0.726	1.585
$\beta_0(.95)$	1.957	0.175	6.251	0.652	1.718
$\beta_1(.10)$	1.635	0.418	4.468	0.692	1.773
$\beta_1(.25)$	1.321	0.248	4.552	0.662	1.215
$\beta_1(.50)$	1.032	0.115	4.501	0.706	1.046
$\beta_1(.75)$	0.870	0.130	4.467	0.674	0.924
$\beta_1(.90)$	0.793	0.204	4.456	0.709	0.802
$\beta_1(.95)$	0.722	0.215	4.541	0.697	0.753
$\beta_2(.10)$	0.031	0.202	-0.132	0.350	-0.018
$\beta_2(.25)$	-0.011	0.111	-0.118	0.327	-0.023
$\beta_2(.50)$	0.093	0.062	-0.095	0.375	-0.011
$\beta_2(.75)$	0.005	0.070	-0.127	0.314	0.001
$\beta_2(.90)$	-0.077	0.069	-0.156	0.394	-0.003
$\beta_2(.95)$	-0.140	0.088	-0.124	0.344	0.008

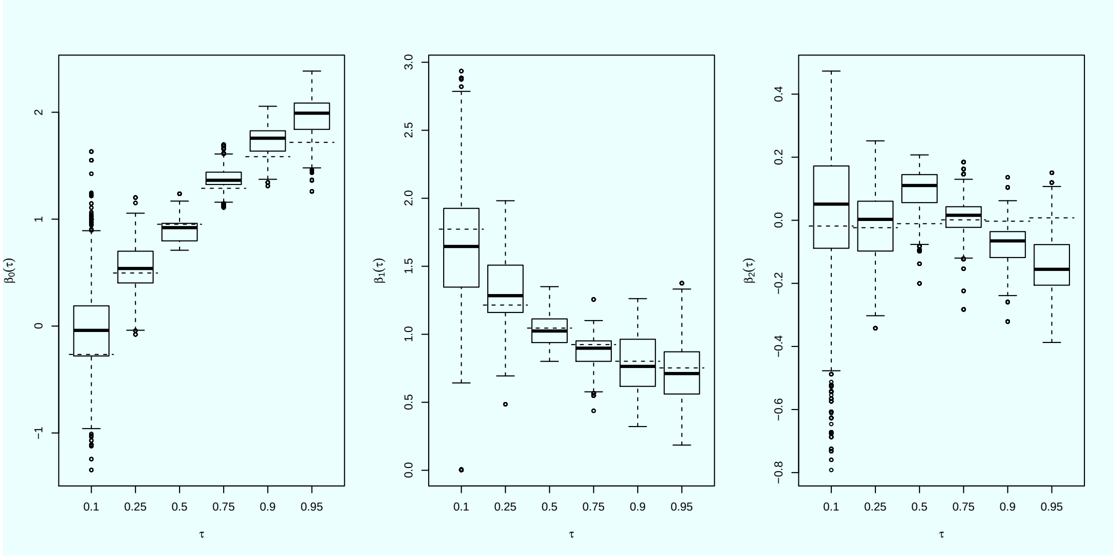


FIGURE 2.5: Boxplots of $\beta(\tau)$ at $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$ from simulated example 3, where horizontal dashed lines denote the true values).

2.5.5 Analysis of Length of Stay (LoS) in Days

The data is extracted from the Worcester Heart Attack Study with 500 observations (Hosmer et al. 2008). For simplicity but without loss of generality, we focus on exploring the relationship of LoS associated with age, gender, BMI (Body Mass Index) and hr (Initial Heart Rate), which are detailed in Table 2.11. The distribution of LoS is skewed and one is usually more interested in long stay or short stay than an average stay (Borghans et al., 2014, Wolkewitz et al., 2017 and among others). We aim to investigate how these factors affect the long LoS, so that we restrict the analysis under $\tau \in \{0.76, 0.78, \cdots, 0.94, 0.96\}$.

TABLE 2.11: Statistical description of the factors in data.

	Max.	Min.	Mean	S.D.	Skew.	Kurt.
Age	104	30	69.85	14.491	-0.378	-0.637
Age × Gender	104	0	29.89	37.456	0.525	-1.600
BMI	44.839	13.045	26.614	5.406	0.527	0.378
hr	186	35	87.018	23.586	0.563	0.441

Therefore, we fit a quantile regression model of the form

$$Q_\tau(Y|X) = \beta_0(\tau) + \beta_1(\tau)\text{Age} + \beta_2(\tau)(\text{Age} \times \text{Gender}) + \beta_3(\tau)\text{BMI} + \beta_4(\tau)\text{hr}.$$

Figure 2.6 displays four boxplots of posterior regression parameters across τs. Overall, the joint effect of gender and age on the outcome of interest is not significant. A closer look at the boxplots reveals that BMI has an increasing negative effect on LoS with τ towards extreme values, as the posterior mean of $\beta_3(\tau)$ reaches $-0.014, -0.120, -0.257$ with $\tau = 0.92, 0.94, 0.96$, respectively. That is, when BMI increases by 1 unit (kg/m^2), the distribution of LoS hasn't been changed much until the extreme upper tail, where LoS decreases with increase of τ. Similarly, heart rate has a significant and positive effect on LoS. Particularly, the positive effect of heart rate on LoS is increasing with τ.

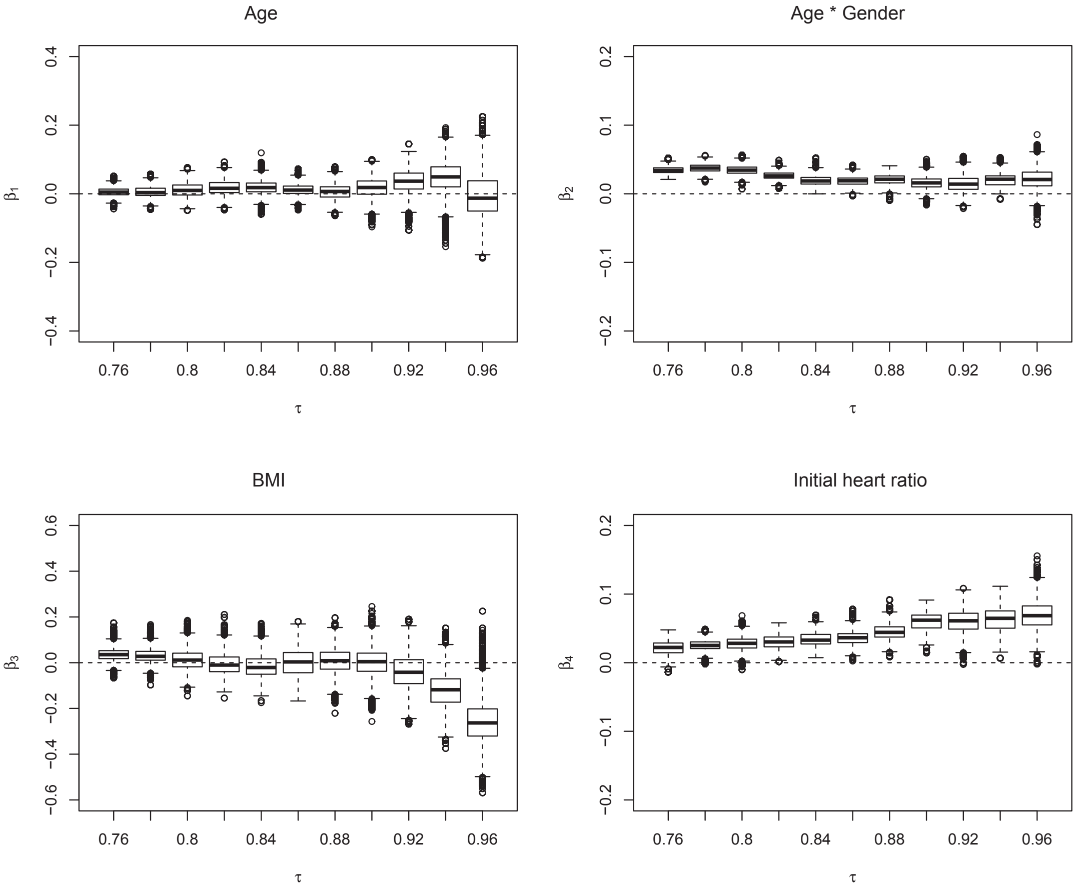


FIGURE 2.6: Boxplots of posterior regression parameters for LoS data across τs with $\tau \in \{0.76, 0.78, 0.80, \cdots, 0.92, 0.94, 0.96\}$.

2.6 Chapter Summary

Discrete responses are common in many disciplines. Regression analysis of discrete responses has been an active and promising area of research. Data with discrete responses are often analyzed incorrectly with ordinary least squares regression or regression for mean. We propose Bayesian quantile regression and Bayesian expectile regression for discrete responses. This is achieved by using a discrete asymmetric Laplace distribution and discrete asymmetric normal distribution to form the likelihood function respectively. The method is shown robust numerically and coherent theoretically. The Bayesian approach which is fairly easy to implement and provides complete univariate and joint posterior distributions of parameters of interest. The posterior distributions of the unknown model parameters are obtained by using M-H algorithm implemented in R. We have shown the usefulness of this approach through two simulated examples and one real data analysis. The extensions of the proposed approach to spatial and random effects models would represent interesting areas of development.

Chapter 3

Improved Local Quantile Regression

In this chapter we investigate a new kernel-weighted likelihood smoothing quantile regression method. The likelihood is based on a normal scale-mixture representation of the asymmetric Laplace distribution (ALD). This approach enjoys the same good design adaptation as the local quantile regression (Spokoiny et al., 2014), particularly for smoothing extreme quantile curves, and ensures non-crossing quantile curves for any given sample. The performance of the proposed method is evaluated via extensive Monte Carlo simulation studies and one real data analysis.

3.1 Introduction

Parametric quantile regression (Koenker, 2005) has been used in a number of disciplines to explore the relationship between the response and covariates at both the center and extremes of the conditional distribution and obtain a more comprehensive analysis of the relationship between variables. While a parametric model is possibly misspecified, non-parametric models, on the other hand, require fewer assumptions about the data and offer a more flexible way of modelling a relationship than parametric models, consequently avoid model misspecification when a parametric model is not available, which is common in wide applications (Wand and Jones, 1995; Fan and Gijbels, 1996; Takezawa, 2005). One of the popular nonparametric smoothing techniques is kernel smoothing.

Nonparametric kernel smoothing quantile regression has attracted much attention in the literature (Chaudhuri, 1991; Hardle and Mammen, 1993; Fan and Gijbels, 1996; Yu and Jones, 1998; Cai and Xu, 2008; Dette and Volgushev, 2008; Dabo-Niang and Laksaci, 2012; Schaumburg, 2012; Kong and Xia, 2015; among others).

However, the performance of kernel smoothing techniques, in spite of their advantages over parametric models in dealing with model misspecification, depends on smoothing parameter or bandwidth selection. While a global bandwidth such as the rule of thumb (Yu and Jones, 1998) is generally useful, a point-wise bandwidth, which depends on the values of covariate X or the design set should be considered for the complexity of the underlying regression functions. In particular, bandwidth selection in nonparametric smoothing quantile regression requires not only design adaption but also quantile adaption. Spokoiny, Wang and Härdle (henceforth SWH) (Spokoiny et al., 2014) developed a kernel-weighted likelihood quantile regression with point-wise bandwidth selection and promising performance in practice.

But SWH's approach may not guarantee non-crossing quantile curves for any given sample (calculated for various percentile $\tau \in (0,1)$), which is a common problem in the estimation of conditional and structural quantile functions due to lack of monotonicity. Note that, monotonicity (for each x in the design set, it's a monotone function of percentile value τ) guarantees non-crossing quantile curves, but not vice versa. Such a phenomenon violates the basic principle of probability theory, that is, the associated distribution functions should be monotone increasing. Various methods were presented to address or avoid the quantile crossing in parametric quantile regression, but with few on nonparametric quantile regression. Recently, Jones and Yu (2007) improved double kernel smoothing for quantile regression, Using spline-based constraints easily allows us to incorporate non-crossing conditions, as in Bondell et al. (2010) or Muggeo et al. (2013), for quantile estimation. Liu and Wu (2011) dealt with this issue via simultaneous multiple quantile smoothing, Qu and Yoon (2015) applied inequality constrains to ensure the monotonicity over quantiles.

In this chapter, we explore a local quantile regression based on a normal scale-mixture representation of asymmetric Laplace distribution (ALD) and show that this method has the similar property of SWH's procedure but much better-adaptive for smoothing extreme quantile curves. Moreover, quantile function is monotone with respect to τ for all x, which is satisfied by the proposed method, but SWH's method, which may also be non-crossing practically but without theoretical justification. Therefore, the proposed method enjoys both design adaptation and non-crossing quantile curves simultaneously. This chapter is organized as follows. We first review SWH's approach in Section 3.2, then propose a new local likelihood smoothing based on a normal scale-mixture representation of ALD and show that this approach satisfies the propagation condition (Spokoiny and Vial, 2009) in Section 3.3. In Section 3.4 we elaborate the proposed adaptive bandwidth selection rule and point out that the rule is able to avoid the problem of quantile curves crossing, especially for estimating extreme quantiles. Section 3.5 illustrates the numerical performance of the proposed method. Section 3.6 provides concluding remarks and discusses future work.

3.2 Kernel-Weighted Likelihood for Local Quantile Regression

Spokoiny et al. (2014) developed an interesting nonparametric quantile regression method: local quantile regression, which provides point-wise bandwidth selection and exhibits promising performance in practice. SWH claimed that their bandwidth selection rule is adaptive and novel, although the regression estimator named qMLE in their equation (8) is simply equivalent to a local polynomial quantile regression or a type of kernel-based weighting 'check function' approach, such as the local linear single-kernel approach of Yu and Jones (1998).

Let (X, Y) be the random variables, where Y is a continuous random variable and X is a univariate regressor $X \in \mathbb{R}^1$. Let $F_Y\left(Y|X\right)$ be the cumulative distribution function of Y given X. Let $Q_\tau\left(Y|X\right) = \inf\left\{Y : F_Y\left(Y|X\right) \geq \tau\right\}$ be the inverse function, which is also the value of a that minimizes the expected loss function:

$$Q_\tau\left(Y|X\right) = \underset{a}{\operatorname{argmin}}\, E\rho_\tau\left(Y - a\right), \tag{3.1}$$

where, $\tau \in (0, 1)$ and $\rho_\tau(\cdot)$ is an asymmetric loss function that satisfies $\rho_\tau(u) = u\left(\tau - I(u < 0)\right)$ with $I(\cdot)$ an indicator function.

Under the quantile non-parametric model $Y = f(X) + \epsilon$, given data in the form $\{X_i, Y_i\}_{i=1}^n$, where X_i and Y_i are independent scalar observations of X and Y, respectively. The τth conditional quantile of Y given X is estimated by

$$\hat{f}(X) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(X_i)\right). \tag{3.2}$$

SWH took advantage of the link between the minimization of the sum of the loss function in Eq.(1.3) and the maximum likelihood theory is given by the asymmetric Laplace distribution. For a random variable $Y \sim \text{ALD}(\mu, \sigma, \tau)$, its density function can be written as

$$f(y; \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\left[\tau - \mathrm{I}(y \leq \mu)\right]\right\}, \qquad y \in (-\infty, +\infty) \tag{3.3}$$

where, $0 < \tau < 1$ is skew parameter, $\sigma > 0$ is scale parameter, and $-\infty < \mu < \infty$ is location parameter.

Based on a ALD log-likelihood, SWH considered

$$L_{SWH}(\boldsymbol{\theta}) \equiv \log\{\tau(1-\tau)\} \sum_{i=1}^{n} I - \sum_{i=1}^{n} \rho_\tau \left(Y_i - f_\theta(X_i)\right), \tag{3.4}$$

with $0 < \tau < 1$ the level of the quantile. Then they fit $f(x)$ at point x by the local polynomial approach $Y_i = \boldsymbol{\psi}_i^T \boldsymbol{\theta} + \epsilon_i$, with basis $\boldsymbol{\psi}_i = \{1, (X_i - x), (X_i - x)^2/2!, \cdots, (X_i - x)^p/p!\}^T$ and $\boldsymbol{\theta} = (\theta_0, ..., \theta_p)^T$. Therefore, the local log-likelihood at x is given by

$$L_{SWH}(W, \boldsymbol{\theta}) \equiv \log \tau(1-\tau) \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta}\right) w_i, \tag{3.5}$$

where the weights W are chosen via a kernel function $w_i = K\left(\frac{X_i - x}{h}\right)$, where h is a bandwidth controlling the degree of localization. Note that, Eq.(3.5) is similar to the global log-likelihood in Eq.(3.4), but each summand in $L_{SWH}(W, \boldsymbol{\theta})$ is multiplied with the weight w_i, so only the points from the local vicinity of x contribute to $L_{SWH}(W, \boldsymbol{\theta})$.

The corresponding local quantile MLE (they named it as qMLE) at x is then given via the maximization of $L_{SWH}(W, \boldsymbol{\theta})$ in Eq.(3.4)

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{SWH}(x) &\equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \, L_{SWH}(W, \boldsymbol{\theta}) \\ &= \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta} \right) w_i. \end{aligned} \tag{3.6}$$

3.3 Local Quantile Regression with An Alternative Likelihood for Smoothing

Figure 3.1 displays the performance of SWH's approach, showing the bandwidth sequence (upper panel) and the smoothed 50% quantile curve (lower panel) based on the Lidar dataset (available in R package *'SemiPar'*), which adapts the data well. And this is also true for other moderate or central quantile curves. However, it can be seen from smoothing extreme quantile curves in Figure 3.2 here, the proposed bandwidth selection rule is lack of good adaptation and then results in the over-smoothing phenomenon. Figure 3.2 displays the smoothed 99% and 1% quantile curves using SWH's method and shows that when the curves start to switch smoothness, the rule is not adaptive so that the estimated curves are too smoothing out of the data ranges. A possibly theoretical interpretation for this problem is: when $\tau \to 0$, the weighted 'check function' $\rho_\tau(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta}) w_i$ takes constant 0 if $Y_i > \boldsymbol{\psi}_i^T \boldsymbol{\theta}$ (also, when $\tau \to 1$ and if $Y_i < \boldsymbol{\psi}_i^T \boldsymbol{\theta}$). This may result in that the proposed significant test always picks constant bandwidth for smoothing extreme quantile curves although this is not a problem for the local quantile regression estimation equation. We want to point out that this over-smoothing problem will be solved by a new version of adaptive bandwidth selection rule.

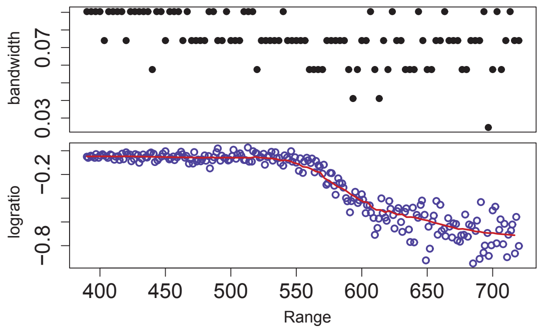
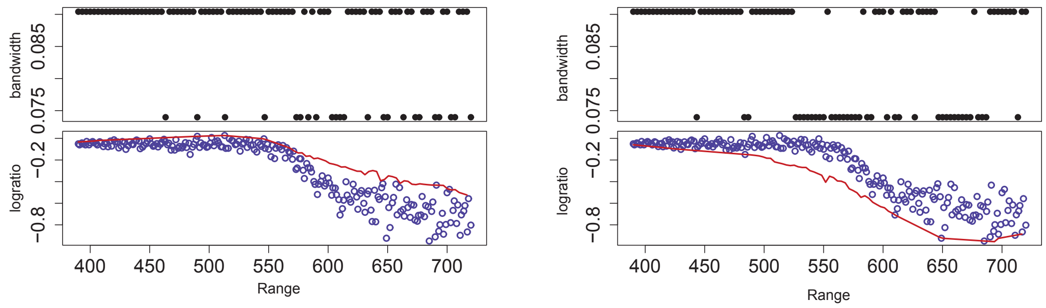


FIGURE 3.1: The bandwidth sequence (upper panel) and the adaptive estimation of 0.50 quantile (lower panel) for the Lidar dataset by SWH's kernel-weighted likelihood function.



(A) $\tau = 0.99$

(B) $\tau = 0.01$

FIGURE 3.2: The bandwidth sequences (upper panels) and smoothed quantile curves (lower panels) for the Lidar dataset using SWH's kernel-weighted likelihood.

Moreover, there is no guarantee of this approach to avoid quantile crossing. Therefore we propose an alternative adaptive bandwidth selection rule based on a normal scale-mixture representation (henceforth NSM) of ALD and show that this alternative version has the similar property of SWH's procedure but adapts much better for smoothing extreme quantile curves.

Reed and Yu (2009) and Kozumi and Kobayashi (2011) note that under the assumption of ALD-based 'working likelihood', the quantile regression model error $\epsilon \sim \text{ALD}(0, 1, \tau)$ can be represented as a scale mixture of normal variable, that is,

$$\epsilon = \mu z + \delta\sqrt{z}e, \tag{3.7}$$

where $\mu = \frac{1-2\tau}{\tau(1-\tau)}$, $\delta^2 = \frac{2}{\tau(1-\tau)}$, $z \sim Exp(1)$ and $e \sim N(0,1)$, and z and e are independent. Hence, SWH's model ($Y_i = f(X_i) + \epsilon_i$) could be re-written as

$$Y_i = f(X_i) + \mu z_i + \delta\sqrt{z_i}e_i. \tag{3.8}$$

That is, for given $\boldsymbol{z} = (z_1, z_2, ..., z_n)$,

$$Y_i \sim N\left(f(X_i) + \mu z_i,\ \delta^2 z_i\right), \tag{3.9}$$

i.e., the joint conditional density of $Y = (Y_1, Y_2, ..., Y_n)$ is given by

$$l\left(Y|\boldsymbol{z}, X\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\delta\sqrt{z_i}} \exp\left\{-\frac{(Y_i - f(X_i) - \mu z_i)^2}{2\delta^2 z_i}\right\}. \tag{3.10}$$

Clearly, if $\boldsymbol{z}$ is fixed in advance, then the local log-likelihood (SWH's Eq.(7)) can be replaced by a normal scale-mixture representation of ALD :

$$\begin{aligned} L_{NSM}(W, \boldsymbol{\theta}) \quad &\equiv \quad -\log(\sqrt{2\pi}\delta)\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\log(z_i)\,w_i \\ &- \quad \frac{1}{2\delta^2}\sum_{i=1}^{n}\frac{(Y_i - f(X_i) - \mu z_i)^2}{z_i}w_i - \sum_{i=1}^{n} z_i w_i, \end{aligned} \tag{3.11}$$

where the weights W are chosen via a kernel function $w_i = K\left(\frac{X_i - x}{h}\right)$, while h is a bandwidth controlling the degree of localization. Similar to Eq.(3.5), the local log-likelihood in Eq.(3.11) depends on the central point x via the structure of the basis vectors $\boldsymbol{\psi}_i$ and via the weights w_i.

Now, once a local pth-degree polynomial $\boldsymbol{\psi}_i^T\boldsymbol{\theta}$ is used to approximate $f(x)$ at $X = x$, the corresponding local qMLE at x could be defined via maximization of $L_{NSM}(W, \boldsymbol{\theta})$ above:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}(x) \quad &\equiv \quad \left(\tilde{\theta}_0(x), \tilde{\theta}_1(x), ..., \tilde{\theta}_p(x)\right) \\ &= \quad \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_{NSM}(W, \boldsymbol{\theta}) \\ &= \quad \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{\psi}_i^T\boldsymbol{\theta} - \mu z_i)^2}{\delta^2 z_i} w_i, \end{aligned} \tag{3.12}$$

where $\tilde{\theta}_0(x)$ estimates $f(x)$, and $\tilde{\theta}_m(x)$ estimates the m^{th} derivative of $f(x)$. Further, let $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, .., \boldsymbol{\psi}_n)^T$ and $\boldsymbol{w}_k = diag\left(\frac{w_1^{(k)}}{\delta^2 z_1}, ..., \frac{w_n^{(k)}}{\delta^2 z_n}\right)$, we have

$$\tilde{\boldsymbol{\theta}}_k(x) = \left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \boldsymbol{\psi}\boldsymbol{w}_k \left(Y + \mu\boldsymbol{z} + \delta\boldsymbol{z}^{1/2}\boldsymbol{e}\right), \tag{3.13}$$

where the design matrix $\boldsymbol{\psi}$ consists of the columns $\boldsymbol{\psi}_i = \{1, (X_i - x), \cdots, (X_i - x)^p/p!\}^T$.

We note that the $L_{NSM}(W, \boldsymbol{\theta})$ involves the specification of vector $\boldsymbol{z}$, and we point out that $\boldsymbol{z}$ could be fixed in advance via a sample from a data-driven inverse Gaussian distribution, and our extensive experiments in Section 3.5 show that the selection of the sample has no effect on the estimation. In fact, note that the joint likelihood function of $(Y, \boldsymbol{z})$ is give by

$$f(Y, \boldsymbol{z}|X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\,\tau\sqrt{z_i}} \exp\left\{-\frac{(Y_i - f(X_i) - \mu z_i)^2}{2\tau^2 z_i}\right\} \prod_{i=1}^{n} \exp(-z_i).$$

Therefore, the conditional density of $f(\boldsymbol{z}|Y)$ is given by

$$\begin{aligned} f(\boldsymbol{z}|Y) &\propto f(Y, \boldsymbol{z}) \\ &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{z_i}} \exp\left(-\frac{1}{2}\left[\frac{(Y_i - f(X_i))^2}{\delta^2} z_i^{-1} + \left(\frac{\mu^2}{\delta^2} + 2\right) z_i\right]\right). \end{aligned} \tag{3.14}$$

That is, $z_i, z_2, ..., z_n$ are i.i.d. with a generalized inverse Gaussian (GIG) distribution:

$$\begin{aligned} f(\boldsymbol{z}|Y) &\propto z_i^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[\frac{(Y_i - f(X_i))^2}{\delta^2} z_i^{-1} + \left(\frac{\mu^2}{\delta^2} + 2\right) z_i\right]\right) \\ &\sim GIG\left(\frac{1}{2}, \eta_i, \zeta_i\right), \end{aligned} \tag{3.15}$$

where $\eta_i^2 = \frac{(Y_i - f(X_i))^2}{\delta^2}$ and $\zeta_i^2 = \frac{\mu^2}{\delta^2} + 2$.

3.4 Performance of Adaptive Bandwidth Selection and Non-Crossing Estimation

3.4.1 Adaptive Bandwidth Selection

There are several methodologies for automatic smoothing parameter selection. One class of methods chooses the smoothing parameter value to minimize a criterion that incorporates both the tightness of the fit and model complexity. Such a criterion can usually be written as a function of the error mean square, and a penalty function designed to decrease with increasing smoothness of the fit. Examples of specific criteria are generalized cross-validation (Craven and Wahba, 1979) and the Akaike information criterion (AIC)(Akaike, 1973). These classical selectors have two undesirable properties when used with local polynomial and kernel estimators: they tend to under-smooth and tend to be non-robust in the sense that small variations in the input data can change the choice of smoothing parameter value significantly. Hurvich et al. (1998) obtained several bias-corrected AIC criteria that limit these unfavorable properties and perform comparably with the plug-in selectors (Ruppert et al., 1995).

The adaptive bandwidth selection rule in SWH's paper is different from the rule-of-thumb rule of Yu and Jones (1998) and AIC rule of Cai and Xu (2008). It does add a nice option to the bandwidth selection menu for practitioners. In this chapter, we perform the local quantile curve estimation following the similar bandwidth selection procedures, but based on a normal scale-mixture representation of ALD.

First, we fix a finite ordered set of candidates of bandwidth $h_1 < h_2 < \cdots < h_K$, where h_1 is very small. According to SWH, the bandwidth sequence can be taken geometrically increasing of the form $h_k = ab^k$ with fixed $a > 0$, $b > 1$, and $n^{-1} < ab^k < 1$ for $k = 1, \cdots, K$. For each $k \leq K$, an ordered weighting scheme $W^{(k)} = \left(w_1^{(k)}, w_2^{(k)}, \cdots, w_n^{(k)}\right)$ is chosen via a kernel function $w_i^{(k)} = K\left(\frac{X_i - x}{h_k}\right)$ leading to the local quantile estimator at x, $\tilde{\boldsymbol{\theta}}_k(x)$, as:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_k(x) &= \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_{NSM}(W^{(k)}, \boldsymbol{\theta}) \\ &= \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(Y_i - \boldsymbol{\psi}_i^T \boldsymbol{\theta} - \mu z_i)^2}{\delta^2 z_i} w_i^{(k)}. \end{aligned} \quad (3.16)$$

Then, we start with the smallest bandwidth h_1. For any $k > 1$, compute the local qMLE $\tilde{\boldsymbol{\theta}}_k(x)$ and check whether it is consistent with all the previous estimators $\tilde{\boldsymbol{\theta}}_l(x)$ for $l < k$. We use a localized likelihood ratio test, i.e. the difference $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right)$ to reject $\tilde{\boldsymbol{\theta}}_k(x)$, where $\tilde{\boldsymbol{\theta}}_l(x)$ maximize the log-likelihood $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) = \sup_\theta L_{NSM}\left(W^{(l)}, \boldsymbol{\theta}\right)$ defined in Eq.(3.11) with bandwidth h_l and $L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right)$ is the other local likelihood under $\tilde{\boldsymbol{\theta}}_k(x)$ with bandwidth $h_k(l < k)$. The difference checks whether $\tilde{\boldsymbol{\theta}}_k(x)$ belongs to the confidence set $\varepsilon_l(\zeta)$ of $\tilde{\boldsymbol{\theta}}_l(x)$:

$$\varepsilon_l(\zeta) := \left\{\boldsymbol{\theta} \in \Theta : L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{NSM}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right) \leq \zeta_l\right\}.$$

If the consistency check is negative, the procedure terminates and selects the latest accepted estimator. The adaptation algorithm can be summarized as follows:

Algorithm 3.1

Step 1: Start with $\hat{\boldsymbol{\theta}}_1(x) = \tilde{\boldsymbol{\theta}}_1(x)$.

Step 2: For $k \geq 2$, $\tilde{\boldsymbol{\theta}}_k(x)$ is accepted and $\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_k(x)$, if $\tilde{\boldsymbol{\theta}}_{k-1}(x)$ was accepted and

$$L_{New}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_l(x)\right) - L_{New}\left(W^{(l)}, \tilde{\boldsymbol{\theta}}_k(x)\right) \leq \zeta_l, \qquad l = 1, ..., k-1.$$

where the choice of critical values $\zeta_l, l = 1, ..., k-1$ are based on the propagating conditions in Theorem 3.1 below.

Step 3: Otherwise, $\hat{\boldsymbol{\theta}}_k(x) = \hat{\boldsymbol{\theta}}_{k-1}(x)$.

The adaptive estimator $\hat{\boldsymbol{\theta}}(x)$ is the latest accepted estimator after all K steps:

$$\hat{\boldsymbol{\theta}}(x) = \hat{\boldsymbol{\theta}}_K(x).$$

Moreover, all the estimators $\tilde{\boldsymbol{\theta}}_k(x)$ should be consistent to each other and the procedure should not terminate at any intermediate step $k < K$. This effect is called as 'propagation'. Hence, under the assumptions **(A1)-(A3)** in Appendix B, and then according to Serdyukova (2012), the propagation conditions (PC) for this approach also satisfies:

Theorem 3.1. *(Theoretical choice of the critical values.) Assume Assumptions B.1-B.3, given* $\alpha \in (0, 1]$ *and* $r > 0$*, the critical values* $\zeta_1, \cdots, \zeta_K$ *satisfy*

$$\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)^T \left(\boldsymbol{\psi} w_k(x) \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \leq \alpha C(p, r), \tag{3.17}$$

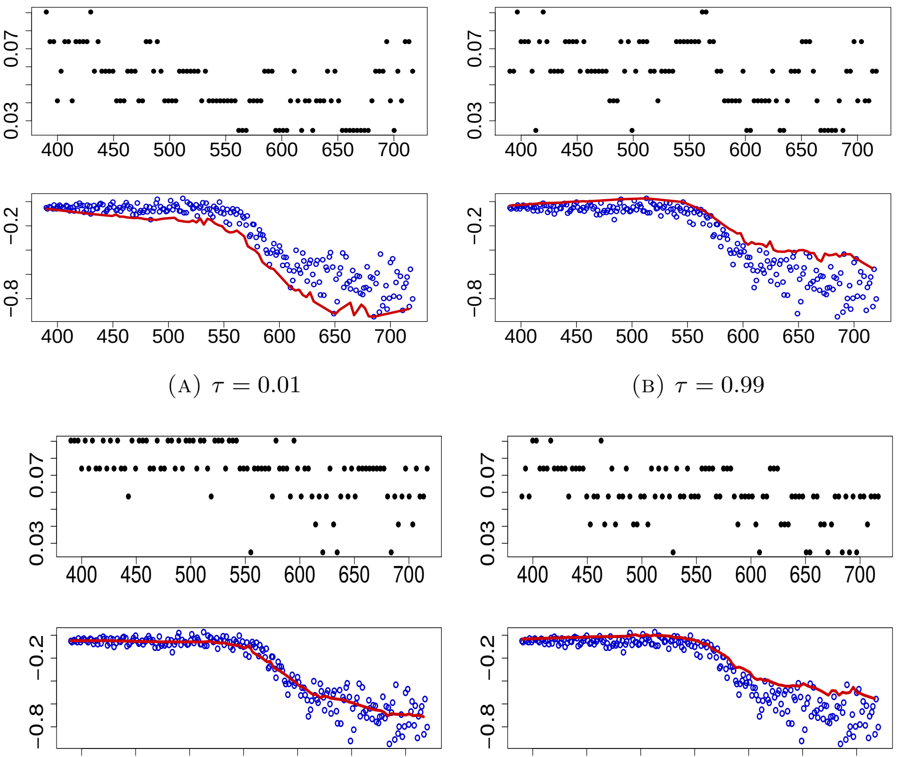
for all $k = 2, \cdots, K$*, where* $C(p, r) = 2^r\Gamma(r+p/2)/\Gamma(p/2)$*, with the choice of the critical values of the form*

$$\zeta_l = \frac{4}{a}\left\{r(K-l)\log b + \log\frac{K}{\alpha} - \frac{p}{4}\log(1-4\mu) - \log(1-b^{-r}) + \bar{C}(p,r)\right\}, l = 1, ..., k-1$$

where $a \in (0, 1/4)$ *is an arbitrary constant,* $b > 1$ *and* $\bar{C}(p, r) = \log\left\{\frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$.

The critical values are selected to ensure the desired propagation condition which effectively means a 'no alarm' property, that is the selected adaptive estimator coincides in the most cases that the estimator $\tilde{\boldsymbol{\theta}}_k(x)$ corresponding to the largest bandwidth. The critical values enter implicitly in the propagation condition: if the false alarm event $\left\{\tilde{\theta}_k(x) \neq \hat{\theta}_k(x)\right\}$ happens too often, it is an indication that some of the critical values ζ_l are too small.

An advantage of the proposed alternative normal scale-mixture likelihood function over SWH's method is that the derived bandwidth has better adaptation when τ tends to 0 or 1. Figure 3.3 displays the bandwidth sequence (upper panel) and smoothed quantile curves for quantiles 1% (3.3a) and 99% (3.3b) based on the Lidar dataset, which provides much better fitting than those curves presented in Figure 3.2. The dependency structure changing on smoothness is more adaptive than the bandwidth sequence in Figure 3.2. This alternative normal scale-mixture likelihood method also works well for other moderate or central quantile curves. Figure 3.3 shows that the method gives quite similar estimates to SWH's method for $\tau = 0.5$ (3.3c) and 0.9 (3.3d) quantile curves.



(A) $\tau = 0.01$

(B) $\tau = 0.99$

(C) $\tau = 0.50$

(D) $\tau = 0.90$

FIGURE 3.3: The bandwidth sequences (upper panels) and smoothed quantile curves (lower panels) for the Lidar dataset via the alternative normal scale-mixture likelihood.

3.4.2 Non-crossing Quantile Curve Estimation

The proposed bandwidth selection rule in SWH's method seems to have no quantile crossing phenomenon when several smoothed quantile curves are provided together. This indicates the advantage of the local bandwidth selection rule. Whereas most of published articles on this topic, which include constrained smoothing spline (He, 1997; Bondell et al., 2010), double-kernel smoothing (Yu and Jones, 1998; Jones and Yu, 2007) and monotone constraint on conditional distribution function (Hall et al., 1999; Dette and Volgushev, 2008), among others, focus on the development of new methods rather than adaptive bandwidth selection for avoiding quantile crossing. SWH showed that the adaptive bandwidth selection rule may not suffer quantile crossing issue, even with 'local constant' kernel smoothing quantile regression

$$\hat{q}_\tau(x) = \underset{a}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - a\right) K_h \left(x - X_i\right).$$

This may be true practically, but without a theoretical justification. Under our proposed approach, the justification of non-crossing quantiles could be outlined as below.

Recall the nonparametric quantile regression model $Y = f(X) + \epsilon$, where $Q_\tau(\epsilon) = 0$. Given data $\{X_i, Y_i\}_{i=1}^n$, and under the local polynomial approach, $\tilde{\theta}_0(x)$ estimates $f(x)$, with

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{NSM} &\equiv \left(\tilde{\theta}_0, \tilde{\theta}_1, \cdots, \tilde{\theta}_p\right) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}}\, L_{NSM}(W, \boldsymbol{\theta}),\end{aligned}$$

where the likelihood function $L_{NSM}(W, \boldsymbol{\theta})$ is expressed in Eq.(3.11) and $\tilde{\theta}_m(x)$ estimates the m^{th} derivative of $f(x)$.

That is, the derivative of $L_{NSM}(W, \boldsymbol{\theta})$ over $\tilde{\theta}_0(x)$ satisfies $\sum_{i=1}^n \frac{w_i}{z_i}(Y_i - \boldsymbol{\psi}_i^T \tilde{\boldsymbol{\theta}}_{NSM} - \mu z_i) = 0$. Therefore, $\tilde{\theta}_0(x)$ can be expressed as,

$$\tilde{\theta}_0(x) = \frac{\sum_{i=1}^n \frac{w_i}{z_i}\left(Y_i - \mu z_i - \sum_{j=1}^p \tilde{\theta}_j \frac{(X_i - x)^j}{j!}\right)}{\sum_{i=1}^n \frac{w_i}{z_i}}.$$

For each x, we aim to check the derivative of $\tilde{\theta}_0(x)$ over $\tau \in (0, 1)$. If $\frac{d\tilde{\theta}_0}{d\tau} > 0$, then $\tilde{\theta}_0$ is an increasing function of τ.

Note that $\mu = \frac{1-2\tau}{\tau(1-\tau)}$, therefore, we have

$$\begin{aligned}\frac{d\tilde{\theta}_0(x)}{d\tau} &= \frac{1}{\sum_{i=1}^n \frac{w_i}{z_i}} \sum_{i=1}^n \frac{-z_i w_i \frac{d\mu}{d\tau}}{z_i} \\ &= \frac{1}{\sum_{i=1}^n \frac{w_i}{z_i}} \sum_{i=1}^n \frac{-z_i w_i \frac{-2(\tau-1/2)^2-1/2}{\tau^2(1-\tau)^2}}{z_i} \\ &= \frac{1}{\sum_{i=1}^n \frac{w_i}{z_i}} \sum_{i=1}^n w_i \frac{2(\tau-1/2)^2+1/2}{\tau^2(1-\tau)^2} \\ &> 0. \end{aligned} \tag{3.18}$$

That is, $\hat{f}(x) \equiv \tilde{\theta}_0(x)$ is a strictly monotonic function of τ over x.

3.5 Numerical examples

In this section, we implement the proposed method via extensive Monte Carlo simulation studies and one real data analysis. All numerical experiments are carried out on one Inter Core i5-3470 CPU (3.20GMHz) processor and 8 GB RAM.

3.5.1 Simulation 1

In this simulation study, we aim to summarize our numerical results on choosing the critical values by the propagation condition as described in Section 3.1. We check the critical values at different quantile levels $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$, and for different choices of α and r. We also study how bandwidth sequence affects the critical values.

Table 3.1 shows the critical values with several choices of α and r with $\tau = 0.2$ and $m =$ 5000 Monte Carlo samples, and a bandwidth sequence $(5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$ scaled to the interval $[0, 1]$.

Table 3.2 shows the critical values for different τs with $\alpha = 0.25, r = 0.5$ and $m = 5000$ Monte Carlo samples, and a bandwidth sequence $(5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$ scaled to the interval $[0, 1]$.

Table 3.3 shows the critical values for the following alternative bandwidth sequences, with $\alpha = 0.25, r = 0.5, \tau = 0.8$ and $m = 5000$ Monte Carlo samples.

$$\eta_1 = (5, 7, 10, 13, 17, 21, 24, 28, 36, 45)/365$$
$$\eta_2 = (10, 13, 17, 21, 24, 28, 36, 45, 49, 60)/365$$
$$\eta_3 = (2, 3, 5, 7, 10, 13, 17, 21, 24, 28)/365$$

It is clear to show from Table 3.1 that critical values decrease when α increases, and increase when r increases. Table 3.2 shows that critical values behave similarly for symmetric τ. Overall, although the critical values differ for different bandwidth sequences, α, r and τ, they indicate the same patterns (finite and decreasing), which indicate that the adaptation algorithm can be completed in maximum $K = 6$ steps, as the values of critical values decrees to zero in 6-step.

Table 3.1: Critical values with different α and r ($\tau = 0.2$)

α	r	Critical values					
0.25	0.5	16.971	11.539	8.133	3.584	0.044	0.000
0.25	0.75	20.218	13.743	9.336	3.131	0.000	0.000
0.25	1	24.676	16.270	9.308	4.214	1.561	0.000
0.5	0.5	12.823	9.619	7.205	3.703	0.949	0.000
0.75	0.5	11.249	7.222	4.244	0.181	0.000	0.000

Table 3.2: Critical values with different τ ($\alpha = 0.25, r = 0.5$)

τ	Critical values					
0.05	10.357	7.605	4.888	1.248	0.000	0.000
0.25	15.782	11.332	8.440	4.354	0.908	0.000
0.50	21.714	15.427	10.351	3.594	0.000	0.000
0.75	15.283	10.932	8.396	3.949	0.840	0.000
0.95	10.789	7.686	4.943	1.208	0.000	0.000

Table 3.3: Critical values with different bandwidth sequences ($\alpha = 0.25, r = 0.5, \tau = 0.8$)

η	Critical values					
η_1	11.002	6.508	3.089	0.000	0.000	0.000
η_2	23.187	13.810	7.775	3.690	0.000	0.000
η_3	6.871	4.737	2.046	0.389	0.000	0.000

3.5.2 Simulation 2

In this simulation study, we compare the performance of our proposed approach to SWH's method as well as two other bandwidth selection techniques. One proposal comes from Ng and Maechler (2007), in which they considered constrained quantile estimations using linear or quadratic splines (implemented with R function *cobs* in Package *cobs*), and the other is from Yu and Jones (1998), in which they considered a rule of thumb bandwidth (implemented with R function *lprq* in Package *quantreg*).

We generate one training data of size 2000 and 500 test data sets of size 500 from the model

$$Y = m(X) + \sigma(X)\varepsilon, \tag{3.19}$$

where the univariate input X follows a uniform distribution on $[4, 4]$ and $m(X)$ is a non-linear function of X

$$m(X) = (1 - X + 2X^2)e^{-0.5x^2},$$

and the scale factor $\sigma(X)$ is linearly increasing in X with the form

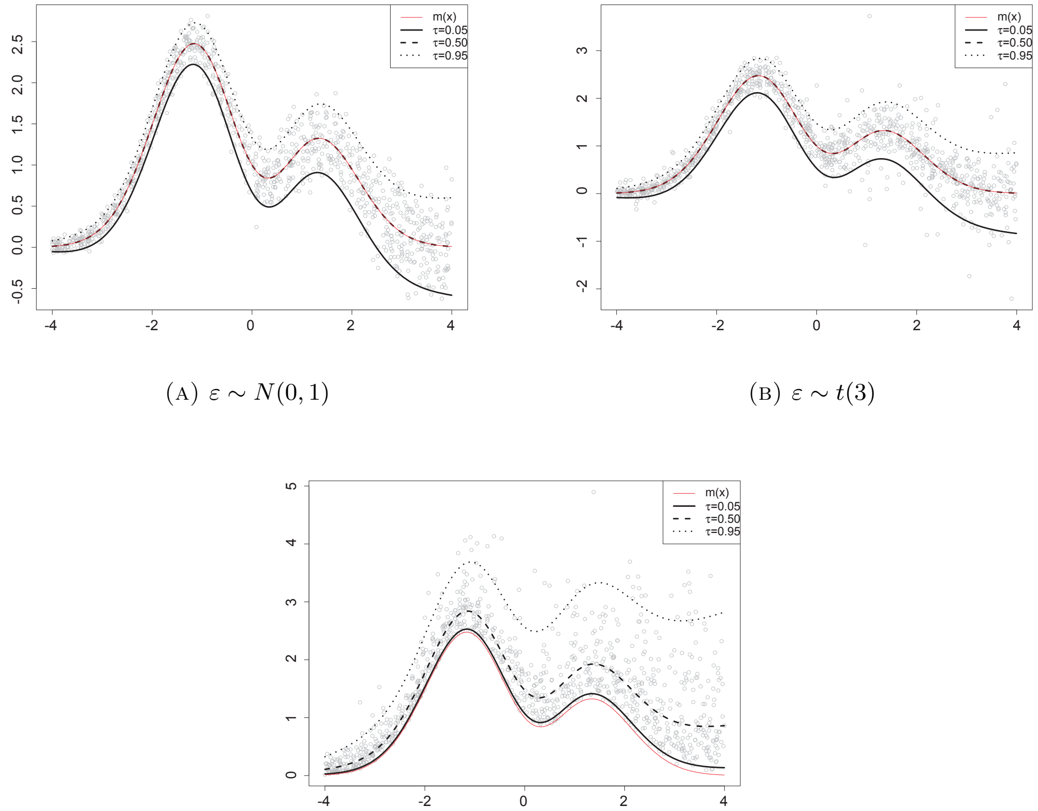
$$\sigma(X) = \frac{1}{5}(1 + 0.2x).$$

Therefore, Eq.(3.19) is a heteroskedastic model.

In this simulation, we consider three different types of random errors for ε: $N(0, 1)$, $t(3)$ and $\chi^2(3)$, respectively. Therefore, the true τ-th conditional quantile function of Y given $X = x$ can be expressed as

$$Q_Y(\tau|x) = m(x) + \sigma(x)F_\tau^{-1}(\varepsilon),$$

where $F_\tau^{-1}(\varepsilon)$ is the τ-th quantile of ε. Fig. 3.4 presents the training data generated under this scenario with their true τ-th conditional quantile functions $Q_Y(\tau|x), \tau \in c(0.05, 0.50, 0.95)$. Note that, the nonlinear function $m(X)$ in the right figure is not identical to the true conditional median function $Q_Y(0.50|x)$ as the random error $\chi^2(3)$ is an asymmetric distribution.



(A) $\varepsilon \sim N(0,1)$

(B) $\varepsilon \sim t(3)$

(C) $\varepsilon \sim \chi^2(3)$

FIGURE 3.4: Simulated training data and true conditional quantile functions with $\tau \in c(0.05, 0.50, 0.95)$.

We aim to compare the prediction power of the above-mentioned four methods for the prediction of the conditional quantile function by 500 test data sets, in terms of three measurements, namely, the root mean square error (RMSE), the mean absolute errors (MAE), and the Theil-U statistic, which is a relative accuracy measure that compares the forecast results with the naïve forecast (Theil, 1966):

$$RMSE(\tau) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(Q_{Y_i}(\tau|x) - \hat{Q}_{Y_i}(\tau|x)\Big)^2},$$

$$MAE(\tau) = \frac{1}{n}\sum_{i=1}^{n}\Big|Q_{Y_i}(\tau|x) - \hat{Q}_{Y_i}(\tau|x)\Big|,$$

$$TheiU(\tau) = \sqrt{\frac{\sum_{i=2}^{n}\left(\frac{\hat{Q}_{Y_i}(\tau|x)-Q_{Y_i}(\tau|x)}{Q_{Y_{i-1}}(\tau|x)}\right)^2}{\sum_{i=2}^{n}\left(\frac{Q_{Y_i}(\tau|x)-Q_{Y_{i-1}}(\tau|x)}{Q_{Y_{i-1}}(\tau|x)}\right)^2}},$$

where $\hat{Q}_{Y_i}(\tau|x)$ is the prediction of the true conditional quantile $Q_{Y_i}(\tau|x)$. The smaller the measurement value is, the better the method is. The three measurements are implemented with R function *av.res* in package *AnalyzeTS*.

The superiority of the proposed normal-scale mixture approach is demonstrated in Table 3.4 which summarizes the results for three values of τs: 0.05, 0.50, and 0.95, based on the 500 replications. Note that, Simulation 2 is implemented with critical values simulated from $ALD(0,1,\tau)$ (coincide with the likelihood) with α=0.25, r=0.5 and η= (5,7,10,13,17,21,24,28,36,45)/365. The bold face values show that both SWH's method and the proposed normal scale-mixture approach are superior to LPQR and COBS, while the proposed approach performs slightly better than SWH. It is encouraging to see that the proposed approach approximates well under Gaussian error and also provides excellent results under the circumstance of heavy tail and asymmetric distributions, such as $t(3)$ and $\chi^2(3)$.

Table 3.4: Average value of the evaluation indices for 500 test data of size 500.

	$\varepsilon \sim N(0,1)$				$\varepsilon \sim t(3)$				$\varepsilon \sim \chi^2(3)$			
Indices	LPQR	COBS	SWH	NSM	LPQR	COBS	SWH	NSM	LPQR	COBS	SWH	NSM
$\tau = 0.05$												
RMSE	0.364	0.254	0.168	**0.157**	0.399	0.274	0.226	**0.213**	0.432	0.239	0.162	**0.154**
MAE	0.234	0.176	0.128	**0.121**	0.273	0.205	0.173	**0.163**	0.269	0.162	0.121	**0.116**
Thei U	17.773	12.414	8.196	**7.667**	19.293	13.264	10.896	**10.286**	20.974	11.640	7.863	**7.480**
$\tau = 0.5$												
RMSE	0.178	0.184	0.163	**0.140**	0.184	0.172	0.141	**0.139**	0.210	0.198	0.176	**0.170**
MAE	0.140	0.144	0.128	**0.114**	0.144	0.137	0.107	**0.103**	0.171	0.161	0.139	**0.132**
Thei U	8.524	8.865	7.839	**7.131**	8.942	8.403	6.875	**6.741**	10.246	9.695	8.587	**8.241**
$\tau = 0.95$												
RMSE	0.258	0.210	0.159	**0.157**	0.283	0.245	0.205	**0.195**	0.367	**0.324**	0.331	0.326
MAE	0.193	0.153	0.125	**0.123**	0.226	0.190	0.162	**0.153**	0.272	0.261	**0.250**	0.261
Thei U	12.507	10.176	7.735	**7.600**	8.983	9.553	**6.862**	7.570	16.743	**14.798**	15.159	14.852

Note: The bandwidth h_τ at τ that controls the complexity of the LPQR model is selected by the rule of thumb in Fan and Gijbels (1996).

3.5.3 Real-world data application

In this section we demonstrate the efficacy of our the proposed alternative approach with one benchmark example that comes from the second and third health examination surveys of the USA (National Center for US Health Examination Surveys, 1970; 1973). Taken together these provide data on the anthropometry of children between the ages of 6 years and under 18 years, with from 400 to 600 children of each sex seen in each year of age (Cole, 1998). Here, along with Yu and Jones (1998), the weights and ages of 4011 US girls were analysed.

Figure 3.5 displays weight against age for a sample of 4011 US girls, where age is a univariate regressor $X \in R^1$ for simplicity. From Figure 3.5, it is evident that the distribution is left-skewed and presents long tails, suggesting that focusing on the centre is not sufficient for a comprehensive description of a weight distribution. Such observation motivates the use of quantile regression, where a complete picture of weight distribution is captured by conditional quantiles.

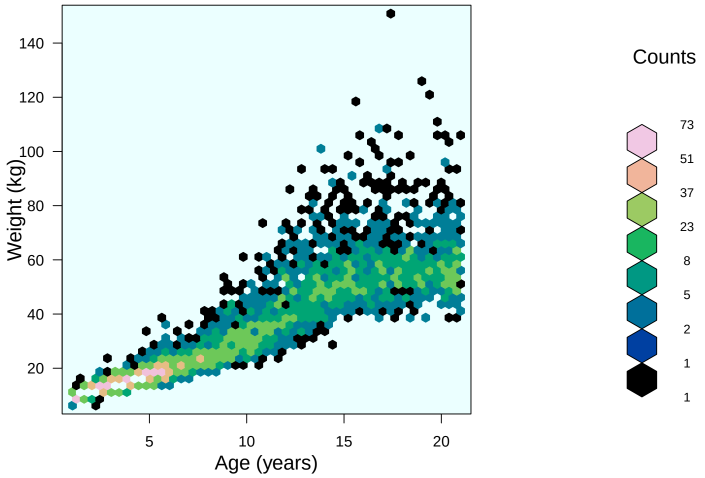
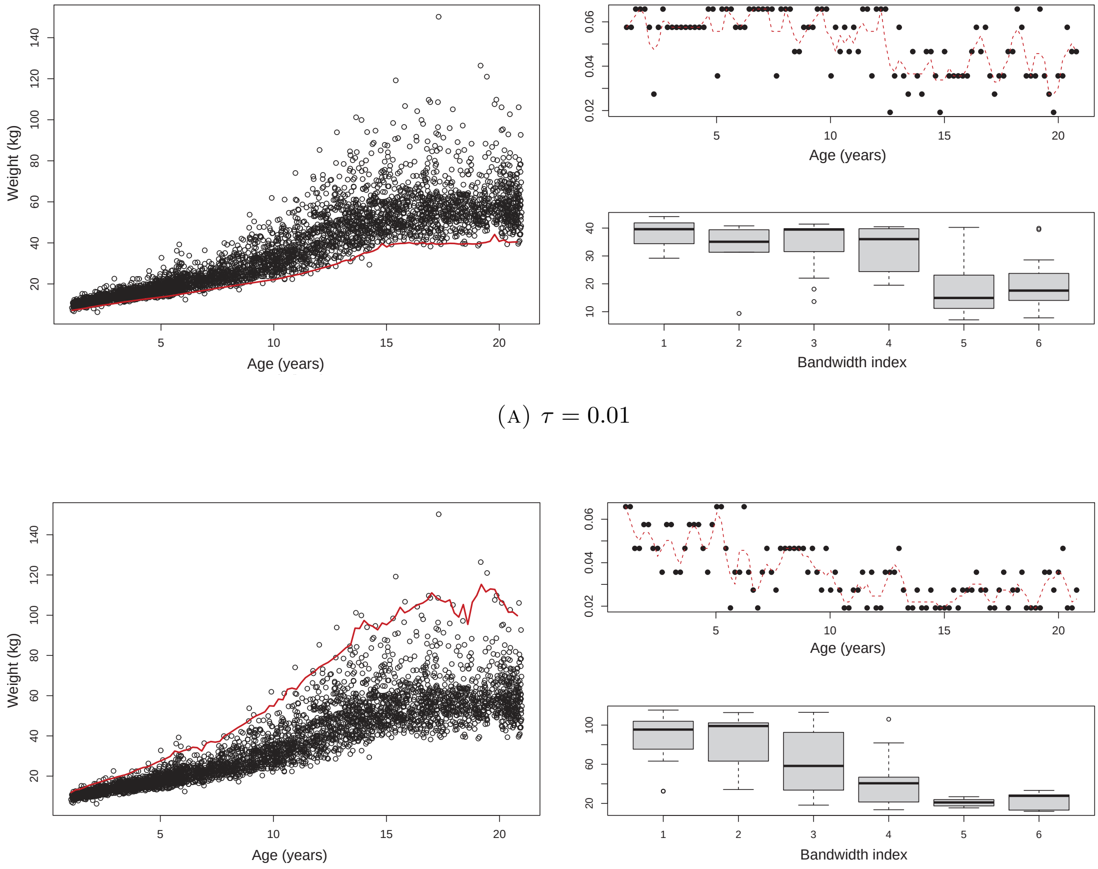


Figure 3.5: US Health Examination Surveys data : weight against age for a sample of 4011 US girls.

We then continue by inspecting the relation between weight and age in the sample. In Figure 3.6, we display the bandwidth sequence (upper right panels), boxplot of adapted bandwidth (lower right panels) showing the relationship between the adapted estimator and the bandwidth index, and smoothed quantile curves for quantile 99% (3.6b) and 1% (3.6a) respectively by using the alternative normal scale-mixture likelihood function. Both adaptations show that the proposed bandwidth selection is well-adapted over the data distribution, which provides smooth fitting and better adaptation when τ tends to extreme quantiles. Furthermore, Figure 3.7 shows that the non-quantile crossing property holds for the rule in Section 3.2, which is based on the alternative normal scale-mixture likelihood function.



(A) $\tau = 0.01$

(B) $\tau = 0.99$

Figure 3.6: Smoothed quantile curves for US Health Examination Surveys with $\tau = 0.01$ and $\tau = 0.99$ via alternative normal scale-mixture likelihood. The bandwidth sequence (upper right); boxplot of block residuals adaptive bandwidth (lower right).

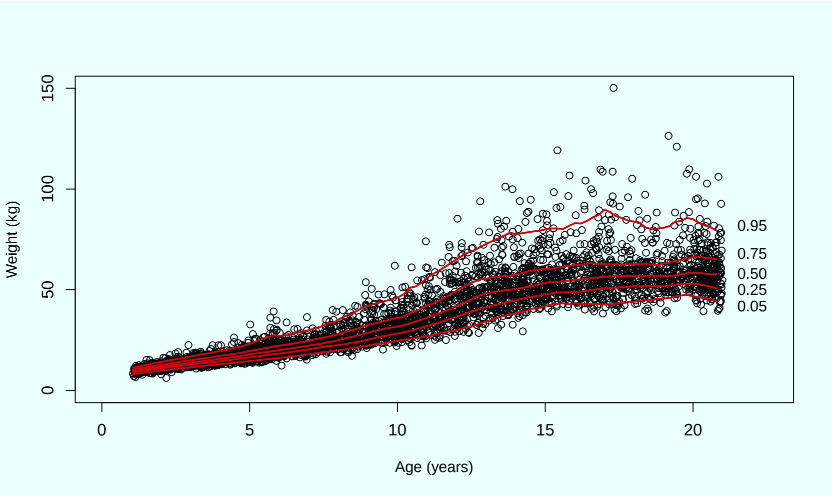


Figure 3.7: Smoothed quantile curves for US Health Examination Surveys with $\tau = c(0.05, 0.25, 0.5, 0.75, 0.95)$ via alternative normal scale-mixture likelihood function.

3.6 Chapter Summary

The kernel-weighted likelihood function (3.5) in SWH's paper is a local ALD-based likelihood function. The ALD-based inference has nowadays become a powerful tool for formulating different quantile regression techniques, particularly for the development of different Bayesian inference techniques for quantile regression. The ALD-based inference for non-Bayesian methods includes Taylor and Yu (2016b) in financial risk analysis, Geraci and Bottai (2007) in longitudinal data analysis and among others.

The local ALD-based likelihood approach in this Chapter uses an alternative ALD-type of likelihood. The resulting automatic bandwidth selection rule not only enjoys the propagation condition of SWH (which postulates that the risk is smaller than the upper bound for the risk of the estimator $\tilde{\boldsymbol{\theta}}_k(x)$) but also guarantees non-quantile curve crossing. Theoretical results also claim that the proposed adaptive procedure performs well, which would minimize the local estimation risk for the problem at hand. We illustrate the performance of the procedure by comparing the Lidar dataset with SWH's approach and analyzing an extended real data application. In particular, we show that the performance of the adaptive procedure is promising in practice, especially for smoothing extreme quantile curves.

Chapter 4

Modelling Tails for Collinear Data with Outliers: Quantile Profile Regression

In this chapter we present a statistical approach to distinguish and interpret the complex relationship between several predictors and the tail of the distribution of a response variable in the presence of high correlation between the predictors.

Covariates which are highly correlated create collinearity problems when used in a standard multiple regression model. Many methods have been proposed in the literature to address this issue. A very common approach is to create an index which aggregates all the highly correlated variables of interest, but it is more informative to look specifically at each predictor to better understand their roles in the statistical analysis. In this paper we illustrate how the complex relationships between the predictors can be de-constructed and analysed using profile regression, a Bayesian non-parametric model for clustering responses and covariates simultaneously. While profile regression is a powerful tool to model the relationship between a response variable and covariates, there is no guarantee that the standard approach of using a mixture of Gaussian distributions for the response model will identify the underlying clusters correctly. In particular, the interest in many practical problems lie in the tails of an asymmetric response, such as obesity in the case of weight distribution, or the number of patients with high glucose levels. In this chapter, we address this by modelling the response variable with an asymmetric Laplace distribution, allowing us to model more accurately for clusters which are asymmetric and predict more accurately for extreme values of the response variable and/or outliers.

Our novel mixture modelling approach is demonstrated on both simulated and real data. In these analyses, our method performs more accurately when compared to the Gaussian mixture model for profile regression.

4.1 Introduction

A well known issue in many statistical applications, when trying to assess meaningful relationships between predictors and response variables through regression models, is the potential multicollinearity of the predictors. A common approach in this case is to examine each predictor separately, to avoid instability in the estimates, but compromising the possibility of learning about the complex relationships involving several predictors at the same time. An alternative approach is to combine the correlated variables into summary indexes and to assess the relationship of these with the outcome of interest, but this approach loses information on the single variables included in the summary.

Dirichlet process mixture models have been proposed as alternatives to regression models when dealing with multicollinearity (Dunson et al., 2008; Molitor et al., 2010). In particular we will use profile regression, a semi-parametric Bayesian method where covariate profiles are allocated into clusters and associated via a regression model with a relevant outcome. This method was implemented by Liverani et al. (2015) in the R package *PReMiuM* and applied in a variety of areas, including, for example, environmental epidemiology (Pirani et al., 2015; Liverani et al., 2016) and genetics (Papathomas et al., 2012).

In this chapter we propose a new profile regression model with a quantile regression submodel to allow a careful modelling of the data when the interest is on lower or upper tails of the distribution of the response profiles rather than their mean. We propose a new quantile regression model. Quantile regression models were first introduced by Koenker and Bassett (1978) and have been applied to a wide range of applications in biostatistics, including survival analysis, ecology, earnings inequality and mobility, income and wealth distribution, value at risk and mutual fund investment styles (Knight and Ackerly, 2002; Geraci and Bottai, 2007). Quantile regression models aim at estimating either the conditional median or other quantiles of the response variable. Their main advantage over least-squares regression is their flexibility for modelling data with heterogeneous conditional distributions. Moreover, quantile regression models are robust to outliers.

We propose a new method that we named 'quantile profile regression'. Quantile profile regression includes a Bayesian mixture of asymmetric Laplace distributions (ALD), which were proposed by Yu and Moyeed (2001) for Bayesian inference for quantile regression based on a 'working likelihood'. The closest work to ours are by Kottas and Krnjajić (2009) and Taddy and Kottas (2010), which constructed general classes of semiparametric and nonparametric distributions for likelihood using Dirichlet process mixture models. However, Kottas and Krnjajić (2009)'s interests were in the error distribution of a quantile regression and their mixture over the scale parameters, while Taddy and Kottas (2010) used DPMM for the joint distribution of the response and covariates but focused on the estimation of the regression parameters. Different from the above methods, in this paper we are interested in using a mixture of ALDs for the response which links covariate profiles to clusters (allocation parameters) and other poosible fixed factors via a 'regression' model. The model is not a standard direct regression function of covariates, but it allows the complex relationships between predictors and the response variables to be explained.

Another close proposal is by Franczak et al. (2014). They proposed the use of shifted asymmetric Laplace distributions for model-based clustering and provided an Expectation Maximisation algorithm. Their mixture model was multivariate and aimed at classical classification problems, and for certain selected examples they outperform the Gaussian mixture models. They did not study the potential relationship between predictors and covariates, while we assume the distribution of the response variable and the covariates to be cluster dependent.

The inference for quantile profile regression is carried out by Markov chain Monte Carlo (MCMC) using the conjugate Gibbs sampler for different quantiles. Our novel mixture modelling approach is demonstrated on both simulated and real data. In these analyses, our mixture of asymmetric Laplace distributions performs favourably when compared to the Gaussian mixture model for profile regression.

The chapter is organised as follows. Section 4.2 describes the Dirichlet process mixture model for Bayesian clustering. Profile regression employing a likelihood function that is based on the asymmetric Laplace distribution is developed in Section 4.3. Section 4.4 and 4.5 conduct a simulation study and a real data example respectively to illustrate the novel approach and compare the results to normal profile regression.

4.2 Dirichlet Process Mixture Model

Dirichlet process mixture models are defined for data $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$, regarded as exchangeable or as independently drawn from an unknown distribution. This distribution is modelled as a mixture of distributions of the form $F(\theta)$, with the mixing distribution over θ being G. The prior for this mixing distribution is a Dirichlet process with concentration parameter α and base distribution G_0 (Ferguson, 1973).

$$\begin{aligned} Y_i|\theta_i &\sim F(\theta) && (4.1)\\ \theta_i|G &\sim G && (4.2)\\ G &\sim DP(G_0, \alpha). && (4.3) \end{aligned}$$

An infinite mixture model will not face the misspecification of parameters in contrast to finite models, especially when using a model structure which is far from the real one, and hence will generate more stable solutions.

4.2.1 Profile Regression

We will focus on the Dirichlet process mixture model described in Liverani et al. (2015). This model links a response vector $\mathbf{Y}$ with the covariate vector $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_d})$ non-parametrically through clustering. Also, the approach enables the potential supplemental fixed effects $\mathbf{W}$, which have a global effect on the response. It is worth noting that the allocated clusters are based on the joint effects of $\mathbf{X}$ and $\mathbf{Y}$, implicitly handling latent high dimensional interactions which would be quite challenging to capture via classical approaches.

Consider a response variable Y_i and a covariate profile $X_i = (x_{i,1}, \ldots, x_{i,d})$ for i in $1, 2, \ldots, n$. The observed data follows an infinite mixture distribution, where mixture component c has density conditional on some component specific parameters $\mathbf{\Theta}_c$ and global parameters $\mathbf{\Lambda}$. Therefore, the proposed model is given by the a joint probability model for the outcome Y_i and profile X_i, where these probability models are conditionally independent:

$$f(Y_i, X_i|\mathbf{\Theta}, \mathbf{\Lambda}, \mathbf{W}_i) = \sum_{c=1}^{\infty} \psi_c f(Y_i|\mathbf{\Theta}_c, \mathbf{\Lambda}, \mathbf{W}_i) f(X_i|\mathbf{\Theta}_c, \mathbf{\Lambda}) \tag{4.4}$$

where $\mathbf{\Theta} = (\psi_1, \mathbf{\Theta}_1, \psi_2, \mathbf{\Theta}_2, \cdots)$, and the weight of mixture component c is given by ψ_c. The mixture weights as $\boldsymbol{\psi} = \{\psi_c, c \geq 1\}$ follow a stick breaking distribution which is given by

$$\begin{aligned}
\psi_c &= V_c \prod_{l<c}(1 - V_l) \quad \text{for } c \in \mathbb{Z}^+ \setminus \{1\} && (4.5)\\
\psi_1 &= V_1 && (4.6)\\
V_c &\sim \text{Beta}(1, \alpha) \quad \text{i.i.d. for } c \in \mathbb{Z}^+. && (4.7)
\end{aligned}$$

In order to make inference using mixture models, it is common and convenient to bring in a vector of latent allocation variables $\mathbf{Z} = (Z_1, \ldots, Z_n)$, such that $Z_i = c$ identifies the allocation of individual i to cluster c. Posterior inference on $\mathbf{Z}$ then offers us with information concerning the clustering of the observations. The inference is carried out via Markov Chain Monte Carlo using the stick-breaking construction of the Dirichlet process and the slice sampler.

There is a wide range of choices for the response sub-model $f(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i)$ and the profile sub-model $f(X_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda})$, including normal, Bernoulli, Binomial, Poisson, Multinomial and Weibull distributions (Liverani et al., 2015). We refer to the profile regression model with the normal distribution for the response variable as 'normal profile regression'.

The blocked infinite DPMM algorithm can now be defined using the following blocked Gibbs updates to sample from the relevant conditionals. This sampler (Liverani et al., 2015) uses a combination of Gibbs and Metropolis-within-Gibbs steps to sample from the infinite mixture (only retaining the parameters of a finite number of mixture components including all those to which individuals are allocated at each sweep).

Suppose we are at sweep t of the sampler. Update as follows:

1. Compute Z^* and the set A.

2. Sample $(\mathbf{V}^A_{t+1}, \mathbf{\Theta}^A_{t+1}, \tilde{\mathbf{Z}}, \mathbf{U}_{t+1}) \sim p(\mathbf{V}^A, \mathbf{\Theta}^A, \mathbf{Z}, \mathbf{U}|\mathbf{V}^P_t, \mathbf{V}^I_t, \mathbf{\Theta}^P_t, \mathbf{\Theta}^I_t, \alpha_t, \Lambda_t, \Theta_0, \mathbf{D})$.

    (a) $\tilde{\mathbf{V}}^A \sim p(\mathbf{V}^A|\mathbf{Z}_t, \alpha_t)$

    (b) $\tilde{\mathbf{\Theta}}^A \sim p(\mathbf{\Theta}^A|\mathbf{Z}_t, \Lambda_t, \Theta_0, \mathbf{D})$

    (c) $(\mathbf{V}^A_{t+1}, \mathbf{\Theta}^A_{t+1}, \tilde{\mathbf{Z}}) \sim p(\mathbf{V}^A, \mathbf{\Theta}^A, \mathbf{Z}|\mathbf{V}^P_t, \mathbf{\Theta}^P_t, \alpha_t, \Lambda_t, \Theta_0, \mathbf{D})$

    (d) $\mathbf{U}_{t+1} \sim p(\mathbf{U}|\mathbf{V}^A_{t+1}, \tilde{\mathbf{Z}})$

3. Compute U^*. Recompute Z^* and the set A.

4. Sample $(\alpha_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{V}^I_{t+1}) \sim p(\alpha, \mathbf{V}^P, \mathbf{V}^I|\mathbf{\Theta}^P_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{\Theta}^A_{t+1}, \mathbf{\Theta}^I_t, \mathbf{U}_{t+1}, \tilde{\mathbf{Z}}, \Lambda_t, \Theta_0, \mathbf{D})$, computing C^* and the set P in the process.

    (a) $\alpha_t \sim p(\alpha|\mathbf{V}^A_{t+1}, \tilde{\mathbf{Z}})$

    (b) $\mathbf{V}^P_{t+1} \sim p(\mathbf{V}^P|\alpha_{t+1}, \mathbf{U}_{t+1}, \tilde{\mathbf{Z}})$

5. Sample $(\mathbf{\Theta}^P_{t+1}, \mathbf{\Theta}^I_{t+1}) \sim p(\mathbf{\Theta}^P, \mathbf{\Theta}^I|\mathbf{V}^A_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{V}^I_{t+1}, \mathbf{\Theta}^A_{t+1}, \mathbf{U}_{t+1}, \tilde{\mathbf{Z}}, \Lambda_t, \Theta_0, \mathbf{D})$.

    (a) $\mathbf{\Theta}^P_{t+1} \sim p(\mathbf{\Theta}^P|\Theta_0)$

6. Sample $\Lambda_{t+1} \sim p(\Lambda|\mathbf{V}^A_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{V}^I_{t+1}, \mathbf{\Theta}^A_{t+1}, \mathbf{\Theta}^P_{t+1}, \mathbf{\Theta}^I_{t+1}, \mathbf{U}_{t+1}, \tilde{\mathbf{Z}}, \Lambda_t, \Theta_0, \mathbf{D})$.

    (a) $\Lambda \sim p(\Lambda|\mathbf{\Theta}^A_{t+1}, \tilde{\mathbf{Z}}, \mathbf{D})$

7. Sample $\mathbf{Z}_{t+1} \sim p(\mathbf{Z}|\mathbf{V}^A_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{V}^I_{t+1}, \mathbf{\Theta}^A_{t+1}, \mathbf{\Theta}^P_{t+1}, \mathbf{\Theta}^I_{t+1}, \mathbf{U}_{t+1}, \alpha_{t+1}, \Lambda_{t+1}, \Theta_0, \mathbf{D})$.

    (a) $\mathbf{Z}_{t+1} \sim p(\mathbf{Z}|\mathbf{V}^A_{t+1}, \mathbf{V}^P_{t+1}, \mathbf{\Theta}^A_{t+1}, \mathbf{\Theta}^P_{t+1}, \mathbf{U}_{t+1}, \Lambda_{t+1}, \mathbf{D})$

Note that $\mathbf{U} = (U_1, U_2, \cdots, U_n)$ are introduced auxiliary variables and A, P and I are disjoint sets that partition $\mathbb{Z}^+$, which are elaborated in the Appendix C. The key idea is that by doing joint updates, we can marginalise out an infinite number of variables when necessary, to ensure that we are always sampling from conditional distributions that depend only upon a finite number of parameters. Moreover, although the sampler is written as a blocked Gibbs sampler, where it is not possible to sample directly from full conditionals (for example in the update of $\mathbf{\Theta}$, depending upon the choices of f and P_{Θ_0}), Metropolis-within-Gibbs steps are applied.

Due to the problem of "label switching", i.e the labels associated with each cluster change during the MCMC iterations, we cannot simply assign each observation to the cluster that maximizes the average posterior probability. Methods that deal with label switching, like the relabelling algorithm of Stephens (2000), require the number of clusters K to be fixed. Using the Dirichlet process mixture models, we allow the number of clusters to vary from one MCMC sample to the next. One possible solution is to choose the partition based on a posterior similarity matrix. At each iteration of the sample, we record pairwise cluster membership and construct a score matrix, with entries equal to 1 for pairs belonging to the same cluster and 0 otherwise. Averaging these matrices over the whole MCMC run leads to a similarity matrix S, which can be then used to identify an optimal partition.

4.3 Quantile Profile Regression

We extend profile regression to allow for asymmetric Laplace distributions for the response variable. We name this model Bayesian profile quantile regression. Let the response sub-model be

$$f(Y_i; \mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) = f(Y_i | \theta_{Z_i}, \beta, \sigma_Y, W_i) = \frac{\tau(1-\tau)}{\sigma_Y} \exp\left\{ -\rho_\tau \left(\frac{Y_i - \lambda_i}{\sigma_Y} \right) \right\} \tag{4.8}$$

that is, $Y_i | Z_i, \mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i \sim \text{ALD}(\lambda_i, \sigma_Y; \tau)$, where $\lambda_i = \theta_{Z_i} + \beta^T W_i$ and $\mathbf{\Lambda} = (\beta, \sigma_Y)$ contains the global parameters and $\mathbf{\Theta} = (\theta_1, \theta_2, \ldots)$ contains the cluster specific parameters. The parameter τ refers to the quantile of interest and it is set, not estimated from the model, depending on the aims of the analysis. For example, if a population of males has the 90% quantile of the weight distribution corresponding to obesity and we aim to investigate how some correlated predictors are related to obesity, then we could set $\tau = 0.9$.

4.3.1 Inference Quantile Profile Regression

We discuss here the details of the sampling from the posterior distribution of the parameters of the ALD. See Liverani et al. (2015) for details of the samplers for all other parameters of the model.

The prior distributions of σ_Y and λ_c are given by

$$\sigma_Y \sim \text{IG}(a, b) \tag{4.9}$$

$$\lambda_c \sim \text{N}(\mu_0, \sigma_0). \tag{4.10}$$

We can derive a Gibbs sampler. The joint posterior distribution of $p(\lambda_i, \sigma_Y|\mathbf{D})$, with $\mathbf{D} = (\mathbf{Y}, \mathbf{X})$, is given by:

$$\begin{aligned} f(\lambda_i, \sigma_Y|\mathbf{D}) &\propto \frac{1}{\sigma_Y} \exp\left\{-\frac{1}{\sigma_Y}(y_i - \lambda_i)(\tau - I(y_i \le \lambda_i))\right\} \\ &\times \frac{1}{\sigma_Y^{a+1}} \exp\left\{-\frac{b}{\sigma_Y}\right\} \times \frac{1}{\sigma_0} \exp\left\{-\frac{(\lambda_i - \mu_0)^2}{2\sigma_0^2}\right\} \\ &\propto \frac{1}{\sigma_Y^{a+2}} \exp\left\{-\frac{1}{\sigma_Y}(b + (y_i - \lambda_i)(\tau - I(y_i \le \lambda_i)))\right\} \exp\left\{-\frac{(\lambda_i - \mu_0)^2}{2\sigma_0^2}\right\} \end{aligned}$$

where the prior distribution of σ_Y is given by $\sigma_Y \sim \text{IG}(a, b)$, and the prior distribution of λ_c is $\lambda_c \sim \text{N}(\mu_0, \sigma_0)$. From the joint posterior distribution, we derive the full conditional density of σ_Y which, conditional on λ_i, is proportional to

$$\frac{1}{\sigma_Y^{a+2}} \exp\left\{-\frac{1}{\sigma_Y}(b + (y_i - \lambda_i)(\tau - I(y_i \le \lambda_i)))\right\}$$

so that we have

$$\sigma_Y|\mathbf{D}, \lambda_i \sim \text{IG}(a + 1, b + (y_i - \lambda_i)(\tau - I(y_i \le \lambda_i))) \tag{4.11}$$

The marginal posterior distribution of λ_i, conditional on σ_Y is given by

$$\begin{aligned} f(\lambda_i|\mathbf{D}, \sigma_Y) &\propto (y_i - \lambda_i)(\tau - I(y_i \le \lambda_i)) + \frac{1}{\sigma_0^2}\left(\lambda_i^2 - 2\lambda_i\mu_0\right) \\ &\propto \begin{cases} -\lambda_i\tau + \frac{1}{\sigma_0^2}\left(\lambda_i^2 - 2\lambda_i\mu_0\right) & \text{if } y_i \ge \lambda_i \\ \lambda_i(1 - \tau) + \frac{1}{\sigma_0^2}\left(\lambda_i^2 - 2\lambda_i\mu_0\right) & \text{if } y_i < \lambda_i \end{cases} \\ &\propto \begin{cases} \frac{1}{\sigma_0^2} - \left(\tau + \frac{2\mu_0}{\sigma_0^2}\right)\lambda & \text{if } y_i \ge \lambda_i \\ \frac{1}{\sigma_0^2} + \left(1 - \tau - \frac{2\mu_0}{\sigma_0^2}\right)\lambda & \text{if } y_i < \lambda_i \end{cases} \end{aligned}$$

Therefore, the Gibbs sample for the posterior of λ_i can be sampled from the normal distribution as follows:

$$\lambda_i|\mathbf{D}, \sigma_Y \sim \begin{cases} \mathrm{N}\left(\mu_0 + \frac{\tau\sigma_0^2}{2}, \sigma_0^2\right) \cdot I(y_i \geq \lambda_i) & \text{if } y_i \geq \lambda_i \\ \mathrm{N}\left(-\mu_0 + \frac{(1-\tau)\sigma_0^2}{2}, \sigma_0^2\right) \cdot I(y_i < \lambda_i) & \text{if } y_i < \lambda_i \end{cases} \tag{4.12}$$

4.3.2 Posterior Predictive Distribution

A common target of inference is not necessarily the partition itself, but how the estimated parameters might allow us to make predictions for future observations. For example we might want to group new observations with existing observations, or, in the case of profile regression, make a prediction about the response if only the covariates of a new observation had been observed. One way to do this is to use posterior predictions, where posterior predictive distributions for quantities of interest can be derived from the whole MCMC run, taking the uncertainty over clustering into account.

As for profile regression, we can sample the posterior predictive distribution of pseudo-profiles. The pseudo-profiles are predictive scenarios determined by the covariates. At each iteration the predictive subjects are allocated to one of the existing clusters in accordance with their own covariate profiles. We can then derive the posterior predictive distribution of the response variable for each pseudo-profile. To compute these distributions, we implemented with computing the full posterior predictive distribution of the pseudo-profiles and then identified the quantiles of interest, which is detailed below.

We compute the posterior probability $p(\tilde{Z}_s^r = c|X_s, \mathbf{\Theta}_r, Y_i, X_1, ..., X_N)$ for each pseudo-profile, where $\tilde{Z}_s^r$ corresponds to each predictive scenario s at each sweep r of the MCMC sampler. With these probabilities we construct a cluster-averaged estimate of θ for each particular pseudo-profile at each sweep. Specifically,

$$\hat{\theta}_s^r = \sum_{c=1}^{\infty} p(\tilde{Z}_s^r = c|X_s, \mathbf{\Theta}_r, Y_i, X_1, ..., X_N)\theta_c^r. \tag{4.13}$$

Looking at the density of these predictions over MCMC sweeps gives us an estimate of the effect of a particular pseudo-profile, and its comparison to other pseudo profiles, allowing us to derive a better understanding of the role of specific covariates.

4.4 Simulation Study

In this section we provide the results of the implementation of quantile profile regression for simulated data. First we simulate data from the ALD and show that the method proposed is more effective than normal profile regression to retrieve generating parameters for asymmetric data. Then we show that if we are interested in a specific quantile of the distribution, even when the generating mechanism is Gaussian, quantile profile regression makes more accurate predictions than normal profile regression.

Figure 4.1 shows the first set of simulated data. Five clusters were generated by drawing independent samples from the asymmetric Laplace distribution (ALD) for the outcome Y and the normal distribution for the covariate X as follows.

$$Y_i \sim \text{ALD}(\theta_{Z_i} + \beta^T W_i, \sigma_Y; q) \tag{4.14}$$

$$X_i \sim \text{normal}(\mu_{Z_i}, \gamma^2_{Z_i}) \tag{4.15}$$

with $i = 1, 2, \ldots, 2300$. As the profile sub-model $p(X_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda})$ is Gaussian with parameters μ_{Z_i} and $\gamma^2_{Z_i}$, the cluster-specific parameters contained in $\mathbf{\Theta}$ are $(\theta_1, \theta_2, \ldots, \theta_5) = (-200, 0, 3, 40, 150)$, $(\mu_1, \mu_2, \ldots, \mu_5) = (0, 6, -8, -3, 5)$ and $(\gamma^2_1, \gamma^2_2, \ldots, \gamma^2_5) = (6, 7, 4, 10, 17)$. When the observation i belongs to cluster c the allocation variable $Z_i = c$. The sizes of the five simulated clusters were 600, 200, 400, 300 and 800 observations respectively. The coefficients β were set equal to 0, therefore omitting the fixed effects. We set $\sigma_Y = 1$.

We use quantile profile regression and normal profile regression, implemented in the R package *PReMiuM*. We set the same priors for both models and keep the hyperparameters constant. The parameters θ_c have a t-distribution with 7 degrees of freedom, mean 0 and scale 2.5. The shape and scale of σ_Y and γ^2 are 2.5 and 2.5. The prior on the mean vector for μ_c has the empirical covariate means as mean and the inverse of the diagonal matrix with elements equal to square of empirical range for each covariate, multiplied by the number of covariates, as precision matrix. The Gamma prior on the Dirichlet parameter α has a shape parameter of 2 and rate of 1. For all simulations below, we ran 20,000 iterations of burn-in and 20,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters (not shown). See Hastie et al. (2015) for more details on convergence for this type of model.

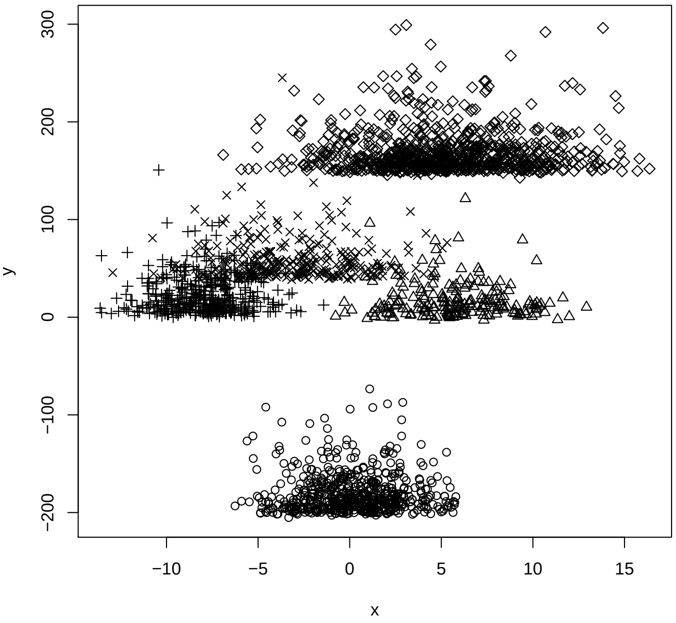


FIGURE 4.1: Data simulated by Eq.(4.14) and (4.15). The five generating clusters can be identified by the different symbols used for the data points.

We initially simulated the ALD data with $q = 0.05$ and run the algorithm using different settings of the parameter τ. Table 4.1 shows the mean of the posterior distributions of θ applying the proposed quantile profile regression (with $\tau = 0.05$ and 0.95) and normal profile regression, as well as commonly-used clustering methods, such as classification and regression trees (CART) and Density-based spatial clustering of applications with noise (DBSCAN). The first row gives the generating values of the parameter θ for the five clusters. The second row gives the posterior means for the five clusters obtained applying quantile profile regression with parameter $\tau = 0.05$, the third row applying normal profile regression, the fourth row applying quantile profile regression with parameter $\tau = 0.95$, and the last two rows applying CART and DBSCAN. For the fourth, six clusters were identified by the method, while only three clusters were identified by DBSCAN.

Quantile profile regression provides more accurate estimations of the generating parameters. On the other hand, when the data is generated with $q = 0.95$ (simulations and results not shown), the reverse happens and the accuracy is highest for $\tau = 0.95$. As the choice of τ is driven by the application and chosen at priori, we only show results for $q = 0.05$ below, without loss of generality.

Table 4.1: Posterior means of θ.

	1	2	3	4	5	6
θ	-200.00	0.00	3.00	40.00	150.00	
quantile $\tau = 0.05$	-200.01	0.52	5.00	35.39	149.92	
normal	-181.40	16.52	21.72	51.50	167.89	
quantile $\tau = 0.95$	-164.21	28.87	32.97	59.71	89.63	209.15
CART	-57.43	-8.14	26.86	82.78	138.04	
DBSCAN	-180.17	43.32	166.73			

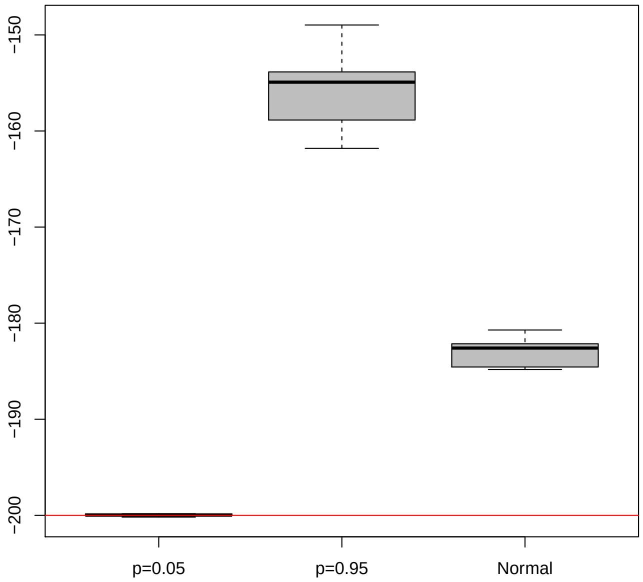


Figure 4.2: Boxplots of the posterior mean of θ_1 over 100 runs for quantile profile regression with $\tau = 0.05$, $\tau = 0.95$ and normal profile regression. The horizontal line marks the generating value $\theta_1 = -200$.

Therefore, as we are in a setting where our interest is in the lowest quantiles of the data, we concentrate on the estimation of θ_1, the parameter of the cluster corresponding to the lowest values of the outcome Y. Figure 4.2 shows the boxplots of the posterior mean of θ_1 over 100 runs of quantile profile regression and normal profile regression against its generating value of -200. Quantile profile regression consistently performs more accurately than the alternative methods. This is not due to the unfair advantage that we know that $q = 0.05$, as Figure 4.3 shows that quantile profile regression outperforms normal profile regression also for $q = 0.1$ and $q = 0.025$.

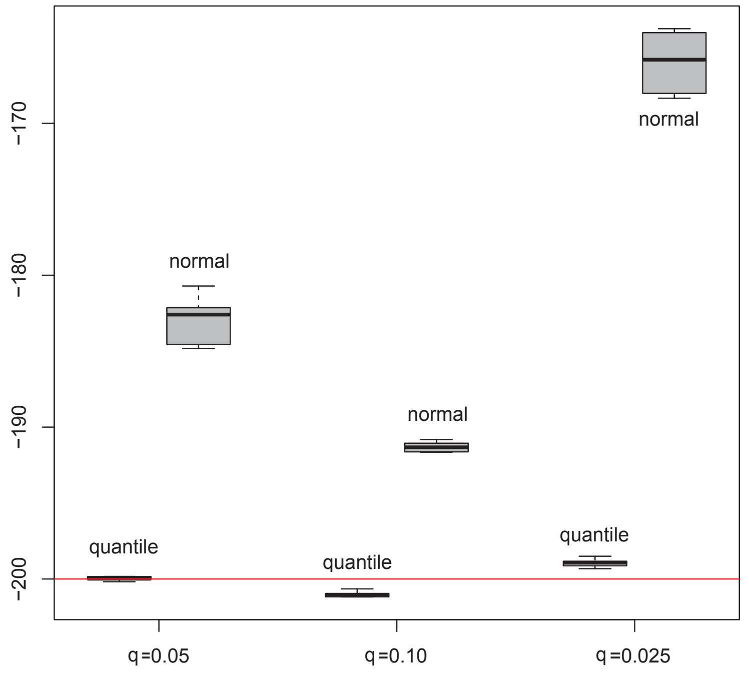


Figure 4.3: Boxplots of the posterior mean of θ_1 over 100 runs for quantile profile regression with $\tau = 0.05$ and normal profile regression, repeated for different generating values of $q = 0.05, 0.10, 0.025$. The horizontal line marks the generating value $\theta_1 = -200$.

Moreover, the results were also robust to the addition of an outlying observation which took the values $x = 15$ and $y = -320$. This is shown in Figure 4.4.

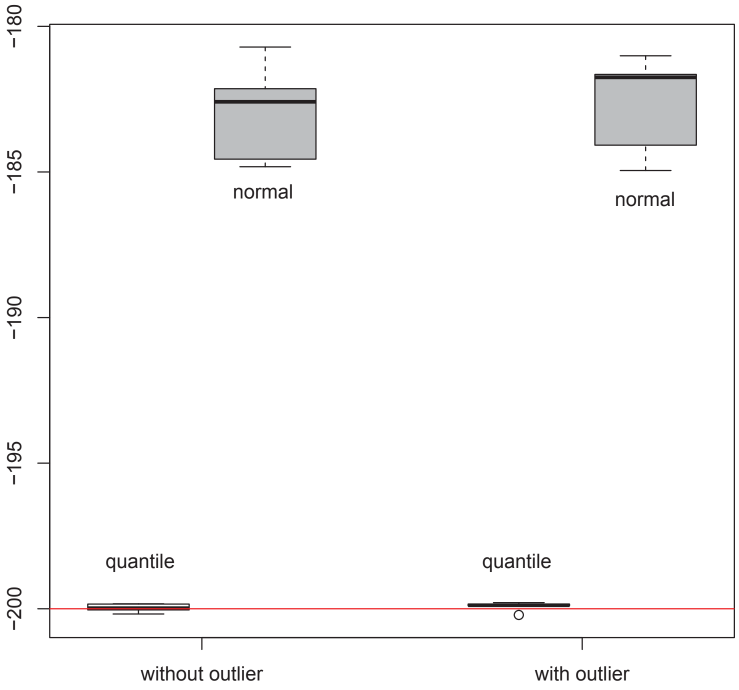


FIGURE 4.4: Boxplots of the posterior mean of θ_1 over 100 runs for quantile profile regression with $\tau = 0.05$ and normal profile regression, comparing the results on the original data and adding an outlier at $x = 15$ and $y = -320$. The horizontal line marks the generating value $\theta_1 = -200$.

We also simulated Y from a normal distribution as follows

$$Y_i \sim \text{normal}(\theta_{Z_i} + \beta^T W_i, \sigma_Y^2) \tag{4.16}$$

$$X_i \sim \text{normal}(\mu_{Z_i}, \gamma_{Z_i}^2) \tag{4.17}$$

with $i = 1, 2, \ldots, 2300$. The cluster-specific parameters contained in $\boldsymbol{\Theta}$ are $(\theta_1, \theta_2, \ldots, \theta_5) = (-6, -2, 0, 3, 6)$, $(\mu_1, \mu_2, \ldots, \mu_5) = (-3, 0, 6, -8, 5)$ and $(\gamma_1^2, \gamma_2^2, \ldots, \gamma_5^2) = (10, 6, 7, 4, 17)$. The sizes of the five simulated clusters were 300, 600, 200, 400 and 800 observations respectively, as for the previous simulation. The coefficients β were set equal to 0, therefore omitting the fixed effects. We set $\sigma_Y^2 = 1$. The data is shown in Figure 4.5. We used the same prior settings as above and for all simulations below we ran 20,000 iterations of burn-in and 20,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters (not shown).

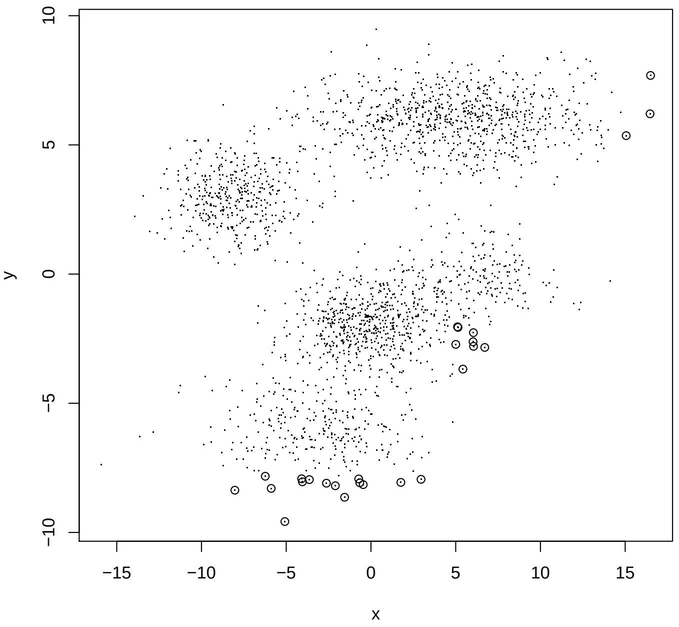


FIGURE 4.5: The data simulated by two uncorrelated normal distributions. The circled and the circled observations are the ones of interest for prediction by quantile profile regression.

As expected, normal profile regression slightly outperforms quantile profile regression when estimating the posterior distribution of θ_c and predicting the outcome Y (Table 4.2).

TABLE 4.2: Posterior means of θ.

	1	2	3	4	5
θ	-6.00	-2.00	0.00	3.00	6.00
quantile $\tau = 0.05$	-5.97	**-1.98**	-0.12	3.03	5.94
normal	**-5.99**	-1.80	**-0.05**	3.13	**6.01**
quantile $\tau = 0.50$	-5.93	-2.10	-0.19	**3.02**	6.03
quantile $\tau = 0.95$	-5.98	-1.94	-0.13	2.94	5.96

However, when the prediction concerns the lowest values of Y, quantile profile regression outperforms normal profile regression. We compare the predictive power of quantile profile regression against normal profile regression using the root mean square error (RMSE) and mean absolute error (MAE) of the predicted values with respect to the observed outcome. These measures of goodness of fit are given by

$$
\begin{aligned}
RMSE &= \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \mathcal{E}^{(M)}(i)\right)^2}{n}} \\
MAE &= \frac{\sum_{i=1}^{n}|Y_i - \mathcal{E}^{(M)}(i)|}{n}
\end{aligned}
\tag{4.18}
$$

where $\mathcal{E}(i)$ denotes the mean for the posterior predictive distribution for Y_i.

Firstly we separate the entire sample into training data and validation data, which is referred as the points crossed in Figure 4.5. We compute the average root mean square error and the average mean absolute error for the prediction of Y for the validation data. They correspond to the lowest observations of Y for different values of X. We carry out quantile profile regression, normal profile regression, CART and standard OLS regression to predict the circled values of Y and repeat this 100 times. Parameters in CART are set by default in R package **'rpart'**. See Table 4.3 for the results. These findings also generalise to the lowest values of Y for different values of X. Quantile profile regression outperforms normal profile regression, CART and standard OLS regression significantly and consistently when predicting in the tails of the distribution.

Table 4.3: Average root mean square error and average mean absolute error over 100 repetitions of the predictions of observations marked by circles in Figure 4.5, obtained using quantile profile regression with $\tau = 0.05$, normal profile regression, CART and standard OLS regression.

	RMSE	MAE
quantile 0.05	**4.02**	**3.09**
normal	5.69	5.03
OLS	5.55	4.86
CART	6.22	5.67

4.5 The English Longitudinal Study of Ageing (ELSA) Analysis

We conducted an analysis with data from the English Longitudinal Study of Ageing (ELSA). ELSA is a longitudinal cohort study of adults aged 50 or older which commenced in 1998, with data collection taking part every two years (Steptoe et al., 2012). The data used in our study are from the nurse visit conducted at Wave 2 of ELSA (2004-2005), a total of 7,666 people took part in this visit where biological data were collected for the first time. The data are available for download from the UK Data Service at http://dx.doi.org/10.5255/UKDA-SN-5050-9.

The aim of the applied portion of our study is to ascertain how cardiometabolic risk factors might be associated with high blood glucose in people who do not currently have a diagnosis of diabetes. Research has shown that high blood glucose levels are an important predictor of incident diabetes (Tabák et al., 2012). However, it has been shown that only considering high blood glucose in prediction of diabetes risk may be overly simplistic as many other cardiometabolic risk factors that are highly correlated with high blood glucose (Haffner et al., 1990; Li et al., 2009) are also associated with diabetes risk (Ford, 2005; Kolberg et al., 2009). Thus we wanted to determine how cardiometabolic risk predictors may cluster together with an outcome of high blood glucose in people who do not have a current diagnosis of diabetes using quantile profile regression modelling. In theory we would expect to see higher levels of blood glucose associated with higher cardiometabolic risk factors.

We removed inaccurate data and missing values (n = 4,474) as we needed all relevant information to conduct our analysis. We also removed data for people with a current diagnosis of diabetes (n = 333), leaving data for 2,859 participants. The variables of interest were: mean systolic blood pressure (‘SYSVAL’), mean diastolic blood pressure (‘DIAVAL’), mean arterial pressure (‘MAPVAL’), cholesterol level (‘CHOL’), high-density lipoprotein level (‘HDL’), triglycerides level (‘TRIG’), low-density lipoprotein level (‘LDL’), C-reactive protein level (CRP: ‘HSCRP’), mean waist (‘WSTVAL’), mean waist/hip ratio (‘WHVAL’) and valid BMI (‘BMIVAL’). See Table 4.4 for summary statistics of the covariates.

Table 4.4: Summary of covariate categories.

	Min	Mean	Max
SYSVAL	80.00	133.74	214.00
DIAVAL	36.50	76.32	118.00
MAPVAL	51.50	95.46	144.00
CHOL	2.10	6.04	12.30
HDL	0.50	1.57	3.40
TRIG	0.40	1.52	4.50
LDL	0.70	3.78	9.20
HSCRP	0.20	3.51	151.00
WSTVAL	61.25	94.15	171.60
WHVAL	0.64	0.88	1.26
BMIVAL	16.02	27.52	55.97

These variables are all quantitative and continuous and they will be included in our model as covariates. Variables used in this analysis are highly correlated, as shown in Table 4.5, thus providing a strong rationale for the use of quantile profile regression. Within our model we adjusted for gender by including it as fixed effect in the model. We opted to focus on the 95% quantile because we wanted to determine how cardiometabolic risk factors might link with high blood glucose.

Table 4.5: Correlation of covariate categories.

	SYSVAL	DIAVAL	MAPVAL	CHOL	HDL	TRIG	LDL	HSCRP	WSTVAL	WHVAL	BMI
SYSVAL	1										
DIAVAL	$0.64^{@}$	1									
MAPVAL	0.89^{+}	0.92^{*}	1								
CHOL	0.08	0.12	0.11	1							
HDL	0.00	0.02	0.02	0.42^{A}	1						
TRIG	0.10	0.12	0.12	0.25	-0.34^{A}	1					
LDL	0.06	0.09	0.08	0.94^{*}	0.22	0.10	1				
HSCRP	0.06	0.02	0.04	-0.04	-0.11	0.05	-0.02	1			
WSTVAL	0.17	0.20	0.21	-0.10	-0.42^{A}	0.31^{A}	-0.06	0.16	1		
WHVAL	0.16	0.14	0.17	-0.13	-0.42^{A}	0.27	-0.08	0.11	$0.78^{@}$	1	
BMIVAL	0.16	0.22	0.21	-0.01	-0.28	0.27	0.00	0.16	$0.79^{@}$	0.35^{A}	1

Note: $A, @, +, *$ denote the correlation between covariates under the significant level 30%, 60%, 80%, 90%, respectively.

We carry out the analysis using quantile profile regression and normal profile regression. The response submodel is given by $p(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) \equiv \text{ALD}(\theta_{Z_i} + \beta^T W_i, \sigma_Y; \tau)$ for quantile profile regression and by $p(Y_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}, \mathbf{W}_i) \equiv \text{normal}(\theta_{Z_i} + \beta^T W_i, \sigma_Y^2)$ for normal profile regression. In both cases the profile sub-model is given by $p(X_i|\mathbf{\Theta}_{Z_i}, \mathbf{\Lambda}) \equiv \text{normal}(\mu_{Z_i}, \gamma_{Z_i}^2)$. We used the same priors as for the simulated data above and our results were not sensitive to the choice of prior. We ran 20,000 iterations of burn-in and 20,000 iterations after that. We obtain good convergence diagnostics on the trace, density and autocorrelation for various parameters.

Quantile profile regression identified four clusters of 1,432, 436, 760 and 231 observations respectively. Figure 4.6 shows boxplots of the posterior distribution for θ_c while Table 4.6 shows the number of observations in each cluster and the posterior mean of θ_c and μ_c for each of the four clusters. We can also look in more detail at the relationship between the response variable and the covariates, as shown in Figure 4.7. We are interested in high values of Y because they correspond to hyperglycemia. In particular the third and fourth clusters have higher than average glucose levels. When examining values of Y, values of blood glucose 5.6 mmol/L are considered as high risk for the development of diabetes (Centers for Disease Control and Prevention, 2011). The blood glucose levels of clusters two, three and four (with credible intervals) were all higher than 5.6 mmol/L indicating these three groups could be considered high risk for diabetes. Figure 4.8 shows that the 95% credible interval for β, which quantifies the linear relationship between gender and the response, is (-0.39, 0.10).

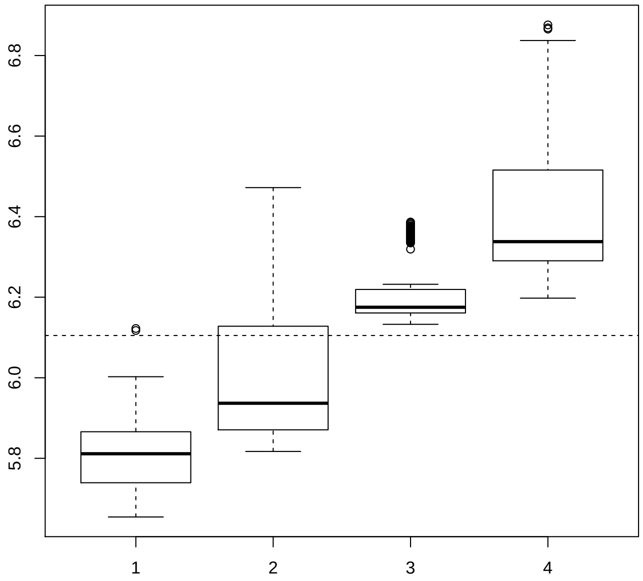


FIGURE 4.6: Boxplots of the posterior distribution of θ_c for the 4 clusters identified by quantile profile regression. The horizontal dashed line is the overall posterior mean.

TABLE 4.6: Size, posterior mean of θ_c and posterior mean of μ_c for each cluster.

	size	θ_c	SYSVAL	DIAVAL	MAPVAL	CHOL	HDL	TRIG	LDL	HSCRP	WSTVAL	WHVAL	BMIVAL
1	1432	5.82	130.35	75.47	93.76	6.00	1.65	1.24	3.78	1.57	90.26	0.87	26.06
2	436	5.99	139.77	76.82	97.80	6.18	1.60	1.55	3.87	4.85	95.95	0.88	29.04
3	760	6.22	135.28	77.25	96.59	6.09	1.43	1.97	3.76	3.12	98.24	0.91	28.50
4	231	6.40	138.20	77.76	97.93	5.88	1.47	1.73	3.63	14.23	101.57	0.90	30.52

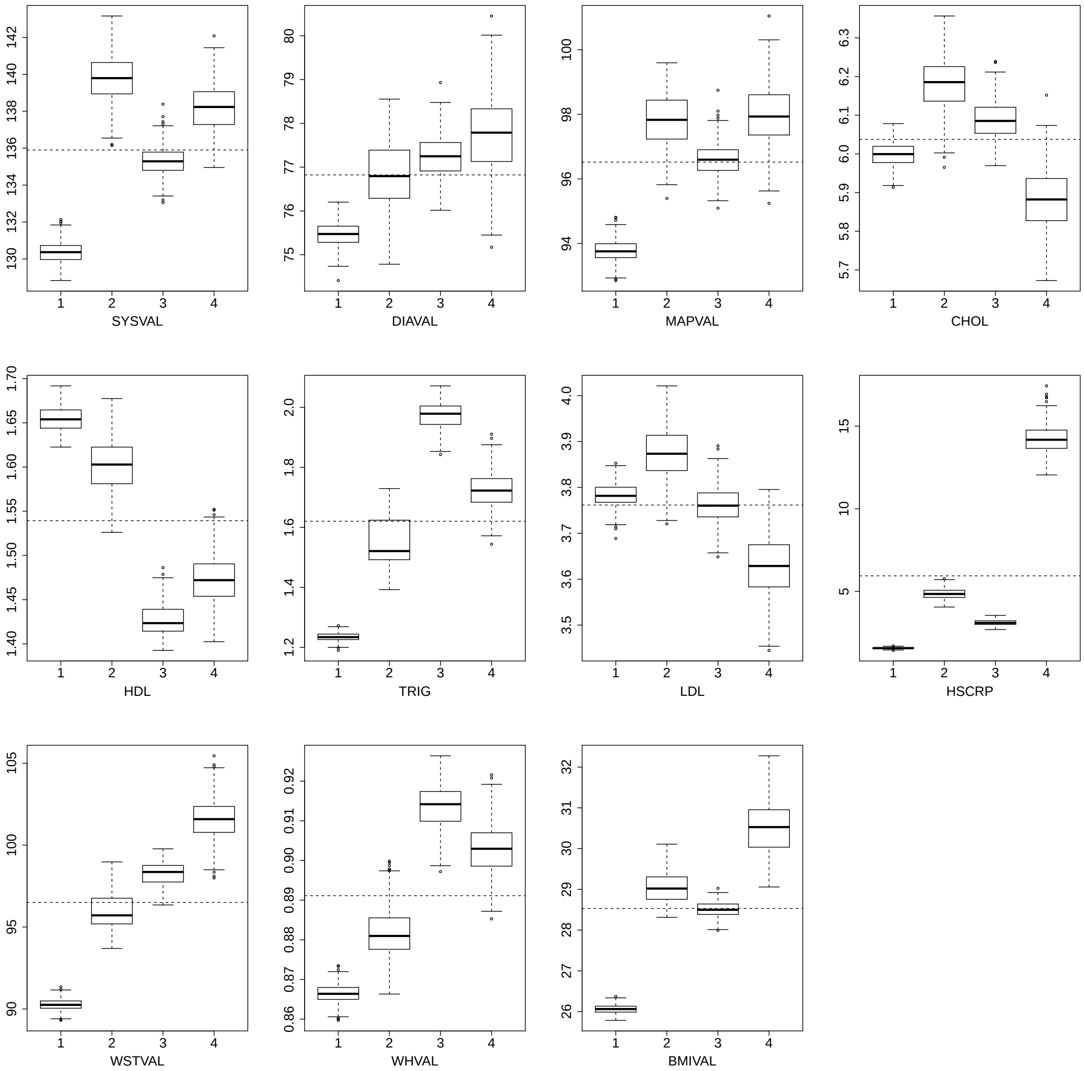


FIGURE 4.7: Boxplots of the posterior distribution of all covariates for the four clusters identified by quantile profile regression. The horizontal dashed line is the overall posterior mean for each covariate.

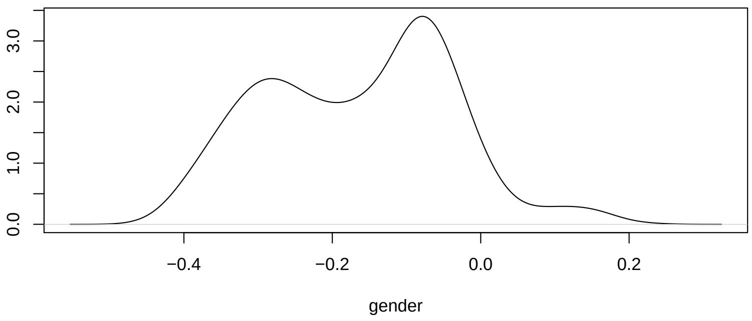


FIGURE 4.8: Posterior distribution of β (gender).

In assessing the cardiometabolic profile of each cluster we can look to Figure 4.7. Cluster 1 (lowest blood glucose levels) generally showed low cardiometabolic risk. Cluster 2 (who had moderately high levels of blood glucose) showed a relatively high risk profile for other cardiometabolic risk factors, with high levels of cholesterol, high/low-density lipoprotein level and blood pressure with borderline/high anthropometric indicators. Cluster 3 (who showed a high blood glucose profile) had borderline blood pressure and anthropometric indicators but a high risk cholesterol and triglyceride profile. Cluster 4 (the highest blood glucose levels) showed very high levels of all cardiometabolic indicators. These results indicate that quantile profile regression modelling is able to discriminate between different levels of blood glucose based on the presence of cardiometabolic risk factors in a way that is theoretically sound (i.e., low blood glucose levels are associated with low cardiometabolic risk and high blood glucose levels are associated with higher cardiometabolic risk) while also being sensitive enough the reveal different groups that could be of clinical interest (e.g., cluster 4 indicated levels of extremely high inflammation through raised CRP which could be of interest to clinicians).

These data indicate that quantile profile regression could be a useful tool for identifying clusters of people based on shared cardiometabolic risk factors. As this analysis was cross sectional we cannot infer whether these clusters would predict incidence of type 2 diabetes, however this modelling tool shows promise for application in the context of illness risk.

We are interested in high values of Y because they correspond to hyperglycemia. First we extract the observations with $Y >= 6$ as validation data, part of which are listed in Table 4.7. We can show the higher accuracy of quantile profile regression over normal profile regression when predicting values of Y in validation data. The RMSE and the MAE obtained comparing these predictions to the observed values are given in Table 4.8. As for simulated data, quantile profile regression proves to be more accurate when predicting values around the quantiles of interest.

TABLE 4.7: A sample of the observations that $Y >= 6$.

SEX	SYSVAL	DIAVAL	MAPVAL	CHOL	HDL	TRIG	LDL	HSCRP	WSTVAL	WHVAL	BMIVAL	FGLU
1	116.50	60.50	79.00	3.30	1.10	0.60	1.90	4.10	88.20	0.78	27.12	6.00
0	112.50	71.00	84.50	4.40	1.10	2.50	2.20	0.60	102.85	0.91	28.18	6.10
1	127.50	62.50	84.00	6.70	1.20	4.00	3.70	7.00	88.40	0.83	27.74	6.10
0	142.00	78.00	99.50	4.50	1.40	1.80	2.50	6.80	122.20	1.03	30.84	6.10
0	142.50	87.00	105.50	3.20	1.20	1.00	1.50	2.10	117.50	0.97	34.61	6.10
1	150.50	97.50	115.00	6.60	1.40	2.70	4.00	4.80	115.70	0.97	36.06	6.20
0	172.50	90.00	117.50	5.80	1.50	1.00	3.80	38.30	97.60	0.93	27.07	6.60
0	106.00	77.00	86.50	4.80	1.00	1.10	3.30	11.10	83.60	0.84	20.07	6.40
0	132.50	78.50	96.50	6.20	1.50	2.30	3.70	2.10	107.75	1.02	30.42	6.20

TABLE 4.8: The RMSE and the MAE for the prediction of the Y values such that $Y >= 6$ applying quantile profile regression and normal profile regression.

	RMSE	MAE
quantile 0.95	1.20	0.93
normal	1.89	1.73

4.6 Chapter Summary

We have proposed a new method for collinear data which is more accurate than existing methods when the modelling interest is in the tails of the distribution. The method is an extension of profile regression, a Bayesian clustering model, and it was applied to simulated and real data and it provided a significant increase in accuracy with considerable reduction in the residuals, especially under extreme quantiles, compared to an estimation with the normal mixture model.

The method proposed is not a standard regression approach, so it does not allow to estimate the effect of each predictor on the outcome, but it allows to explain the complex relationships between predictors and the response variables. This is demonstrated in Sections 4.4 and 4.5, and also explored more extensively by Molitor et al. (2010);

Hastie et al. (2013); Molitor et al. (2014); Mattei et al. (2016); Liverani et al. (2016); Coker et al. (2016). Profile regression is able to disentangle the complex relationships between predictors and response variables and can be used to evaluate how changes in the predictors might affect the response variable.

A limitation of the model proposed in its present form is that the asymmetric Laplace distribution is included for the response variable but not for the predictors, so it does not account for interest in the tails of the distribution of the covariates. This is the topic of future work.

Chapter 5

Conclusions and Future Research

This thesis has proposed several new developments for quantile regression to address common challenges such as discrete responses, quantile non-crossing, and clustering problems. Clear advantages over existing methods include a coherent Bayesian approach, a normal scale-mixture representation of ALD that theoretically guarantee quantile non-crossing and a Bayesian clustering model that discover complex relationship among high correlated covariates. The main contributions and future research topics are listed below.

5.1 Main Contributions

Bayesian regression beyond the mean for discrete responses are proposed in Chapter 2. This method is proposed via the development of discrete probability mass functions for likelihood functions. Bayesian quantile regression for discrete responses is first developed. This method is then extended to Bayesian expectile regression for discrete responses. This method provides a direct Bayesian approach to these regression models, therefore interpretations of regression results becomes easy and intuitive. In particular, this approach has proven to be coherent irrespective of the true distribution of the response and also proper with regarding to improper priors for unknown model parameters.

In Chapter 3, a new kernel-weighted likelihood smoothing method is proposed to address important challenges that may arise in quantile regression, such as lack of accuracy at extreme quantiles and quantile crossing problems. An automatic data-driven method for selecting these bandwidths is proposed, which not only enjoys the propagation condition but also guarantees quantile non-crossing. Theoretical results claim that the proposed adaptive procedure performs well, which would minimize the risk of localized estimation for the problem at hand. Several advantages of the proposed approach over the existing method are discussed.

In Chapter 4, quantile profile regression is provided to distinguish and interpret the complex relationship between several predictors and the tail of the distribution of a response variable in the presence of high correlation between the predictors. It allows the effects of changes in predictors on the response variable to be evaluated. An MCMC-based computation technique with an additional Gibbs sampler is developed. Several advantages of the proposed approach over existing normal profile regression are discussed.

5.2 Recommendations for Future Research

The work considered in Chapter 2 opens the door to new research directions for discrete responses in quantile regression by adopting discrete probability mass functions for likelihood functions. There are many possible extensions such as extending the proposed method to Bayesian semi-parametric quantile regression and expectile regression for discrete responses. Taking quantile regression as an example, one could consider a non-parametric mixture with a Dirichlet process prior, denoted by $\mathrm{DP}(\alpha, G_0)$, with parameter α and base distribution G_0, for the mixing distribution. Then one could introduce a latent mixing scale parameter σ associated with response observation Y, with the p.m.f of the DALD expressed:

$$\begin{aligned}&\phi(y;\mu,\sigma,\tau)\\ &= \begin{cases} (\tau-1)\left[\exp\{-\frac{\tau}{\sigma}\}-1\right]\exp\left\{-\rho_\tau(\frac{y-\mu}{\sigma})\right\}, & y \geq \mu, \\ \tau\left[\exp\{\frac{1-\tau}{\sigma}\}-1\right]\exp\left\{-\rho_\tau(\frac{y-\mu}{\sigma})\right\}, & y < \mu. \end{cases}\end{aligned}$$

Or,

$$\phi(y;\mu,p,q) = \begin{cases} \log q\left[p^{y-\mu}(p-1)\right], & y \geq \mu, \\ \log p\left[q^{-(y-\mu)}(q-1)\right], & y < \mu, \end{cases}$$

where the parameters (p, q), $(0 < p, q < 1)$, are related to τ and σ via the relationships $p = \exp\left\{-\frac{\tau}{\sigma}\right\}$ and $q = \exp\left\{\frac{\tau-1}{\sigma}\right\}$ or $\tau = \log p/\log(pq)$ and $\sigma = -1/\log(pq)$.

Then, for $i = 1, \cdots, n$, the model can be expressed in the hierarchical form:

$$\begin{aligned} &Y_i|\boldsymbol{\beta}, \sigma \sim \text{DALD}(Y; \boldsymbol{X}_i^T\boldsymbol{\beta}, \sigma_i, \tau) \\ &\sigma_i|G \sim G \\ &G|\alpha, d \sim \text{DP}(\alpha, G_0) \\ &\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}) \end{aligned} \tag{5.1}$$

with a Gamma prior on α and an inverse Gamma distribution for G_0 with mean $d/c - 1(c > 1)$. Here, let $c = 2$, which yields an infinite variance for G_0 and work with a Gamma prior for d. In this case, the extension of methods for prior specification, posterior inference can also be implemented based on MCMC techniques (Kottas and Krnjajić 2009; Taddy and Kottas 2010).

The kernel-weighted likelihood smoothing quantile regression method reported in Chapter 3 can be extended to the d-dimensional case $X \in \mathbb{R}^d$, with $d > 1$, under the non-parametric additive modelling framework (Yu and Lu, 2004). That is, let Y be a real-valued dependent variable and $X = \left(X^{(1)}, \cdots, X^{(d)}\right) \in \mathbb{R}^d$ as a vector of explanatory variables. Let $f(x)$ be a d-dimensional τth quantile regression function of Y given $X = x$. Suppose that the τth quantile function $f(\boldsymbol{x})$ is modelled as an additive function of $\left(x^{(1)}, \cdots, x^{(d)}\right)$,

$$f(x) = \sum_{l=1}^{d} f^{(l)}(x^{(l)}), \tag{5.2}$$

where each $f^{(l)}(x^{(l)})$ can be fitted by the proposed approach in Section 3 and the whole $f(x)$ can be further derived via backfitting algorithm used in Yu and Lu (2004). Without loss of generality, consider a local linear regression with $p = 2$, for $l = 1, \cdots, d$:

$$(\hat{a}^{(l)}, \hat{b}^{(l)}) = \underset{a,b}{\text{argmin}} \sum_{i=1}^{n} \rho_\tau \left(Y_i - a - b(X_i^{(l)} - x^{(l)})\right) \left(\frac{X_i^{(l)} - x^{(l)}}{h^{(l)}}\right).$$

where $K(\cdot)$ is a kernel function and $h^{(l)}(l = 1, \cdots, d)$ is the bandwidth for estimating $f^{(l)}(x^{(l)})$ in the setting above.

One can also extend the idea of quantile profile regression in Chapter 4 to account for interest in the tails of the distribution of the covariates, also take account of the extension to discrete responses through the combination with the developments in Chapter 2 .

Appendix A

Appendix A

Appendix A aims to show the proof of Theorem 2.1, that is, the posterior distribution is proper with regards to improper priors for the unknown parameters.

Proof of Theorem 2.1 A parametrization of the DALD in Eq.(2.5) leads to the following alternative form,

$$\phi(y;\mu,p,q) = \begin{cases} p^{y-\mu}(1-p)\log q, & y \geq \mu, y \in \mathbb{Z} \\ q^{y-\mu}(1-q)\log p, & y < \mu, y \in \mathbb{Z} \end{cases} \tag{A.1}$$

where the parameters p and q $(0 < p, q < 1)$ are related to τ via the relationships $p = \exp\{-\tau\}$ and $q = \exp\{1-\tau\}$.

Lemma A.1. *The p.m.f. $\phi(t)$ defined in Eq.(A.1) is bounded by $p^{|t|}(1-q)\log p$ and $q^{|t|}(1-p)\log q$.*

Proof of Lemma A.1. Expand $\phi(t)$ as a mixture of g, consider $0 < q \leq p < 1$,

$$\begin{aligned}\phi(t) &= p^{|t|}(1-p)\log q\mathbb{I}(t \geq 0) + q^{|t|}(1-q)\log p\mathbb{I}(t < 0) \\ &\leq (1-q)\log p\left(p^{|t|}\mathbb{I}(t \geq 0) + q^{|t|}\mathbb{I}(t < 0)\right) \\ &\leq p^{|t|}(1-q)\log p.\end{aligned}$$

Also,

$$\begin{aligned}\phi(t) &= p^{|t|}(1-p)\log q\mathbb{I}(t \geq 0) + q^{|t|}(1-q)\log p\mathbb{I}(t<0)\\ &\geq (1-p)\log q\left(p^{|t|}\mathbb{I}(t\geq 0) + q^{|t|}\mathbb{I}(t<0)\right)\\ &\geq q^{|t|}(1-p)\log q.\end{aligned}$$

Now, it is known that $g(t;a) = a^{|t|}(a-1)\log a$ with $0 < a < 1$ is a increasing function of t. Therefore, $\phi(t)$ has upper bound $h(p,q)q^{|t|}$ and lower bound $h(q,p)p^{|t|}$. The same procedure may be easily adapted to $q \geq p$. □

Lemma A.2. *For any constant $a(0 < a < 1)$ and sample size $n > m$,*

$$\int \prod_{k=0}^{m} |\beta_k|^{r_k} \prod_{i=1}^{n} \exp\{(\log a)|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}|\}d\boldsymbol{\beta} < \infty.$$

Proof of Lemma A.2. Without loss of generality and consider $m = 1$ for simplicity, then $\boldsymbol{X}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 X_{1i}$,

$$\begin{aligned}&\int_{\mathbb{R}^2} |\beta_0|^{r_0}|\beta_1|^{r_1}\exp\left\{(\log a)\sum_{i=1}^{n}|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}|\right\}d\beta_0 d\beta_1\\ &\leq \int_{\mathbb{R}^2} |\beta_0|^{r_0}\exp\{(\log a)|\beta_0 + \beta_1 X_{11} - Y_1|\}|\beta_1|^{r_1}\\ &\times \exp\{(\log a)|\beta_0 + \beta_1 X_{12} - Y_2|\}d\beta_0 d\beta_1.\end{aligned}$$

Since the double-integration $\int_{\mathbb{R}^2}|U|^{r_0}\exp(-|U+V+c_1|)|V_1|^{r_1}\exp(-|U+V+c_2)dUdV$ is finite for any constants c_1, c_2, $r_0 \geq$ and $r_1 \geq 0$, Lemma 2 is proved. □

Theorem 2.1 below establishes that in the absence of any realistic prior information we could legitimately use an improper uniform prior distribution for all the components of $\boldsymbol{\beta}$.

Theorem 2.1 *Assume the posterior is given by Eq.(2.8) and $\pi(\boldsymbol{\beta}) \propto 1$, then all posterior moments of $\boldsymbol{\beta}$ in Eq.(2.9) exist.*

Proof of Theorem 2.1. We need to prove that

$$\int_{\mathbb{R}^{m+1}} \prod_{k=0}^{m} |\beta_k|^{r_k}\exp\left\{(\log a)\sum_{i=1}^{n}|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}|\right\}d\boldsymbol{\beta},$$

is finite. According to Lemma A.1 and Lemma A.2, it suffices to be proved. □

Property in Section 2.4 Under Bayesian inference of expectile μ for the discrete random variable Y, it can be proved that the posterior distribution is proper with regards to improper priors for the unknown parameters.

Similar to Lemma 1, the p.m.f $\phi^{(E)}(t)$ defined in Eq.(2.13) is bounded. Consider $\theta \leq 0.5$,

$$\begin{aligned}
\phi^{(E)}(t) &= k\sqrt{\frac{\pi}{\theta}}\left[\Phi\left(\sqrt{2\theta}(t+1)\right) - \Phi\left(\sqrt{2\theta}t\right)\right]\mathbb{I}_{t>0} \\
&\quad + k\sqrt{\frac{\pi}{1-\theta}}\left[\Phi\left(\sqrt{2(1-\theta)}(t+1)\right) - \Phi\left(\sqrt{2(1-\theta)}t\right)\right]\mathbb{I}_{t\leq 0} \\
&\leq k\sqrt{\frac{\pi}{\theta}}\left(\left[\Phi\left(\sqrt{2\theta}(t+1)\right) - \Phi\left(\sqrt{2\theta}t\right)\right]\right)\mathbb{I}_{t>0} \\
&\quad + k\sqrt{\frac{\pi}{\theta}}\left(\left[\Phi\left(\sqrt{2(1-\theta)}(t+1)\right) - \Phi\left(\sqrt{2(1-\theta)}t\right)\right]\right)\mathbb{I}_{t\leq 0} \\
&\leq k\sqrt{\frac{\pi}{\theta}}\left[\Phi\left(\sqrt{2(1-\theta)}(t+1)\right) - \Phi\left(\sqrt{2(1-\theta)}t\right)\right] \\
&= k\sqrt{\frac{2\pi(1-\theta)}{\theta}}\Phi'(t).
\end{aligned}$$

Also,

$$\begin{aligned}
\phi^{(E)}(t) &= k\sqrt{\frac{\pi}{\theta}}\left[\Phi\left(\sqrt{2\theta}(t+1)\right) - \Phi\left(\sqrt{2\theta}t\right)\right]\mathbb{I}_{t>0} \\
&\quad + k\sqrt{\frac{\pi}{1-\theta}}\left[\Phi\left(\sqrt{2(1-\theta)}(t+1)\right) - \Phi\left(\sqrt{2(1-\theta)}t\right)\right]\mathbb{I}_{t\leq 0} \\
&\geq k\sqrt{\frac{\pi}{1-\theta}}\left(\left[\Phi\left(\sqrt{2\theta}(t+1)\right) - \Phi\left(\sqrt{2\theta}t\right)\right]\right)\mathbb{I}_{t>0} \\
&\quad + k\sqrt{\frac{\pi}{1-\theta}}\left(\left[\Phi\left(\sqrt{2(1-\theta)}(t+1)\right) - \Phi\left(\sqrt{2(1-\theta)}t\right)\right]\right)\mathbb{I}_{t\leq 0} \\
&\geq k\sqrt{\frac{\pi}{1-\theta}}\left[\Phi\left(\sqrt{2\theta}(t+1)\right) - \Phi\left(\sqrt{2\theta}t\right)\right] \\
&= k\sqrt{\frac{2\pi\theta}{1-\theta}}\Phi'(t),
\end{aligned}$$

where $k = \frac{2}{\sqrt{\pi}}\frac{\sqrt{\theta(1-\theta)}}{\sqrt{\theta}+\sqrt{1-\theta}}$, $\Phi(\cdot)$ denotes the c.d.f. of the standard normal distribution. Now, it is known that $\sqrt{\frac{\theta}{1-\theta}} \leq \sqrt{\frac{1-\theta}{\theta}}$ for $\theta \leq 0.5$. Therefore, $\phi^{(E)}(t)$ has upper bound $k\sqrt{\frac{2\pi(1-\theta)}{\theta}}\Phi'(t)$ and lower bound $k\sqrt{\frac{2\pi\theta}{1-\theta}}\Phi'(t)$. The same procedure may be easily adapted to $\theta \geq 0.5$.

Then, for $n > m$, we also have

$$\int \prod_{k=0}^{m} |\beta_k|^{r_k} \prod_{i=1}^{n} \exp\{-(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2\} d\boldsymbol{\beta} < \infty.$$

Similarly, consider $m = 1$ for simplicity, then $\boldsymbol{X}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 X_{1i}$,

$$\begin{aligned}
&\int_{\mathbb{R}^2} |\beta_0|^{r_0} \beta_1|^{r_1} \exp\{-(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})^2\} d\beta_0 d\beta_1 \\
&\leq \int_{\mathbb{R}^2} |\beta_0|^{r_0} \exp\{-(\beta_0 + \beta_1 X_{11} - Y_1)^2\} |\beta_1|^{r_1} \\
&\times \exp\{-(\beta_0 + \beta_1 X_{12} - Y_2)^2\} d\beta_0 d\beta_1.
\end{aligned}$$

As is known to all that the double-integration is finite for any constants.

Therefore, assume the likelihood is given by Eq.(2.8) and $\pi(\boldsymbol{\beta}) \propto 1$, then it can be proved that all posterior moments of $\boldsymbol{\beta}$ in Eq.(2.9) exist.

Appendix B

Appendix B

Appendix B aims to elaborate the proof Theorem 3.1 in Chapter 3.

Recall: $\boldsymbol{w}_k = diag\left(\frac{w_1^{(k)}}{\delta^2 z_1}, ..., \frac{w_n^{(k)}}{\delta^2 z_n}\right)$.

Assumption B.1. *Consider a finite sequence of scales* $w_k = diag\left(w_1^{(k)}, \cdots, w_n^{(k)}\right)$*, the* $p \times n$ *matrix* $\boldsymbol{\psi}^T w_1$*is of full row rank.*

Assumption B.2. *For any fixed* x *and the method of localization with* $w_i^{(k)}(x) \geq 0$*, the following relation holds:*

$$w_1(x) \leq w_2(x) \leq \cdots \leq w_K(x).$$

Assumption B.3. *Assume that the true regression model*

$$Y_i = f_0(X_i) + \mu_0 z_{0,i} + \delta_0^2 \sqrt{z_{0,i}} e_i,$$

considering the regression model (3.8), where $\boldsymbol{Z}_0 = diag\left(\delta_0^2 z_{0,1}, \cdots, \delta_0^2 z_{0,n}\right)$ stands for the unknown true covariance matrix, with $z_{0,i}$ is the true value of Eq.(3.8), there exists $\eta \in [0, 1)$ such that

$$1 - \eta \leq \frac{\delta_0^2 z_{0,i}}{\delta^2 z_i} \leq 1 + \eta \quad \text{for all } i = 1, \cdots, n.$$

Assuming Assumption B.3, the true covariance matrix $\boldsymbol{Z}_0 \preceq \boldsymbol{Z}(1+\eta)$, and the conditional variance of the estimate $\tilde{\boldsymbol{\theta}}_k(x)$ is bounded with $\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}$: as follows :

$$\begin{aligned}
\operatorname{Var}\left(\tilde{\boldsymbol{\theta}}_k(x)\right) &= \left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{Z}_0\boldsymbol{w}_k\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&\preceq (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{Z}\boldsymbol{w}_k\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{Z}^{-1/2}w_k^2\boldsymbol{Z}^{-1/2}\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&\preceq (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\boldsymbol{\psi}\boldsymbol{Z}^{-1/2}w_k\boldsymbol{Z}^{-1/2}\boldsymbol{\psi}^T\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1}\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\boldsymbol{\psi}\boldsymbol{w}_k\boldsymbol{\psi}^T\right)^{-1} \\
&= (1+\eta)\left(\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\frac{w_i^{(k)}}{\delta^2 z_i}\right)^{-1}. \qquad \text{(B.1)}
\end{aligned}$$

According to the basic property of quadratic equation, consider a simple example $\left(\frac{1}{z_1}+\frac{1}{z_2}\right)^{-1}$ and there always holds $\left(\frac{1}{z_1}+\frac{1}{z_2}\right)^{-1} = \frac{z_1z_2}{z_1+z_2} \leq z_1+z_2$, with $z_1, z_2 > 0$. The same procedure may be easily adapted to Eq.(B.1) as follows:

$$\begin{aligned}
\operatorname{Var}\left(\tilde{\boldsymbol{\theta}}_k(x)\right) &\preceq (1+\eta)\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)}\delta^2 z_i \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} z_i. \qquad \text{(B.2)}
\end{aligned}$$

Therefore, the unconditional variance of the estimate $\tilde{\boldsymbol{\theta}}_k(x)$ as follows is bounded with $\boldsymbol{\psi}w_k\boldsymbol{\psi}^T$

$$\begin{aligned}
\mathbf{V}_k(x) &\equiv E\left[\operatorname{Var}\tilde{\boldsymbol{\theta}}_k(x)\right] \\
&= E\left[(1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} z_i\right] \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} E\left[z_i\right] \\
&= (1+\eta)\delta^2\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T w_i^{(k)} \\
&= (1+\eta)\delta^2\boldsymbol{\psi}w_k\boldsymbol{\psi}^T. \qquad \text{(B.3)}
\end{aligned}$$

Assumption B.4. *Let for some constants b_0 and b such that $1 < b_0 \leq b$ for any $2 \leq k \leq K$, the matrices $\boldsymbol{B}_k = \Psi\, \boldsymbol{W}_k \Psi^T$ satisfy*

$$b_0 I_p \preceq \boldsymbol{B}_{k-1}^{-1/2} \boldsymbol{B}_k \boldsymbol{B}_{k-1}^{-1/2} \preceq b I_p.$$

Proof. of Theorem 3.1.

$$\begin{aligned}
&\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)^T \left(\boldsymbol{\psi} w_k(x) \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \\
= &\sum_{m=1}^{k-1} \mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \tilde{\boldsymbol{\theta}}_m(x)\right)^T \left(\boldsymbol{\psi} w_k \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_k(x) - \tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_m(x)\right\} \qquad \text{(B.4)}
\end{aligned}$$

The event $\left\{\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_m(x)\right\}$ happens if for some $l = 1, \cdots, m$, $T_{l,m+1} > \zeta_l$, Hence,

$$\left\{\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_m(x)\right\} \subseteq \bigcup_{l=1}^{m} \{T_{l,m+1} > \zeta_l\},$$

where $T_{l,m+1} = \left(\tilde{\boldsymbol{\theta}}_l(x) - \tilde{\boldsymbol{\theta}}_{m+1}(x)\right)^T \left(\boldsymbol{\psi} w_l(x) \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_l(x) - \tilde{\boldsymbol{\theta}}_{m+1}(x)\right)$.

Further, combined with the Cauchy-Schwarz inequality, for any positive a:

$$\begin{aligned}
&\mathbb{E}\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \tilde{\boldsymbol{\theta}}_m(x)\right)^T \left(\boldsymbol{\psi} w_k \boldsymbol{\psi}^T\right) \left(\tilde{\boldsymbol{\theta}}_k(x) - \tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_m(x)\right\} \\
= \;&\mathbb{E}\left|2L_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_k(x), \tilde{\boldsymbol{\theta}}_m(x)\right)\right|^r I\left\{\hat{\boldsymbol{\theta}}_k(x) = \tilde{\boldsymbol{\theta}}_m(x)\right\} \\
\leq \;&\sum_{l=1}^{m} e^{-\frac{a}{4}\zeta_l} \left\{\mathbb{E}\left[\left|2L_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_k(x), \tilde{\boldsymbol{\theta}}_m(x)\right)\right|^{2r}\right]\right\}^{\frac{1}{2}} \\
&\left\{\mathbb{E}\left[\exp\left\{aL_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_l(x), \tilde{\boldsymbol{\theta}}_{m+1}(x)\right)\right\}\right]\right\}^{\frac{1}{2}}. \qquad \text{(B.5)}
\end{aligned}$$

Among which,

$$\begin{aligned}
&E\left[\left|2L_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_k(x), \tilde{\boldsymbol{\theta}}_m(x)\right)\right|^{2r}\right] \\
= \quad &2r \int_0^\infty P\left\{2L_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_k(x), \tilde{\boldsymbol{\theta}}_m(x)\right) \geq \zeta\right\} \zeta^{2r-1} d\zeta \\
\leq \quad &2r \int_0^\infty P\left\{\gamma \geq \zeta \left[2\left(1+\eta\right)\left(1 + b^{(k-m)}\right)\right]^{-1}\right\} \zeta^{2r-1} d\zeta \\
= \quad &2^{2r}\left(1+\eta\right)^{2r}\left(1 + b^{(k-m)}\right)^{2r} E\left|\chi_p^2\right|^r \\
=\eta = 0 \quad &2^{2r} C(p, 2r)\left(1 + b^{(k-m)}\right)^{2r}, \qquad \text{(B.6)}
\end{aligned}$$

and

$$
\begin{aligned}
& E\left[\exp\left\{aL_{NSM}\left(W^{(k)}, \tilde{\boldsymbol{\theta}}_l(x), \tilde{\boldsymbol{\theta}}_{m+1}(x)\right)\right\}\right] \\
= \quad & \prod_{j=1}^{p}\left[1 - a\lambda_j\left(V_{l,m+1}^{-1/2}\left(\boldsymbol{\psi} w_m \boldsymbol{\psi}^T\right) V_{l,m+1}^{-1/2}\right)\right]^{-1/2} \\
\leq \quad & \left[1 - a\lambda_{max}\left(V_{l,m+1}^{-1/2}\left(\boldsymbol{\psi} w_m \boldsymbol{\psi}^T\right) V_{l,m+1}^{-1/2}\right)\right]^{-p/2} \\
\leq \quad & \left[1 - 2a\left(1+\eta\right)\left(1 + b^{-(m+1-l)}\right)\right]^{-p/2} \\
= \eta = 0 \quad & \left[1 - 2a\left(1 + b^{-(m+1-l)}\right)\right]^{-p/2}. \qquad \text{(B.7)}
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
& E\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)^T \left(\boldsymbol{\psi} w_k(x) \boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \\
\leq \quad & 2^r\sqrt{C(p,2r)}(1-4a)^{-p/4}\sum_{m=1}^{k-1}\sum_{l=1}^{m} e^{-\frac{\mu}{4}\zeta_l}\left(1 + b^{(k-m)}\right)^r \\
\leq \quad & 2^{2r}\sqrt{C(p,2r)}(1-4a)^{-p/4}(1-b^{-r})\sum_{l=1}^{k-1} e^{-\frac{\mu}{4}\zeta_l} b^{r(k-l)}. \qquad \text{(B.8)}
\end{aligned}
$$

For any $l < k < K$, with an arbitrary constant $a \in (0, 1/4)$ the choice of the threshold of the form

$$
\zeta_l = \frac{4}{a}\left\{r(K-l)\log b + \log\frac{K}{\alpha} - \frac{p}{4}\log(1-4\mu) - \log(1-b^{-r}) + \bar{C}(p,r)\right\},
$$

where $\bar{C}(p,r) = \log\left\{\frac{2^{2r}[\Gamma(2r+p/2)\Gamma(p/2)]^{1/2}}{\Gamma(r+p/2)}\right\}$ provides the required PC bounds.

$$
E\left|\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)^T \left(\boldsymbol{\psi} w_k(x) \boldsymbol{\psi}^T\right)\left(\tilde{\boldsymbol{\theta}}_k(x) - \hat{\boldsymbol{\theta}}_k(x)\right)\right|^r \leq \alpha C(p,r), \; for\ all\ k = 2, \cdots, K.
$$

□

Appendix C

Appendix C

Appendix C shows the details of Notations for DPMM sampler in Chapter 4.

Given the allocation variables $\mathbf{Z}$, define

$$Z^* = \max_{1 \leq i \leq n} Z_i.$$

Similarly, given the auxiliary variable $\mathbf{U} = (U_1, U_2, \cdots, U_n)$ and the vector $\mathbf{V}$, define

$$U^* = \min_{1 \leq i \leq n} U_i.$$

and

$$\begin{aligned} C^* &= \min \left\{ c \in \mathbb{Z}^+ : \sum_{l=1}^{c} \psi_l > 1 - U^* \right\} \\ &= \min \left\{ c \in \mathbb{Z}^+ : \sum_{l=1}^{c} \left[V_l \prod_{r<l} (1 - V_r) \right] > 1 - U^* \right\} \end{aligned}$$

The purpose of the variable C^* is to provide an upper limit on which mixture components need updating at each sweep. Specifically, although there are infinitely many component parameters in the model, since $P(Z_i = c | U_i > \psi_c) = 0$, we need only concentrate our updating efforts on those components c for which $\psi_c > U_i$ for some $i = 1, 2, \cdots, n$.

With these definitions in place we make use of the following sets and vectors (which again will change at each sweep)

$$A = \{c \in \mathbb{Z}^+ : c \leq Z^*\}, \mathbf{V}^A = (V_1, V_2, \cdots, V_{Z^*}), \boldsymbol{\Theta}^A = (\Theta_1, \Theta_2, \cdots, \Theta_{Z^*})$$
$$P = \{c \in \mathbb{Z}^+ : Z^* < c \leq C^*\}, \mathbf{V}^P = (V_{Z^*+1}, V_{Z^*+2}, \cdots, V_{C^*}), \boldsymbol{\Theta}^P = (\Theta_{Z^*+1}, \Theta_{Z^*+2}, \cdots, \Theta_{C^*})$$
$$I = \{c \in \mathbb{Z}^+ : Z^* > C^*\}, \mathbf{V}^I = (V_{C^*+1}, V_{C^*+2}, \cdots), \boldsymbol{\Theta}^I = (\Theta_{C^*+1}, \Theta_{C^*+2}, \cdots)$$

Bibliography

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). *Automatic subspace clustering of high dimensional data for data mining applications*, volume 27. ACM.

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." *In: B. N. PETROV and F. CSAKI, eds. Second International Symposium on Information Theory. Budapest: Akademiai Kiado*, 267–281.

Alhamzawi, R. (2012). "Package 'Brq'." *Statistical Modelling*, 12: 279–297.

Alhamzawi, R. and Yu, K. (2013). "Conjugate priors and variable selection for Bayesian quantile regression." *Computational Statistics & Data Analysis*, 64: 209–219.

Alhamzawi, R., Yu, K., Vinciotti, V., and Tucker, A. (2011). "Prior elicitation for mixed quantile regression with an allometric model." *Environmetrics*, 22(7): 911–920.

Anderson, M. J. (2008). "Animal-sediment relationships re-visited: Characterising species' distributions along an environmental gradient using canonical analysis and quantile regression splines." *Journal of Experimental Marine Biology and Ecology*, 366: 16–27.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., and Sander, J. (1999). "OPTICS: ordering points to identify the clustering structure." In *ACM Sigmod record*, volume 28, 49–60. ACM.

Atella, V., Pace, N., and Vuri, D. (2008). "Are employers discriminating with respect to weight?: European Evidence using Quantile Regression." *Economics & Human Biology*, 6: 305–329.

Bassett, G. W. and Koenker, R. W. (1986). "Strong consistency of regression quantiles and related empirical processes." *Econometric Theory*, 2(02): 191–201.

Benoit, D., AlHamzawi, R., Yu, K., and Poel, D. (2014). "bayesQR: Bayesian quantile regression." https://cran.r-project.org/web/packages/bayesQR/index.html. Accessed : 2014-04-18.

Benoit, D. and Van den Poel, D. (2012). "Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution." *Journal of Applied Econometrics*, 27: 1174–1188.

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

Bissiri, P., Holmes, C., and Walker, S. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society: Series B*.
URL http://dx.doi.org/10.1111/rssb.12158

Bondell, H. D., Reich, B. J., and Wang, H. (2010). "Noncrossing quantile regression curve estimation." *Biometrika*, 825–838.

Borghans, I., Hekkert, K. D., den Ouden, L., Cihangir, S., Vesseur, J., Kool, R. B., and Westert, G. P. (2014). "Unexpectedly long hospital stays as an indicator of risk of unsafe care: an exploratory study." *BMJ open*, 4(6): e004773.

Bottai, M., Cai, B., and McKeown, R. (2010). "Logistic quantile regression for bounded outcomes." *Statistics in Medicine*, 169: 309–317.

Briollais, L. and Durrieu, G. (2014). "Application of quantile regression to recent genetic and -omic studies." *Human genetics*, 133: 951–966.

Buchinsky, M. (1995). "Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study." *Journal of Econometrics*, 68(2): 303–338.

Cade, B. S. and Noon, B. R. (2003). "A gentle introduction to quantile regression for ecologists." *Frontiers in Ecology and the Environment*, 1(8): 412–420.

Cai, Z. and Xu, X. (2008). "Nonparametric quantile estimations for dynamic smooth coefficient models." *Journal of the American Statistical Association*, 103: 1595–1608.

Canale, A. and Dunson, D. B. (2011). "Bayesian kernel mixtures for counts." *Journal of the American Statistical Association*, 106(496): 1528–1539.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." *Computers & Geosciences*, 37: 1277–1284.

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

Chamberlain, G. and Imbens, G. W. (2003). "Nonparametric applications of Bayesian inference." *Journal of Business & Economic Statistics*, 21(1): 12–18.

Chaudhuri, P. (1991). "Nonparametric estimates of regression quantiles and their local Bahadur representation." *The Annals of statistics*, 19: 760–777.

Chen, R.-B., Chu, C.-H., Lai, T.-Y., and Wu, Y. N. (2011). "Stochastic matching pursuit for Bayesian variable selection." *Statistics and Computing*, 21(2): 247–259.

Chen, X., Koenker, R., and Xiao, Z. (2009). "Copula-based nonlinear quantile autoregression." *The Econometrics Journal*, 12(s1).

Chernozhukov, V., Gagliardini, P., and Scaillet, O. (2008). *Nonparametric instrumental variable estimation of quantile structural effects*. HEC.

Coad, A. and Rao, R. (2008). "Innovation and firm growth in high-tech sectors: A quantile regression approach." *Research Policy*, 37: 633–648.

Coker, E., Liverani, S., Ghosh, J. K., Jerrett, M., Beckerman, B., Li, A., Ritz, B., and Molitor, J. (2016). "Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County." *Environment international*, 91: 1–13.

Cole, T. (1998). "Fitting smoothed centile curves to reference data." *J. R. Statist. Soc.* A, 151: 385–418.

Cole, T. and Green, P. (1992). "Smoothing reference centile curves: The lms method and penalized likelihood." *Statistics in medicine*, 11: 1305–1319.

Craven, P. and Wahba, G. (1979). "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation." *Numerical Mathematics*, 31: 377–403.

Dabo-Niang, S. and Laksaci, A. (2012). "Nonparametric quantile regression estimation for functional dependent data." *Numerical Mathematics*, 41: 1254–1268.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

Delbaen, F., Bellini, F., Bignozzi, V., and Ziegel, J. (2016). "Risk measures with the CxLS property." *Finance and Stochastics*, 20: 433–453.

Dette, H. and Volgushev, S. (2008). "Non-crossing non-parametric estimates of quantile curves." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3): 609–627.

Dunson, D. B., Herring, A. B., and Siega-Riz, A. M. (2008). "Bayesian Inference on Changes in Response Densities Over Predictor Clusters." *Journal of the American Statistical Association*, 103(484): 1508–1517.

Dunson, D. B. and Taylor, J. A. (2005). "Approximate Bayesian inference for quantiles." *Nonparametric Statistics*, 17(3): 385–400.

Efron, B. (1991). "Resgression percentiles using asymmetric squared error loss." *Statistica Sinica*, 1: 93–125.

Ehm, W., Gneiting, T., Jordan, A., and Krüger (2016). "Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings." *Journal of the Royal Statistical Society, Series B*, 78: 505–562.

Erkanli, A. (1994). "Laplace approximations for posterior expectations when the mode occurs at the boundary of the parameter space." *Journal of the American Statistical Association*, 89(425): 250–258.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." In *Kdd*, volume 96, 226–231.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Fatti, L., Senaoana, E., and Thompson, M. L. (1998). "Bayesian updating in reference centile charts." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 103–115.

Fenske, N., Kneib, T., and Hothorn, T. (2011). "Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression." *J. Am. Statist. Assoc*, 106: 494–510.

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The annals of statistics*, 209–230.

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

Fitzenberger, B., Koenker, R., and Machado, J. (2013). *Economic Applications of Quantile Regression*. Physica-Verlag HD.

Ford, E. S. (2005). "Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome." *Diabetes care*, 28(7): 1769–1778.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). "Mixtures of Shifted AsymmetricLaplace Distributions." *IEEE transactions on pattern analysis and machine intelligence*, 36(6): 1149–1157.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6: 721–741.

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

Gerlach, R., Chen, C., and Chan, N. (2011). "Bayesian Time-Varying Quantile Forecasting for Value-at-Risk in Financial Markets." *Journal of Business & Economic Statistics*, 29: 481–492.

Givens, G. H. and Hoeting, J. A. (2012). *Computational statistics*, volume 710. John Wiley & Sons.

Gower, J. C. and Ross, G. (1969). "Minimum spanning trees and single linkage cluster analysis." *Applied statistics*, 54–64.

Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 711–732.

Haffner, S. M., Stern, M. P., Hazuda, H. P., Mitchell, B. D., and Patterson, J. K. (1990). "Cardiovascular risk factors in confirmed prediabetic individuals: does the clock for coronary heart disease start ticking before the onset of clinical diabetes?" *Jama*, 263(21): 2893–2898.

Hall, P., Wolff, R. C., and Yao, Q. (1999). "Methods for estimating a conditional distribution function." *Journal of the American Statistical Association*, 94(445): 154–163.

Hardle, W. and Mammen, E. (1993). "Comparing nonparametric versus parametric regression fits." *The Annals of Statistics*, 21: 1926–1947.

Hastie, D. I., Liverani, S., Azizi, L., Richardson, S., and Stücker, I. (2013). "A semiparametric approach to estimate risk functions associated with multi-dimensional exposure profiles: application to smoking and lung cancer." *BMC Medical Research Methodology*, 13(129): 463–492.

Hastie, D. I., Liverani, S., and Richardson, S. (2015). "Sampling from Dirichlet process mixture models with unknown concentration parameter: mixing issues in large data implementations." *Statistics and Computing*, 25(5): 1023–1037.

Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*, 57(1): 97–109.

He, X. (1997). “Quantile curves without crossing.” *The American Statistician*, 51(2): 186–192.

He, X. and Ng, P. (1999). “COBS: Qualitatively constrained smoothing via linear programming.” *Computational Statistics*, 14(3): 315–338.

He, X., Ng, P., and Portnoy, S. (1998). “Bivariate quantile smoothing splines.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3): 537–550.

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

Hewson, P. and Yu, K. (2008). “Quantile regression for binary performance indicators.” *Applied Stochastic Models in Business and Industry*, 24: 401–418.

Hinneburg, A., Keim, D. A., et al. (1998). “An efficient approach to clustering in large multimedia databases with noise.” In *KDD*, volume 98, 58–65.

Horowitz, J. L. and Lee, S. (2005). “Nonparametric estimation of an additive quantile regression model.” *Journal of the American Statistical Association*, 100(472): 1238–1249.

— (2007). “Nonparametric instrumental variables estimation of a quantile regression model.” *Econometrica*, 75(4): 1191–1208.

Hosmer, D., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). “Polynomial spline estimation and inference for varying coefficient models with longitudinal data.” *Statistica Sinica*, 763–788.

Hurvich, C., Simonoff, J. S., and Tsai, C. (1998). “Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion.” *Journal of the Royal Statistical Society: B*, 60: 271–293.

Jones, M. and Yu, K. (2007). “Improved double kernel local linear quantile regression.” *Statistical Modelling*, 7(4): 377–389.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kim, T.-H. and Muller, C. (2000). *Two-stage quantile regression*. Univ., Department of Economics.

Kim, T.-H. and White, H. (2003). “Estimation, inference, and specification testing for possibly misspecified quantile regression.” In *Maximum likelihood estimation of misspecified models: twenty years later*, 107–132. Emerald Group Publishing Limited.

Kneib, T. (2013). “Beyond mean regression.” *Statistical modelling*, 13: 275–303.

Knight, C. A. and Ackerly, D. D. (2002). "Variation in nuclear DNA content across environmental gradients: a quantile regression analysis." *Ecology Letters*, 5(1): 66–76.

Koenker, R. (1984). "A note on L-estimates for linear models." *Statistics & probability letters*, 2(6): 323–325.

— (2005). *Quantile Regression*. New York: Cambridge University Press.

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

Koenker, R. and d'Orey, V. (1993). "A Remark on Computing Regression Quantile." *Applied statistics*, 43: 410–414.

Koenker, R. and Hallock, K. (2001). "Quantile regression: An introduction." *Journal of Economic Perspectives*, 15: 43–56.

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

Koenker, R. and Mizera, I. (2004). "Penalized triograms: Total variation regularization for bivariate smoothing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1): 145–163.

Koenker, R., Ng, P., and Portnoy, S. (1994). "Quantile smoothing splines." *Biometrika*, 81: 673–680.

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

Koenker, R. and Xiao, Z. (2002). "Inference on the quantile regression process." *Econometrica*, 70(4): 1583–1612.

— (2006). "Quantile autoregression." *Journal of the American Statistical Association*, 101(475): 980–990.

Kolberg, J. A., Jørgensen, T., Gerwien, R. W., Hamren, S., McKenna, M. P., Moler, E., Rowe, M. W., Urdea, M. S., Xu, X. M., Hansen, T., et al. (2009). "Development of a type 2 diabetes risk model from a panel of serum biomarkers from the Inter99 cohort." *Diabetes care*, 32(7): 1207–1212.

Kong, E. and Xia, Y. (2015). "Uniform Bahadur Representation for Nonparametric Censored Quantile Regression: A Redistribution-of-Mass Approach." *Econometric Theory*.

Kottas, A. and Gelfand, A. E. (2001). "Bayesian semiparametric median regression modeling." *Journal of the American Statistical Association*, 96(456): 1458–1468.

Kottas, A. and Krnjajić, M. (2009). "Bayesian Semiparametric Modelling in Quantile Regression." *Applied Probability & Statistics*, 36: 297–319.

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

Kukush, A., Beirlant, J., and Goegebeur, Y. (2005). "Nonparametric estimation of conditional quantiles."

Lee, D. and Neocleous, T. (2010). “Bayesian quantile regression for count data with application to environmental epidemiology.” *Journal of the Royal Statistical Society: Series C*, 59: 905–920.

Li, C., Ford, E. S., Zhao, G., and Mokdad, A. H. (2009). “Prevalence of pre-diabetes and its association with clustering of cardiometabolic risk factors and hyperinsulinemia among US adolescents.” *Diabetes care*, 32(2): 342–347.

Li, Q., Lin, J., and Racine, J. S. (2013). “Optimal bandwidth selection for nonparametric conditional distribution and quantile functions.” *Journal of Business & Economic Statistics*, 31(1): 57–65.

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

Li, Q. and Racine, J. S. (2008). “Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data.” *Journal of Business & Economic Statistics*, 26(4): 423–434.

Liu, Y. and Wu, Y. (2011). “Simultaneous multiple non-crossing quantile regression estimation using kernel constraints.” *Journal of nonparametric statistics*, 23(2): 415–437.

Liverani, S., Hastie, D., Azizi, L., Papathomas, M., and Sylvia, R. (2015). “PReMiuM: An R Package for Profile Regression Mixture Models using Dirichlet Processes.” *Journal of Statistical Software*, 64(2): 1–30.

Liverani, S., Lavigne, A., and Blangiardo, M. (2016). “Modelling collinear and spatially correlated data.” *Spatial and Spatio-temporal Epidemiology*, 18: 63–73.

Lum, K. and Gelfand, A. (2012). “Spatial Quantile Multiple Regression Using the Asymmetric Laplace Process.” *Bayesian Analysis*, 7: 235–258.

Machado, J. and Santos Silva, J. (2005). “Quantiles for Counts.” *Journal of the American Statistical Association*, 100: 1226–1237.

Machado, J. A. and Mata, J. (2005). “Counterfactual decomposition of changes in wage distributions using quantile regression.” *Journal of applied Econometrics*, 20(4): 445–465.

MacQueen, J. et al. (1967). “Some methods for classification and analysis of multivariate observations.” In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 281–297. Oakland, CA, USA.

Mattei, F., Liverani, S., Guida, F., Matrat, M., Cenée, S., Azizi, L., Menvielle, G., Sanchez, M., Pilorget, C., Lapôtre-Ledoux, B., et al. (2016). “Multidimensional analysis of the effect of occupational exposure to organic solvents on lung cancer risk: the ICARE study.” *Occupational and environmental medicine*, 73(6): 368–377.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). “Equation of state calculations by fast computing machines.” *The journal of chemical physics*, 21(6): 1087–1092.

Molitor, J., Brown, I. J., Chan, Q., Papathomas, M., Liverani, S., Molitor, N., Richardson, S., Van Horn, L., Daviglus, M. L., Dyer, A., Stamler, J., Elliott, P., and Group, I. R. (2014). "Blood Pressure Differences Associated With Optimal Macronutrient Intake Trial for Heart Health (OMNIHEART)–Like Diet Compared With a Typical American Diet." *Hypertension*, 64(6): 1198–1204.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). "Bayesian profile regression with an application to the National Survey of Children's Health." *Biostatistics*, kxq013.

Muggeo, V. M., Sciandra, M., Tomasello, A., and Calvo, S. (2013). "Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology." *Environmental and ecological statistics*, 20(4): 519–531.

Newey, W. and Powell, J. (1987). "Asymmetric Least Squares Estimation and Testing." *Econometrica*, 55: 819–847.

Ng, P. and Maechler, M. (2007). "A fast and efficient implementation of qualitatively constrained quantile smoothing splines." *Statistical Modelling*, 7(4): 315–328.

Noh, H., Ghouch, A. E., and Van Keilegom, I. (2015). "Semiparametric conditional quantile estimation through copula-based multivariate models." *Journal of Business & Economic Statistics*, 33(2): 167–178.

Papathomas, M., Molitor, J., Hoggart, C., Hastie, D., and Richardson, S. (2012). "Exploring data from genetic association studies using Bayesian variable selection and the Dirichlet process: application to searching for gene× gene patterns." *Genetic epidemiology*, 36(6): 663–674.

Parente, P. M. and Silva, J. M. S. (2015). "Quantile Regression with Clustered Data." *Journal of Econometric Methods, forthcoming* (in press).

Peng, L. and Huang, Y. (2010). "Survival analysis with quantile regression models." *Journal of the American Statistical Association*, 103: 637–649.

Pirani, M., Best, N., Blangiardo, M., Liverani, S., Atkinson, R. W., and Fuller, G. W. (2015). "Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles." *Environment International*, 79: 56–64.

Portnoy, S. (2003). "Censored Regression Quantiles." *Journal of the American Statistical Association*, 98: 1001–1012.

Portnoy, S., Koenker, R., et al. (1997). "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators." *Statistical Science*, 12(4): 279–300.

Qu, Z. and Yoon, J. (2015). "Nonparametric estimation and inference on conditional quantile processes." *Journal of Econometrics*, 185(1): 1–19.

Rahman, M. (2016). "Bayesian Quantile Regression for Ordinal Models." *Bayesian Analysis*, 11: 1–24.

Reed, C. and Yu, K. (2009). "A partially collapsed Gibbs sampler for Bayesian quantile regression." Technical report, Brunel University Mathematics Technical Report.

Reich, B. R., Bondell, H., and Wang, H. J. (2010). "Flexible Bayesian quantile regression for independent and clustered data." *Biostatistics*, 11: 337–352.

Robert, C. P. (2004). *Monte carlo methods*. Wiley Online Library.

Roy, D. (2003). "The discrete normal distribution." *Communications in Statistics-theory and Methods*, 32(10): 1871–1883.

Royston, P. and Sauerbrei, W. (2008). *Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables*, volume 777. John Wiley & Sons.

Ruppert, D., Sheather, S., and Wand, M. (1995). "An effective bandwidth selector for local least squares regression." *Journal of the American Statistical Association*, 90: 1257–1270.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge university press.

Schaumburg, J. (2012). "Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory." *Computational Statistics & Data Analysis*, 56: 4081–4096.

Serdyukova, N. (2012). "Spatial adaptation in heteroscedastic regression: Propagation approach." *Electron. J. Stat.*, 6: 861–907.

Sigrist, M. W. (1994). *Air monitoring by spectroscopic techniques*, volume 127. John Wiley & Sons.

Smith, A. F. and Roberts, G. O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods." *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–23.

Smith, L., Reich, B., Herring, A., Langlois, P., and Fuentes, M. (2015). "Multilevel Quantile Function Modeling with Application to Birth Outcomes." *Biometrics*, 71: 508–519.

Smith, M. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Journal of Econometrics*, 75(2): 317–343.

Späth, H. (1980). "Cluster analysis algorithms for data reduction and classification of objects."

Spokoiny, V. and Vial, C. (2009). "Parameter tuning in pointwise adaptation using a propagation approach." *Ann. Statist.*, 37: 2783–2807.

Spokoiny, V., Wang, W., and Härdle, W. (2014). "Local quantile regression." *Journal of Statistical Planning and Inference*, 37: 1109–1129.

Sriram, K., Ramamoorthi, R., Ghosh, P., et al. (2013). "Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density." *Bayesian Analysis*, 8(2): 479–504.

Sriram, K., Shi, P., and Ghosh, P. (2016). "A Bayesian quantile regression model for insurance company costs data." *Journal of the Royal Statistical Society, Series A*, 179: 177–202.

Stephens, M. (2000). "Dealing with label switching in mixture models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809.

Steptoe, A., Breeze, E., Banks, J., and Nazroo, J. (2012). "Cohort profile: the English longitudinal study of ageing." *International journal of epidemiology*, 42(6): 1640–1648.

Tabák, A. G., Herder, C., Rathmann, W., Brunner, E. J., and Kivimäki, M. (2012). "Prediabetes: a high-risk state for diabetes development." *The Lancet*, 379(9833): 2279–2290.

Taddy, M. A. and Kottas, A. (2010). "A Bayesian Nonparametric Approach to Inference for Quantile Regression." *Journal of Business & Economic Statistics*, 28: 357–369,.

Takezawa, K. (2005). *Introduction to nonparametric regression*, volume 606. John Wiley & Sons.

Tasche, D. (2001). "Conditional expectation as quantile derivative." *arXiv preprint math/0104190*.

Taylor, J. and Yu, K. (2016a). "Using Autoregressive Logit Models to Forecast the Exceedance Probability for Financial Risk Management." *Journal of the Royal Statistical Society, Series A*. (In press).

Taylor, J. W. and Yu, K. (2016b). "Using auto-regressive logit models to forecast the exceedance probability for financial risk management." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Theil, H. (1966). "Applied economic forecasting."

Thompson, P., Cai, Y., Moyeed, R., Reeve, D., and Stander, J. (2010). "Bayesian nonparametric quantile regression using splines." *Computational Statistics and Data Analysis*, 54: 1138–1150.

Tsai, I. (2012). "The relationship between stock price index and exchange rate in Asian markets: A quantile regression approach." *Journal of international financial markets, institutions & money*, 22: 609–621.

Tsionas, E. G. (2003). "Bayesian quantile inference." *Journal of statistical computation and simulation*, 73(9): 659–674.

Walker, S. and Mallick, B. K. (1999). "A Bayesian semiparametric accelerated failure time model." *Biometrics*, 55(2): 477–483.

Waltrup, L., Sobotka, F., Kneib, T., and Kauermann, G. (2015). "Expectile and quantile regression-David and Goliath?'." *Statistical modelling*, 15: 433–456.

Wand, M. and Jones, M. (1995). "Kernel smoothing. 1995." *Chapman&Hall, London*.

Wang, W., Yang, J., Muntz, R., et al. (1997). "STING: A statistical information grid approach to spatial data mining." In *VLDB*, volume 97, 186–195.

Wei, Y., Pere, A., Koenker, R., and He, X. (2006). "Quantile regression methods for reference growth charts." *Statistics in medicine*, 25: 1369–1382.

Welsh, A. (1996). "Robust estimation of smooth regression and spread functions and their derivatives." *Statistica Sinica*, 347–366.

Wolkewitz, M., Zortel, M., Palomar-Martinez, M., Alvarez-Lerma, F., Olaechea-Astigarraga, P., and Schumacher, M. (2017). "Landmark prediction of nosocomial infection risk to disentangle short-and long-stay patients." *Journal of Hospital Infection*, 96(1): 81–84.

Wu, T. Z., Yu, K., and Yu, Y. (2010). "Single-index quantile regression." *Journal of Multivariate Analysis*, 101(7): 1607–1621.

Wu, Y. and Liu, Y. (2009). "Stepwise multiple quantile regression estimation using non-crossing constraints." *Statistics and Its Interface*, 2: 299–310.

Wu, Y. and Yin, G. (2015). "Conditional quantile screening in ultrahigh-dimensional heterogeneous data." *Biometrika*, 102: 65–76.

Xiao, Z. and Koenker, R. (2009). "Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models." *Journal of the American Statistical Association*, 104(488): 1696–1712.

Xu, Q., Liu, X., Jiang, C., and Yu, K. (2016). "Quantile autoregression neural network model with applications to evaluating value at risk." *Applied Soft Computing*, 49: 1–12.

Yang, Y. and He, X. (2012). "Bayesian empirical likelihood for quantile regression." *The Annals of Statistics*, 40: 1102–1131.

Yang, Y., Wang, H., and He, X. (2016). "Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood." *International Statistical Review*. (In press).

Yao, Q. and Tong, H. (1996). "Asymmetric least squares regression estimation: A nonparametric approach." *Journal of Nonparametric Statistics*, 6(2-3): 273–292.

Yu, K. and Jones, M. (1997). "A comparison of local constant and local linear regression quantile estimators." *Computational Statistics & Data Analysis*, 25(2): 159–166.

Yu, K. and Jones, M. C. (1998). "Local linear quantile regression." *Journal of the American Statistical Association*, 93: 228–237.

Yu, K., Liu, X., Alhamzawi, R., Becker, F., and Lord, J. (2016). "Statistical methods for body mass index: a selective review." *Statistical methods in medical research*, 0962280216643117.

Yu, K. and Lu, Z. (2004). "Local linear additive quantile regression." *Scandinavian Journal of Statistics*, 31(3): 333–346.

Yu, K., Lu, Z., and Stander, J. (2003). "Quantile regression: applications and current research areas." *The Statistician*, 52: 331–350.

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

Yu, K. and Stander, J. (2007). "Bayesian analysis of a Tobit quantile regression model." *Journal of Econometrics*, 137: 260–276.

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*, 359–377.

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

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

Yue, Y. R. and Rue, H. (2011). "Bayesian inference for additive mixed quantile regression models." *Computational Statistics & Data Analysis*, 55(1): 84–96.

Zhang, J., Zhang, R., and Lu, Z. (2016). "Quantile-adaptive variable screening in ultra-high dimensional varying coefficient models." *Journal of Applied Statistics*, 43: 643–654.

Ziegel, J. (2016). "Coherence and Elicitability." *Mathematical Finance*. (In press).