Regularized and Robust Regression Methods for High-Dimensional Data

A thesis submitted for degree of

Doctorate of Philosophy

by

Hussein Hashem

B.Sc., M.Sc.

Supervised by

Dr. Veronica Vinciotti

Department of Mathematical Sciences

School of Information System, Computing and Mathematics

July- 2014

Abstract

Recently, variable selection in high-dimensional data has attracted much research interest. Classical stepwise subset selection methods are widely used in practice, but when the number of predictors is large these methods are difficult to implement. In these cases, modern regularization methods have become a popular choice as they perform variable selection and parameter estimation simultaneously. However, the estimation procedure becomes more difficult and challenging when the data suffer from outliers or when the assumption of normality is violated such as in the case of heavy-tailed errors. In these cases, quantile regression is the most appropriate method to use. In this thesis we combine these two classical approaches together to produce regularized quantile regression methods.

Chapter 2 shows a comparative simulation study of regularized and robust regression methods when the response variable is continuous. In chapter 3, we develop a quantile regression model with a group lasso penalty for binary response data when the predictors have a grouped structure and when the data suffer from outliers. In chapter 4, we extend this method to the case of censored response variables. Numerical examples on simulated and real data are used to evaluate the performance of the proposed methods in comparisons with other existing methods.

Certificate of Originality

I hereby certify that the work presented in this thesis is my own research and has not been presented for a higher degree at any other university or institute. Any material that could be construed as the work of others is fully cited and appears in the references.

Hussein Hashem

To my parents

Acknowledgements

This thesis would not have been possible without the help and support of a number of people. First, I would like to thank my thesis supervisor, Dr.Veronica Vinciotti, for her support, guidance and for encouraging me to develop my own ideas. Also my special thanks go to Dr. Keming Yu for all his support, and helping me significantly with my research as a second supervisor.

I would like to express my gratitude to the administrator staff and research students at the department of mathematics in Brunel University. Specifically, I would like to thank Dr. Rahim Al-Hamzawi and Dr. Ali Al-Kenani for their useful discussions, advises and support, both as a student and friend.

Finally, I would like to thank my family for the great support during my PhD study. I want to thank my wife for her encouragement, her support and patience over the years.

Author's Publications

1. V. Vinciotti and H. Hashem (2013). Robust methods for inferring sparse network structures. *Computational Statistics and Data Analysis, 67,* 84-94.

2. H. Hashem, V. Vinciotti, R. Al-Hamzawi and K. Yu (2013). Binary quantile regression with group lasso. *Submitted to Statistics: A Journal of Theoretical and Applied Statistics.*

Table of Contents

Abstract ii

Declaration iii

Acknowledgements v

Author's Publication vi

1. Introduction 1

1.1 Regularization regression methods 4

1.2 Robust regression methods 7

1.2.1 Quantile regression methods 13

1.2.2 Robust and regularized regression methods 15

1.3 Thesis outline 16

2. Regularized Robust Regression Methods for Continuous Response Variables 18

2.1 Classical regularized regression methods 18

2.1.1 Ridge regression 18

2.1.2 LASSO 20

2.1.3 Bridge regression 22

2.1.4 Elastic net 22

2.2.5 Adaptive lasso 23

2.1.6 SCAD 23

2.1.7 Group lasso 24

2.2 Robust regularized regression methods 25

2.3 Comparison of robust and regularized regression methods on simulated data 28

2.3.1 Example 1: low- dimensional 29

2.3.2 Example 2: high- dimensional 30

2.4 Comparison of Bayesian regularized quantile regression methods with classical methods on simulated data. 33

2.4.1 Example 3: low- dimensional with sparse coefficients 33

2.4.2 Example 4: high – dimensional with sparse coefficients 34

2.4.3 Example 5: simulation with non-sparse coefficients 35

2.5 Concluding remarks 36

3. Binary quantile regression with group lasso 38

3.1 Introduction 38

3.2 Binary quantile group lasso 40

3.3 Bayesian parameter estimation 41

3.4 Class prediction 49

3.5 Simulation study 50

3.6 Real application 60

3.6.1 Birth weight dataset 60

3.6.2 Colon dataset 61

3.6.3 Labor force participation dataset 61

3.6.4 Splice site detection dataset 62

3.6.5 Cleveland heart dataset 62

3.7 Chapter conclusion 66

4. Tobit quantile regression with group lasso 67

4.1 Introduction 67

4.2 The model 69

4.3 Bayesian parameter estimation 70

4.4 Computing predicted values 74

4.5 Simulation study 77

4.6 Real application 86

4.7 Conclusion 89

5. Conclusions and Future Research **91**

5.1 Main Contributions 91

5.2 Recommendations for Future Research 92

Bibliography **93**

List of Figures

Figure 1.1: Objective (left), ψ (center), and weight (right) functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are $M = 1.345$ for the Huber estimator and $M = 4.685$ for the bisquare……..…..12

Figure 1.2: Check function for three values of θ for quantile regression. For $\theta = 0.5$, positive and negative errors are treated symmetrically, but for the other values of θ, positive and negative errors are treated asymmetrically……..……..……14

Figure 2.1: Comparison of regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for example 1 and the bottom panels the average true positives when $p = 15$ and $n = 100$ ……..……..……30

Figure 2.2: Comparison of regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for example 2 and the bottom panels the average true positives when $p = 100$ and $n = 50$……..……32

Figure 2.3: Comparison of Bayesian quantile regression methods with frequentist methods, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 3 when $p = 15$ and $n = 100$……..…33

Figure 2.4: Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 4 when $p = 100$ and $n = 50$…….... 34

Figure 2.5: Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 5 when $p = 50$ and $n = 100$……35

Figure 2.6: Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 5 when $p = 100$ and $n = 50$……36

Figure 3.1: Some density functions of the errors considered in the simulation study……53

Figure 3.2: Trace plots for some selected $\hat{\beta}_j's$ at r=0.5 and quantile 0.5 for simulation case1. The horizontal line refers to $\frac{\beta_j}{\|\beta_j\|}$ ……..……55

Figure 3.3: Average ROC curves (over 40 replications) of Bayesian binary quantile regression with group lasso (BBQ. grplasso, $\theta = 0.5$), compared with grpreg, glmnet and bayesQR, under a Skewed (left top panel), a Laplace (right top panel), t_1 (left bottom panel) and Kurtotic (right bottom panel) error distribution.. ……60

Figure 4.1: Comparison of Bayesian tobit quantile regression with group lasso (BTQ. grplasso) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso) under normal and kurtotic error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for the simulation study when $p = 15$, n=100 and $\theta = 0.5$.....79

Figure 4.2: Bias and variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a normal distribution for the error, $p = 15$, $n = 100$ and $\theta = 0.5$……...80

Figure 4.3: Bias and variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, p = 15, n = 100 and θ = 0.5 ……..……..81

Figure 4.4: Comparison of Bayesian tobit quantile regression with group lasso (BTQ. grplasso) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso) under normal and kurtotic error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for the simulation study when p = 100, n=100 and θ = 0.5 82

Figure 4.5: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a normal distribution for the error, r = 0.5, p = 100, n = 100 and θ = 0.5.....82

Figure 4.6: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a normal distribution for the error, r= 0.5, p = 100, n = 100 and θ = 0.5……83

Figure 4.7: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a normal distribution for the error, r = 0.95, p = 100, n = 100 and θ = 0.5……..……83

Figure 4.7: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a normal distribution for the error, r = 0.95, p = 100 , n = 100 and θ = 0.5…….. 83

Figure 4.8: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, r = 0.5, p = 100 ,n = 100 and θ = 0.5……..……..84

Figure 4.9: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, r = 0.5 ,p = 100 ,n = 100 and θ = 0.5……..….84

Figure 4.10: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, r = 0.95 , p = 100 , n = 100 and θ = 0.5 ……..…..85

Figure 4.11: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, r = 0.95 , p = 100 and θ = 0.5 . ……..….....85

Figure 4.12: Comparison of Bayesian tobit quantile regression with group lasso (BTQ. grplasso) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso) under normal and kurtotic error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for the simulation study when p = 100, n=100 and θ = 0.5 86

List of Tables

Table 1.1: Objective functions and weight functions for least-squares, Huber, and biweight estimators.....11

Table 3.1: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.5$ and β values as in case (1). BBQ. grplasso: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); grpreg: frequentist mean-based logistic regression model with group lasso penalty, glmnet: frequentist mean-based logistic regression model under a group lasso penalty; bayesQR: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles). Best mean indicated in bold……...……..56

Table 3.2: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.95$ and β values as in case (1). BBQ. grplasso : Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); grpreg: frequentist mean-based logistic regression model with group lasso penalty, glmnet : frequentist mean-based logistic regression model under a group lasso penalty; bayesQR: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles)…….. 57

Table 3.3: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.5$ and β values as in case (2). BBQ. grplasso: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); grpreg: frequentist mean-based logistic regression model with group lasso penalty, glmnet: frequentist mean-based logistic regression model under a group lasso penalty; bayesQR: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles). Best mean indicated in bold ……...….58

Table 3.4: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.95$ and β values as in case (2). BBQ. grplasso : Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); grpreg: frequentist mean-based logistic regression model with group lasso penalty, glmnet : frequentist mean-based logistic regression model under a group lasso penalty; bayesQR: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles). Best mean indicated in bold ……59

Table 3.5: Variables in the labor force participation dataset……...…….. 62

Table 3.6: Variables in the heart disease dataset……...…… 63

Table 3.7: AUC values, averaged over 5 replications (with standard deviations in brackets) on real data: BBQ. grplasso: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); grpreg: frequentist mean-based logistic regression model with group lasso penalty, glmnet: frequentist mean-based logistic regression model under a group lasso penalty; bayesQR: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles)…….... 64

Table 3.8: 95% credible intervals for birth dataset at $\theta = 0.5$……...……..... 65

Table 3.9: 95% credible intervals for birth dataset at $\theta = 0.95$……65

Table 4.1: 99% credible intervals for labor force participation dataset at $\theta = 0.5$. ……..…..88

Table 4.2: 99% credible intervals for labor force participation dataset at $\theta = 0.95$. ……..….89

Chapter 1

Introduction

Variable selection is important for high-dimensional data analysis in many research areas such as biology (Peng et al., 2010), signal processing (Lustig et al., 2008) and collaborative filtering (Koren et al., 2009). For example, microarray experiments allow one to measure thousands of variables (genes, proteins) simultaneously. The data sets generated by these experiments are generally very large in terms of the number of predictors (p) and often small in terms of the number of biological samples (n). In regression analysis, this problem is often termed as the "large p and small n problem" ($p \gg n$) and presents a major barrier to traditional statistical methods.

With the development of computer and data collection technologies, the database sizes continue to grow and various statistical methodologies have been developed over the past several decades to cope with the challenges presented by these data. In particular, there are major challenges in parameter estimation, model and variable selection.

In classical multiple regression, model selection procedures, such as forward, backward, stepwise selection and all subset regression, are not suitable in a high-dimensional data. Furthermore, the least squares method, which is widely used for regression modelling, is not appropriate when the assumption of normality is violated such as in the case of heavy-tailed errors under a large number of predictors. To overcome these drawbacks, several regularized regression methods and robust methods have been proposed for fitting multiple regression models, particularly for the case when $p \gg n$ where the least squares method cannot be used.

Hoerl and Kennard (1970) proposed ridge regression by adding an L_2- penalty to the least squares loss function. Although ridge regression can produce accurate estimates under a large number of predictors, it cannot perform variable selection simultaneously, and hence classical model selection procedures have to be used for selecting an optimal model. In order to overcome this limitation, Tibshirani (1996) proposed LASSO (Least Absolute Shrinkage and Selection Operator), which minimizes the residual sum of squares subject to an L_1 -norm constraint. The lasso penalty results into some coefficients being estimated to exactly zero, thus performing estimation and variable selection simultaneously. Following from the seminal paper of Tibshirani (1996), various extensions of lasso were developed, such as elastic net (Zou and Hastie, 2005), which combines the L_1- penalty (lasso) and the L_2- penalty (ridge), adaptive lasso (Zou, 2006), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), etc. The estimates of regression coefficients by the lasso methods cannot be derived analytically because the L_1- penalty term is not differentiable. To solve this problem, several efficient algorithms were proposed, for example, Lars and coordinate descent algorithms (Efron et al., 2004, Friedman et al., 2010).

Most methods in the literature are focused on the mean regression, which means that the relationship between the response variable and predictor variables is summarized by describing the mean of the response, for each fixed value of the predictors, using a function (conditional mean function) of the response.

Quantile regression, introduced by Koenker and Bassett (1978), can be used when an estimate of the various quantiles (such as the median) of a conditional distribution is of interest. This allows one to see and compare how some quantiles of the response variable may be more affected by some predictor variables than other quantiles.

Modelling quantiles, rather than the mean, makes quantile regression models more robust to outliers, than linear regression (mean regression) models (Reed, 2011). Furthermore, quantile regression provides a more complete picture of the conditional distribution of y given x when both lower and upper or all quantiles are of interest, as in the analysis of body mass index where both lower (underweight) and upper (overweight) quantiles are used to check health standards.

Some methods have combined regularized and robust regression methods in order to perform variable selection in high-dimensional data with outliers. For example, Rosset and Zhu (2007) proposed the Huber lasso method which combines the Huber's criterion loss with a lasso penalty. The LAD-adaptive lasso method is proposed by Wang et al. (2007a), combining the idea of Least Absolute Deviance (LAD) and adaptive lasso. Bradic and Fan (2011) introduce a new penalized quasi-likelihood estimator for robust linear models for high dimensional data. Lambert-Lacroix and Zwald (2011) developed the Huber's Criterion with adaptive lasso which combines the Huber's loss function and adaptive lasso penalty. Arslan (2012) developed and investigated the properties of weighted LAD-lasso method which combines the idea of the Weighted Least Absolute Deviation (WLAD) regression estimation method and the adaptive lasso for robust parameter estimation and variable selection. In chapter 2 we will give an overview and detail of these methods. In the next two sections we will give an overview of regularized and robust regression methods, respectively.

1.1 Regularized regression methods

We start from the classical linear regression model to describe the regularized regression methods. A classical linear regression model has the following form (Hubert and Rousseeuw, 1997)

$y_i = x_i^T \beta + u_i \,, i = 1,\ldots,n$, with $u_i \sim N(0, \sigma^2)$ $, i = 1,\ldots,n$, (1.1)

where y_i is the response for the ith sample, x_i is a p vector of predictors or the explanatory variables, and β is a p vector unknown coefficients which we want to estimate. The most popular estimation method is the Ordinary Least Squares (OLS), in which the coefficients

$\hat{\beta}_{ls} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)$ minimize the residual sum of squares

$$\sum_{i=1}^{n}(y_i - x_i^T \beta)^2 \,. \tag{1.2}$$

All the methods described in this thesis use standardized variables; therefore the intercept β_0 is usually not included in the penalty. This can be done by first centring the inputs and response variables. That is,

$\sum_{i=1}^{n} y_i = 0$, $\sum_{i=1}^{n} x_{ij} = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1,\ldots,p$.

Denote by X the $n \times p$ matrix with each column the values of the corresponding predictor, and similarly let y be the vector of observation for the response variable. Then $\hat{\beta}_{ls}$ satisfies

$$X^T X \, \hat{\beta}_{ls} = X^T y \tag{1.3}$$

and assuming that X has full column rank $p(p \leq n \; and \; X^T X$ is positive definite and can be inverted), we obtain a unique solution for the regression coefficients

$$\hat{\beta}_{ls} = (X^T X)^{-1} X^T y \, .$$

When multicollinearity problems among the predictors are present or when $p > n$, the matrix X and the matrix $(X^T X)$ do not have full rank. Thus, the inverse $(X^T X)^{-1}$ cannot be calculated, equation (1.3) cannot be solved and the OLS estimator has no unique solution (Flexeder, 2010). Even in cases when the estimate can be obtained, there are two reasons why the data analyst is often not satisfied with these OLS estimates (Hastie et al., 2009). The first is prediction accuracy: the OLS tends to give estimators with low biases but high variances and better prediction accuracy can usually be obtained by lowering the variance with a little increased bias. This can be achieved by shrinking or setting some coefficients to exact zero. The second reason is interpretation: with a large number of predictors, we often would like to determine a smaller subset that shows the strongest effects.

There are two standard techniques for improving the OLS estimates (Tibshirani, 1996): the first technique is to use subset selection, such as stepwise procedures. Despite providing interpretable models, stepwise procedures can be extremely variable because they are based on a discrete process where predictors are either retained or dropped from the model. Small changes in the data can result in very different models being selected and this can reduce its prediction accuracy. The second technique is to find estimates of the regression coefficients by minimizing the residual sum of squares plus a penalty involving the size of the βs. These methods may set some βs exactly to zero thus performing also variable selection.

Motivated by these considerations, regularised regression approaches were developed. In these approaches, the coefficients $\hat{\beta}$ are found as the minimum of the penalized least squares loss defined by

$$\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda Pen(\beta), \tag{1.4}$$

where Pen is the penalty function. Several penalty functions exist such as lasso (Tibshirani, 1996), elastic net (Zou and Hastie, 2005) and adaptive lasso (Zou, 2006). We review some of them in chapter 2.

In regularized regression model the problem of choosing the regularization parameter λ is very important and needs to be taken in consideration. Several classical model selection criteria have been applied to select the parameter λ, such as Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), and Generalized Cross-Validation (GCV; Craven and Wahba, 1978) as well as K-fold cross-validation methods (see for example Tibshirani (1996), Fan and Li (2001), Zou (2006), Wang et al.(2007b) and Lazaridis (2008) for applications of these methods in regularized regression models). More in detail, the criteria are defined as follow:

$$Cp = \frac{SSE}{n} + \frac{2\hat{\sigma}^2}{n}\, df \qquad (Cp\text{ , Mallows, 1973}),$$

$$AIC = \left(\frac{SSE}{n\hat{\sigma}^2}\right) + \frac{2}{n}\, df \qquad (\text{AIC, Akaike, 1973}),$$

$$BIC = \left(\frac{SSE}{n\hat{\sigma}^2}\right) + \frac{\log(n)}{n}\, df \qquad (\text{BIC, Schwarz, 1978}).$$

where: SSE is the sum of squared errors of the model with predictors p, $SSE = \left\|y - x^T\hat{\beta}\right\|^2$, n is the number of observations, $\hat{\sigma}^2$ is the estimated conditional variance

$\hat{\sigma}^2 = \frac{\|y - x^T\hat{\beta}\|^2}{n}$, and df are the degrees of freedom, which in this context are the number of non-zero coefficients in $\hat{\beta}$ (Zou et al., 2007). Next, we discuss the cross-validation method.

K-fold cross validation is a popular method for estimating the prediction error and comparing different models. K-fold cross-validation uses one part of the training data to fit the model and a second part to test the model. The general idea of K-fold cross-validation is to divide the data into K-folds and leave one fold out to calculate the prediction error. So we split our data (x_i, y_i) into K equal parts. Then for each $k = 1, \ldots, K$, we remove the kth part from our data set, and fit a model and predict $\hat{f}^{-k}(x, \lambda)$. Let C_k be the indices of observations in the *kth* fold. The cross-validation estimate of the expected test error is (Tibshirani and Tibshirani, 2009)

$$CV(\lambda) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in C_k} \left\| y - \hat{f}^{-k}(x, \lambda) \right\|^2). \tag{1.5}$$

We repeat this for a grid of λ values, and choose the minimizer $\hat{\lambda}$ to be our choice of estimate λ. In expression (1.5), the cross-validation function is written in terms of the squared loss. The cross-validation function can also be written in terms of the log-likelihood function.

1.2 Robust regression methods

The performance of the Ordinary Least Squares (OLS) method can be very poor when the error has a heavy tailed distribution which may arise as a result of outliers. Rousseeuw and Leroy (1987) define three types of outliers that can affect the OLS estimator: vertical outliers, bad leverage points and good leverage points. Vertical outliers are those observations that have outlying values for the response variable y but

are not outlying in the explanatory variables x. Their presence affects the OLS estimation and in particular the estimated intercept. Good leverage points are observations that are outlying in the explanatory variables but that are located close to the regression line. Their presence does not affect the OLS estimation but it affects the estimated standard errors. Finally, bad leverage points are observations that are both outlying in the explanatory variables and located far from the true regression line. Their presence affects significantly the OLS estimation of both the intercept and the slope. Because the OLS is very sensitive to these outliers, robust regression is a form of regression analysis designed to solve some limitations of classical methods in the presence of outliers. Researchers have developed many robust methods to deal with this problem, amongst these Huber's M-Estimators (Huber, 1964), MM-estimators (Yohai, 1987), Least Median of Squares estimators and Least Trimmed Squares estimators (Rousseeuw, 1984), S-estimators (Rousseeuw and Yohai, 1984) and quantile regression methods (Koenker and Bassett, 1978).

The least squares estimator is obtained by minimising a function of the residuals, which is equivalent to considering a likelihood function under an assumption of normal distribution of the errors. M-estimation is based on the idea that, whilst we still want a maximum likelihood estimator, the errors might be better represented by a different, heavier-tailed, distribution. If this probability distribution function is $f(u_i)$ then the maximum likelihood estimator for β is that which maximises the likelihood function

$$\prod_{i=1}^{n} f(u_i) = \prod_{i=}^{n} f(y_i - x_i^T \beta).$$

This means it also maximises the log-likelihood function

$$\sum_{i=1}^{n} log\, f(u_i) = \sum_{i=1}^{n} log\, f(y_i - x_i^T \beta).$$

When the errors are normally distributed it has been shown that this leads to minimising the function $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$. Assuming that the errors are distributed differently leads to the maximum likelihood estimator minimising a different function. Using this idea, an M-estimator can be defined by

$$\min_{\beta} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) \tag{1.6}$$

where $\rho(.)$ is an outlier resistant loss function called the objective function. Possible choices for $\rho(.)$ should have the following properties:

■Always-non negative, $\rho(t) \geq 0$

■ Equal to 0 when its argument is 0, $\rho(0) = 0$

■Symmetric, $\rho(t) = \rho(-t)$

■Monotone in $|t_i|$, $\rho(t_i) \geq \rho(t_i')$ for $|t_i| \geq |t_i'|$.

Some special case are:

* $\rho(t) = t^2$, which gives the OLS estimator.

*$\rho(t) = \begin{cases} t^2 & if \quad |t| \leq M \\ 2M|t| - M^2 & if \quad |t| > M \end{cases}$, which gives the robust Huber estimator (Huber, 1981). M is normally tuned to 1.345.

*$\rho(t) = \begin{cases} 1 - [1 - (\frac{t}{M})^2]^3 & if \quad |t| \leq M \\ 1 & if \quad |t| > M \end{cases}$, which gives the Tukey Biweight estimator. M is normally chosen to 4.685 (Bai, 2004).

*$\rho_\theta(t) = \begin{cases} \theta t & if \quad t \geq 0 \\ -(1-\theta)t & if \quad t < 0 \end{cases}$, where $0 < \theta < 1$. This setting corresponds to quantile regression methods (Koenker and Bassett, 1978).

When $\theta = 0.5$, $\rho_{0.5}(t) = |t|$ gives median regression or least absolute deviations regression (LAD).

Finding an M-estimate requires partial differentiation of $\rho(t)$ with respect to each of the β parameters (Draper and Smith, 1998). Minimizing $\sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \beta}{\sigma})$ is equivalent to solving $\sum_{i=1}^{n} \psi(\frac{y_i - x_i^T \beta}{\sigma}) x_i = 0$, where σ is the standard deviation of the regression model, ψ is the derivative of ρ, $\psi(t) = \frac{dp(t)}{d(t)} = \rho'(t)$, and is called the score function. To facilitate computing, we would like to make this equation similar to the estimating equations for a familiar problem like weighted least squares. Define the weight function with $w_i = w(t_i) = \frac{\psi(t_i)}{t_i} = \frac{\psi(\frac{y_i - x_i^T \beta}{\sigma})}{(\frac{y_i - x_i^T \beta}{\sigma})}$. The estimating equations can then be written as

$$\sum_{i=1}^{n} w_i x_i (y_i - x_i^T \beta) \frac{1}{\sigma} = 0,$$

$$\Longrightarrow \sum_{i=1}^{n} w_i x_i (y_i - x_i^T \beta) = 0.$$

Defining the weight matrix $W = diag(w_i)$, $i = 1, \ldots, n$ as follows:

$$W = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix},$$

the above equations can be combined into the following matrix equation

$$X^T W X \, \hat{\beta}_{wls} = X^T W y.$$

Therefore, the estimator is given by $\hat{\beta}_{wls} = (X^T W X)^{-1} X^T W y$. (1.7)

In practice, this is very similar to the solution for the least squares estimator, but with the introduction of a weight matrix to reduce the influence of outliers. Generally, unlike least squares, equation (1.7) cannot be used to calculate an M-estimate directly from data, since W is unknown as it depends on the residuals. So iterative algorithms are used to solve this problem, where the estimator of β in the last iteration is used to calculate W and then W is used to obtain the estimator of β in the current iteration. This is the so called Iteratively Reweighted Least-Squares (IRLS) algorithm.

Several choices of the objective functions ρ have been proposed by various authors. Three of these are presented in table 1.1 together with the corresponding derivatives score function ψ and the resulting weight w. The objective functions, and the corresponding ψ and weight functions for the three estimators are also given in Figure 1.1 (Fox and Weisberg, 2010).

Table 1.1: Objective functions and weight functions for least-squares, Huber, and biweight sestimators.

Method	Objective Function	Weight Function
Least Square	$\rho_{LS}(t) = t^2$	$w_{LS}(t) = 1$
Huber	$\rho_H(t) = \begin{cases} \frac{1}{2}t^2 & if \quad \lvert t\rvert \le M \\ M\lvert t\rvert - \frac{1}{2}M^2 & if \quad \lvert t\rvert > M \end{cases}$	$w_H(t) = \begin{cases} 1 & if \quad \lvert t\rvert \le M \\ \frac{M}{\lvert t\rvert} & if \quad \lvert t\rvert > M \end{cases}$
Bisquare	$\rho_B(t) = \begin{cases} \frac{M^2}{6}\{1 - [1 - (\frac{t}{M})^2]^3\} & if \quad \lvert t\rvert \le M \\ \frac{M^2}{6} & if \quad \lvert t\rvert > M \end{cases}$	$w_B(t) = \begin{cases} [1 - (\frac{t}{M})^2]^2 & if \quad \lvert t\rvert \le M \\ 0 & if \quad \lvert t\rvert > M \end{cases}$

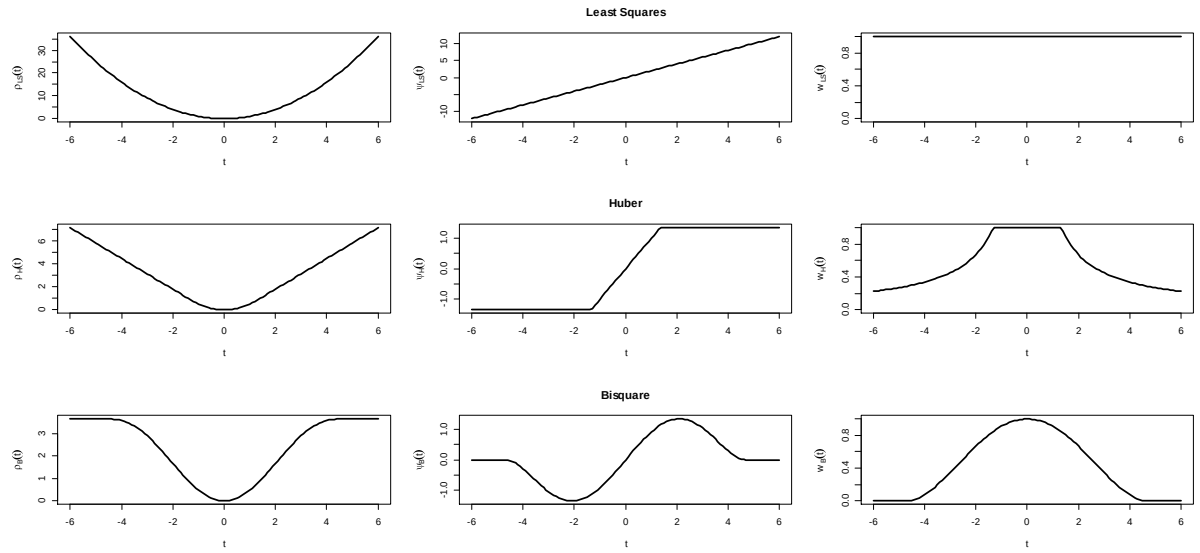

Figure 1.1: Objective (left), ψ (center), and weight (right) functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are $M = 1.345$ for the Huber estimator and $M = 4.685$ for the bisquare.

Figure 1.1 illustrates the different objective functions ρ with the corresponding derivatives score function ψ and the weight functions w for three M estimators: the least-squares estimator, the Huber estimator, and the Tukey bisquare (or biweight) estimator. Both the least-squares and Huber objective functions increase as the residual t departs from 0, but the least-squares objective function increases more quickly. In contrast, the bisquare objective function levels off for $|t| > M$. The robust weight functions give reduced weights at the tails compared to the least squares estimator, which gives weight one to all observations. This means that unusually large residuals have a much smaller effect on the estimate if using the least squares method. As a result M estimators are more robust to heavy-tailed error distributions.

The value M for the Huber and bisquare estimators is called a tuning constant; smaller values of M produce more resistance to outliers. The tuning constant is generally used to give reasonably high efficiency in the normal case. In particular, there are standard values (or ranges) for the tuning constants, resulting in estimators with

95% asymptotic relative efficiency in under the considerations, it was found that $M = 1.345\,\sigma$ for the Huber and $M = 4.685\,\sigma$ for the bisquare are appropriate choices (where σ is the standard deviation of the errors) (Fox and Weisberg, 2010). The standard values have been used in Figure 1.1.

1.2.1 Quantile regression methods

As we discussed in the previous section, we can use different objective function ρ in robust regression methods. A particular choice of the objective function ρ leads to quantile regression which now describe in detail. Let $\theta \in (0,1)$ be the quantile. Assume our model is given by $y_i = x_i^T \beta_\theta + u_i$ and that not the expected value, but the θth quantile of the error term conditional on the predictors is zero, i.e. $Q_\theta(u_i|x_i) = 0$. Then we assume that the θth conditional quantile of y with respect to x follows

$$Q_\theta(y|x) = x^T \beta_\theta.$$

The parameter vector β_θ can be estimated by

$$\hat{\beta}_\theta = \min_\beta \sum_{i=1}^n \rho_\theta(y_i - x_i^T \beta_\theta)\,,$$

where ρ_θ is the check function defined by

$$\rho_\theta(t) = \begin{cases} \theta t & if \quad t \geq 0, \\ -(1-\theta)t & if \quad t < 0 \end{cases} \tag{1.8}$$

or equivalently $\rho_\theta(t) = \frac{|t| + (2\theta - 1)t}{2}$.

Figure 1.2 shows the check function in equation (1.8) for three quantiles $0.25, 0.5\ and\ 0.75$.

The special case of the conditional median ($\theta = 0.5$) is well known and corresponds to

$$\hat{\beta}_{0.5} = \min_{\beta_{0.5}} \sum_{i=1}^{n} |y_i - x_i^T \beta_{0.5}|$$

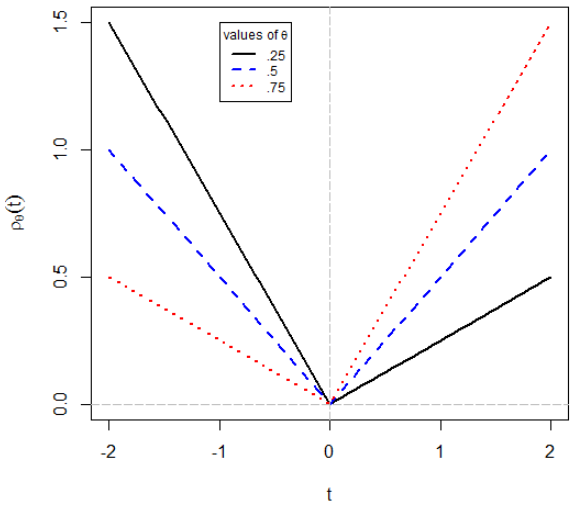

Figure 1.2: Check function for three values of θ for quantile regression. For $\theta = 0.5$, positive and negative errors are treated symmetrically, but for the other values of θ, positive and negative errors are treated asymmetrically.

Bayesian estimation for subset selection in standard mean regression suffers from many problems, for example assigning a prior for each subset in the model space and Bayesian computational efficiency (Alhamzawi and Yu, 2012). These difficulties become more challenging in quantile regression framework when one is interested in assigning prior distributions for the parameters.

Yu and Moyeed (2001) suggested a Bayesian quantile regression method where the errors are ALD distributed. This method is developed by the maximum posterior estimator under the ALD and the check function estimator of Koenker and Bassett (1978). This Bayesian approach has been extended by a number of researchers. For example, Yu and Stander (2007) developed a Bayesian estimation procedure for a

censored quantile regression, Benoit and Poel (2011) considered binary quantile regression from a Bayesian framework.

If we assume the error term u_i in the quantile regression has an Asymmetric Laplace Distribution (ALD) with $\mu = 0$, $\sigma > 0$ and $\theta \in (0,1)$ its pdf is given by (Yu and Zhang, 2005)

$$f(u;\mu = 0,\sigma,\theta) = \frac{\theta(1-\theta)}{\sigma}\exp\{-\rho_\theta(u)\},$$

where $\rho_\theta(u) = \frac{|u|+(2\theta-1)u}{2\sigma}$. It is known that when $p = 0.5$ the above probability density function is changed to the standard symmetric form of the Laplace density, that is

$$f(u;\mu = 0,\sigma,\theta = 0.5) = \frac{1}{4}\exp\{-\frac{u}{2\sigma}\}.$$

The expected value and variance of μ are respectively given by (Yu and Zhang, 2005)

$E(u) = \frac{\sigma(1-2\theta)}{\theta(1-\theta)}$ and $Var(u) = \frac{\sigma^2(1-2\theta+2\theta^2)}{\theta^2(1-\theta)^2}$.

Li and Lin (2010) studied lasso, elastic net and group lasso in quantile regressions for continuous response variable by using Bayesian approaches. We will consider similar approaches for binary and tobit quantile regression under a group lasso penalty in Chapters 3 and 4.

1.2.2 Robust and Regularized regression methods

Regularization methods have been recently considered also for robust and quantile regression methods, so that quantile regression can be applied also for high-dimensional data. The first use of regularization in quantile regression is made by Koenker (2004) which include the lasso penalty on the random effects in a mixed-effect quantile regression model to shrink the random effects towards zero. Wang et al. (2007a) considered the least absolute deviation (LAD) estimate with adaptive lasso

penalty and proved its oracle property. Li and Zhu (2008) have developed quantile regression models under a lasso penalty, the theoretical properties of which are derived in (Belloni and Chernozhukov, 2011). Li et al. (2010) provide a Bayesian formulation of the same problem. Wu and Liu (2009) explained the oracle properties of the SCAD and adaptive lasso regularized quantile regression. Finally, Alhamzawi and Yu (2012) developed a Bayesian adaptive lasso regularized quantile regression model. This is thesis is making a contribution to this literature.

1.2 Thesis Outline

The outline of the thesis is as follows. In Chapter 2, we address the problem of variable selection when the response variable is continuous for high-dimensional data. We briefly present a motivation of the regularized robust regression methods for continuous response variables, review several regularization methods and present a comparative simulation study under different error distributions.

In Chapter 3, we address the problem of variable selection when the response variable is binary for high-dimensional data. We propose quantile regression with a group lasso penalty when the response is binary. We develop a Bayesian procedure for parameter estimation. Simulations and real data analysis are conducted to investigate the effectiveness of the proposed model.

In Chapter 4, we address the problem of variable selection when the response variable is censored for high-dimensional data. Quantile regression with a group lasso penalty approach is extended to a tobit model. We present a Bayesian approach for parameter estimation and illustrate the performance of the proposed method using simulation studies and real data analysis. Moreover, we investigate, the calculation of

predicted values for the tobit model when the error term is distributed as a normal and asymmetric Laplace distribution, respectively.

In Chapter 5, we summarise the conclusions drawn as a result of the research work presented. This chapter also discusses some suggestions for future work.

Chapter 2

Regularized Robust Regression Methods for Continuous Response Variables

This chapter considers the estimation of linear regression parameters using regularization methods when the response variable is continuous and the data is highly dimensional. As discussed in chapter 1, regularization is a method for modelling modern data, which is high-dimensional, sometimes noisy and often contains a lot of unimportant predictors (Rosset, 2003). Regularization methods can improve the predictive error of the model by reducing the variability in the estimates of regression coefficients by shrinking the estimates towards zero. For example lasso, elastic net and adaptive lasso, as discussed in the first chapter, shrink some coefficient estimates to exactly zero, thus providing a form of variable selection. The main aim of this chapter is to study and compare different regularized robust regression methods and Bayesian regularized quantile regression methods for continuous response variables under different error distributions in the case of high-dimensional data.

2.1 Classical regularized regression methods

2.1.1 Ridge regression

Ridge regression introduced by Hoerl and Kennard (1970) is one of the most popular alternative solutions to OLS. This method is used to improve the estimation of regression parameters in the case where the predictor variables are highly correlated. The ridge regression parameter estimates are given by minimizing the residual sum of squares subject to an L_2- penalty on the coefficients. The ridge estimate is given by

$$\hat{\beta}_{ridge} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\}, s.t. \quad \sum_{j=1}^{p}\beta_j^2 \leq t\,, t \geq 0\,. \tag{2.1}$$

Or equivalently, the ridge regression is defined by the following minimisation problem:

$$\hat{\beta}_{ridge} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}\ ,\ \ \lambda \geq 0, \tag{2.2}$$

where the positive scalar λ is a regularization parameter that controls the amount of shrinkage and the penalty function is given by the L_2-norm. The parameter t in (2.1) is clearly related to the parameter λ in (2.2). This means that for a specific value λ there exists a value t such that the estimation equations (2.1) and (2.2) lead to the same solution.

Rewriting the criterion (2.2) in matrix form yields,

$$\hat{\beta}_{ridge} = \min_{\beta}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

Hoerl and Kennard (1970) suggested using all of the available variables and obtaining estimates using:

$$\hat{\beta}_{ridge} = (X^T X + \lambda I_p)^{-1} X^T y\,,$$

where I is the $p \times p$ identity matrix. By adding λI_p to $X^T X$, this results in a regular and invertible matrix. The intercept β_0 is usually not included in the penalty. This can be done by first centring the inputs and response variables.

Contrary to the OLS estimates, the ridge estimator is biased. Hence this regularization method accepts a little bias to reduce the variance and the mean squared error, respectively, of the estimates and possibly improve the prediction accuracy. Hoerl and Kennard (1970) introduced a graphical method known as the ridge trace to help the user determine the optimal value of the regularization parameter λ. In

summary, ridge regression yields more stable estimates of the regression coefficients by shrinking the coefficients. In general, no coefficients are shrunk to exactly zero and therefore the procedure does not give an easily interpretable model. Further regularization methods were proposed, for example lasso and elastic net that perform variable selection and estimation simultaneously.

2.1.2 Lasso

A popular method called Least Absolute Shrinkage and Selection Operator (lasso) was proposed by Tibshirani (1996). The lasso is a penalized least squares method which imposes an L_1- penalty on the regression coefficients. The lasso is a regularization method to estimate coefficients and perform variable selection for high dimensional data, where the number of predictor variables p is potentially much larger than the number of samples n .The intercept β_0 is usually not included in the penalty. This can be done by first centring the inputs and response variables, then fitting a model with no intercept. The lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. The lasso estimate $\hat{\beta}$ is defined by

$$\hat{\beta}_{lasso} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\}, \quad s.t. \quad \sum_{j=1}^{p}|\beta_j| \leq t \quad t \geq 0.$$

An equivalent form of the lasso is,

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_j |\beta_j|\right\},$$

or $$\hat{\beta}_{lasso} = \min_{\beta}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Here $t \geq 0$ (or λ) is a regularized parameter that can be chosen by cross-validation or generalized cross-validation (Tibshirani 1996). For every choice of t, there is a choice of λ that gives the same result. Because of the L_1-penalty, the solution of lasso is usually sparse when a high regularization parameter λ is used and the lasso does both shrinkage and variable selection simultaneously. The estimation of lasso is a convex optimization problem and can be solved by a quadratic programming algorithm for a given λ. This can be computationally expensive since it requires solving the optimization problem for a grid of λs. However, an efficient algorithm introduced by Efron et al. (2004), Least Angle Regression (Lars), is available in the $lars$ R package for computing the entire path solution at a small computational cost.

Although the lasso has shown success in many situations, it has some limitations. Zou and Hastie (2005) consider the following three scenarios:

(a)In the case where the number of predictors is larger than the number of observations, the lasso selects at most n variables before it saturates. Lasso cannot do group selection because of the nature of the convex optimization problem.

(b)If there is a group of variables among which the pairwise correlations are very high, then lasso tends to arbitrarily select only one variable from the group. Group selection is important, for example, in gene selection problems.

(c) If there is high correlation between the predictors, it has been observed that the prediction performance of the lasso is determined by ridge regression.

Case (a) and (b) make the lasso unsuitable as a variable selection method in some situations.

2.1.3 Bridge regression

Bridge regression is a method for the estimation of linear models that minimizes the squared sum of errors subject to the L_q norm of the parameter estimates being less than a constant t. The bridge estimate can be obtained by minimizing (Frank and Friedman, 1993)

$$\hat{\beta}_{bridge} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|^q\}, \qquad \lambda \geq 0$$

or $$\hat{\beta}_{bridge} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2\}, s.t. \quad \sum_{j=1}^{p}\beta_j^q \leq t\,,\ t \geq 0\,.$$

Unfortunately there is no closed form solution for problems of this type. Since bridge regression penalties contains subset selection ($q = 0$), lasso ($q = 1$), and ridge regression ($q = 2$) as special cases, it gives us opportunities to choose between subset regression and ridge regression. For how to estimate the amount q and the regularization parameter λ via generalized cross-validation from the data when $0 \leq q \leq 2$, see Frank & Friedman (1993) and Fu (1998).

2.1.4 Elastic net

A regularization and variable selection method which is used to improve selection when groups of predictors are highly correlated is the elastic net, presented by Zou and Hastie (2005). The elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. The elastic net criterion is defined by

$$\hat{\beta}_{elastic\ net} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j\, x_{ij})^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2\ \}, \qquad (2.3)$$

which depends on two regularized parameters $\lambda_1, \lambda_2 > 0$.

The elastic net penalty is a convex combination of the lasso and ridge penalty and, in constraint form, it is given by $(1-\alpha)\sum_{j=1}^{p}|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2 \leq t$ with $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$. For $\alpha = 1$ we obtain simple ridge regression, whereas for $\alpha = 0$ we obtain the lasso penalty. Equation (2.3) is called the naive elastic net, because it is similar to either ridge regression or the lasso and tends to over shrink in regression problems. Zou and Hastie (2005) propose the elastic net as a useful method in the analysis of microarray data, where the selection of highly correlated groups of predictors is preferred because these groups are biologically interesting.

2.1.5 Adaptive lasso

Zou (2006) proposed a new version of lasso, which is called adaptive lasso. The penalized least squares with adaptive lasso is defined as

$$\hat{\beta}_{adaptive\ lasso} = \min_{\beta}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\widehat{w}_j\,|\beta_j|\}\,.$$

Instead of simply using the absolute value of the parameters as the penalization, adaptive weights are added for penalizing different coefficients differently. Zou (2006) suggested the use of estimated weights, $\widehat{w}_j = \frac{1}{|\hat{\beta}_j|^{\gamma}}$, where $\hat{\beta}_j$ comes from minimizing the OLS or lasso and γ is a user-chosen constant. The choice of $\widehat{w}_j$ is very important and Zou (2006) suggested using OLS while γ can be chosen by K -fold cross-validation. The adaptive lasso selects the true set of nonzero coefficients with probability tending to one.

2.1.6 SCAD

The SCAD (Smoothly Clipped Absolute Deviation) penalty was proposed by Fan and Li (2001). The SCAD penalty is best defined in terms of its firs derivative,

$$p'_{\lambda}(\beta) = \lambda\left\{I\{\beta \leq \lambda\} + \frac{(a\lambda-\beta)+}{(a-1)\lambda} I\{\beta > \lambda\}\right\} \text{ for some } a > 2 \text{ and } \beta > 0,$$

where I is the indicator function, β is vector of unknown parameters and λ is regularized parameter. An important improvement of SCAD over lasso is that large values of β are penalized less than small values of β. Also, unlike traditional variable selection procedures, the SCAD estimator's sampling properties can be established precisely. For example, Fan and Li (2001) demonstrated that as n increases, the SCAD procedure selects the true set of nonzero coefficients with probability tending to one. Fan and Li (2001) also show that the SCAD penalty can be effectively implemented in robust linear and generalized linear models.

2.1.7 Group lasso

As we explained the properties of the lasso penalty, this penalty has the advantage of providing simultaneous parameter estimation and variable selection (Tibshirani, 1996). The original lasso method was extended in a number of directions, amongst which adaptive lasso (Zou, 2006; Alhamzawi et al., 2012) and Cox regularized regression (Tibshirani, 1997). In some cases, the predictors have a natural group structure, such as in the case of a categorical variable being converted into dummy variables. In these cases, the selection of groups of variables is of interest, rather than of individual variables. In order to address this type of problems, Yuan and Lin (2006) developed the group lasso method and a number of authors have subsequently extended it and studied its theoretical properties (Bach, 2008; Huang and Zhang, 2010; Wei and Huang, 2010; Lounici et al., 2011; Sharma et al., 2013; Simon et al., 2013). As we discussed, the elastic net method is suitable when groups of predictors are highly correlated. The group lasso regularized regression (Yuan and Lin, 2006) also handles the predictors when they are grouped together. The group structure of elastic net is unknown when

compared with group lasso where the group structure is completely known in advance. The group lasso regularized regression (Yuan and Lin, 2006) is defined as

$$\hat{\beta}_{group\ lasso} = \min_{\beta} \sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda \sum_{g=1}^{G} \|\beta_g\|_{H_g} ,$$

where: $\beta = (\beta_1^T, \ldots, \beta_G^T)^T$, $x_i = (x_{i1}^T, \ldots, x_{iG}^T)^T$,

β_g: the vector of coefficients of the gth group predictors x_{ig}

$\|\beta_g\|_{H_g} = (\beta_g^T H_g \beta_g)^{\frac{1}{2}}$ $g = 1, \ldots, G.$

$H_g = d_g I_{d_g}$ and d_g the dimension of the vector β_g

G : Number of groups.

As this method will be the focus of this thesis, further details will be given in Chapter 3 and Chapter 4.

2.2 Robust regularized regression methods

When the regression response suffers from outliers, the performance of lasso can be poor. A first attempt to solve this problem has been done by Rosset and Zhu (2007) and Wang et al. (2007a). Rosset and Zhu (2007) combine the idea of Huber's criterion as loss function and lasso penalty. They fix the penalty to be the L_1- penalty and use Huber's loss function with fixed M. That is

$$\hat{\beta}_{Huber\ lasso} = \min_{\beta} \sum_{i=1}^{n} \rho(y_i - x_i^T\beta) + \lambda \sum_{j=1}^{p} |\beta_j| , \qquad (2.4)$$

where $\rho(t) = \begin{cases} t^2 & if \quad |t| \le M \\ 2M|t| - M^2 & if \quad |t| > M \end{cases}$.

The LAD-adaptive lasso method is developed by Wang et al. (2007a) by combining the idea of Least Absolute Deviance (LAD) and adaptive lasso for robust regression shrinkage and selection. The LAD- adaptive lasso can be written as

$$\hat{\beta}_{ladl} = \min_{\beta} \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right| + \lambda \sum_{j=1}^{p} \widehat{w}_j^{ladl} \left| \beta_j \right| ,$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{ladl}, \ldots, \widehat{w}_p^{ladl} \right)$ is a known weights vector. In this model the estimator is robust to outliers because the squared loss has been replaced by the l_1-loss. Lambert-Lacroix and Zwald (2011) proposed the Huber's Criterion with adaptive lasso which combines the idea of Huber's criterion as loss function and adaptive lasso penalty, defined by

$$\hat{\beta}_{Hadl} = \min_{\beta} \mathcal{L}_{\rho}(\beta, s) + \lambda \sum_{j=1}^{p} \widehat{w}_j^{Hadl} \left| \beta_j \right|$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{Hladl}, \ldots, \widehat{w}_p^{Hladl} \right)$ is a known weights vector and the Huber's criterion is defined by

$$\mathcal{L}_{\rho}(\beta, s) = \begin{cases} ns + \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} \beta_j x_{ij}}{s} \right) s & if \quad s > 0, \\ 2M \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right| & if \quad s = 0, \\ +\infty & if \quad s < 0, \end{cases}$$

where $\rho(t)$ is defined as (2.4), $s > 0$ is a scale parameter for the distribution. The $\rho(t)$ definition shows how the loss is quadratic for small residuals but it becomes linear for large residuals, thus penalizing outliers. Also this method has been used for regression problems in a number of applications and has shown robustness against outliers. The constant M depends on the level of noise and outliers in the data and is often set to the value $M = 1.345$.

Bradic and Fan (2011) proposed a new method for robust linear models, which replaces the quadratic loss by a weighted linear combination of convex loss functions, so

$\hat{\beta}_{BF} = \min_{\beta}\{\sum_{i=1}^{n} \rho_w(y_i - x_i^T\beta) + n\sum_{j=1}^{p} p_\lambda\left|\beta_j\right|\}$,

where : $\rho_w = \sum_{k=1}^{K} w_k\rho_k,\ \ with\ \ \rho_1\ ,\ldots,\rho_K$ convex loss functions and $w_1\ ,\ldots,w_K$ positive constants.

p_λ : is a specific penalty function.

Arslan (2012) proposed and investigated the properties of the weighted LAD-lasso method which combines the idea of the weighted least absolute deviation (WLAD) regression and the adaptive lasso for robust parameter estimation and variable selection in regression. The WLAD-lasso regression estimator can be obtained by minimizing the following objective function

$$\hat{\beta}_{wlad} = \min_{\beta} \sum_{i=1}^{n} w_i\left|y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right| + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right|,$$

where w_i are the weights computed from the robust distances of the predictors x_i $RD(x_i) = (x_i - \hat{\mu})^T\hat{\Sigma}^{-1}(x_i - \hat{\mu})$, for $i = 1,\ \ldots\ ,n$, $\lambda_j,\ j = 1,\ \ldots\ ,p$ are the regularized parameters in the adaptive lasso objective function and will be estimated from the data, $\hat{\mu}$ and $\hat{\Sigma}^{-1}$ are robust location and scatter estimators respectively (Hubert and Rousseeuw, 1997).

Recently Li and Lin (2010) studied lasso, elastic net and group lasso in quantile regressions for continuous response variable by using Bayesian approaches. The lasso and elastic net regularized quantile regression for $0 < \theta < 1$ is given by, respectively (Li and Lin, 2010)

$\hat{\beta}_{lq} = \min_{\beta} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T\beta) + \lambda\sum_{j=1}^{p}\left|\beta_j\right|$ and

$$\hat{\beta}_{enq} = \min_{\beta} \sum_{i=1}^{n} \rho_{\theta}(y_i - x_i^T \beta) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 ,$$

where ρ_θ is the check function as defined in (1.8). Li and Lin (2010) derived Gibbs samplers for both methods by using Bayesian approaches. In the next section we will give a comparative simulation study for some of the existing methods.

2.3 Comparison of robust and regularized regression methods on simulated data

In this section, we compare regularized regression methods in low- dimensional ($p = 15, n = 100$) and high-dimensional ($p = 100$, $n = 50$) settings. For both settings we use a classical simulation setting, e.g. Bradic and Fan (2011), where $y = \beta_0 + x\beta + u$, with $\beta_0 = 0$ and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$. We draw the independent variables x from a multivariate normal distribution, $N(0, \Sigma_x)$. The pairwise covariance between x_i and x_j is set to be $(\Sigma_x)_{ij} = r^{|i-j|}$. For the error u, we choose a range of distributions in order to test the robustness of the methods to departures from normality. In particular, we consider the following cases: $u \sim N(0, 1)$, Double Exponential (DE), t-distribution with 1 (t_1) and 3 (t_3) degrees of freedom, Gamma$(3, 1)$ and mixture normal distributions. We design a mixture normal distribution with large outliers, similar to Lambert-Lacroix and Zwald (2011), by drawing 90% of the data from $a\ N(0, 1)$ distribution and 10% from a $N(0, 1000)$ distribution. Under all these cases, we compare the regularized regression methods described in the previous section, namely lasso (Tibshirani, 1996), LAD (Li and Zhu, 2008) and Huber lasso (Rosset and Zhu, 2007), with their adaptive versions (Xu and Ying, 2010; Lambert-Lacroix and Zwald, 2011). For lasso we use the R package $lars$, for elastic net we use the R package $elasticnet$, for LAD and the Huber lasso we use the R implementations

provided by Li and Zhu (2008) and Rosset and Zhu (2007), respectively, for the adaptive lasso we adapt some of the functions in the $parcor$ R package and we code in a similar way the adaptive LAD and adaptive Huber lasso methods. For the adaptive versions of the methods, we define the weights using the corresponding non-adaptive lasso versions with a penalty parameter chosen to optimize a BIC criterion. As for the main penalty parameter, we fix this to the parameter that selects exactly three non-zero coefficients, for each of the six methods. In this way, all methods can be compared at the same level of sparseness and the true positives can be directly compared.

2.3.1 Example 1: low- dimensional

In this section we consider a low-dimensional data set with $p = 15$ and $n = 100$. Figure 2.1 reports the results of the simulation. We consider both the case of low correlation ($r = 0.5$) and that of high correlation ($r = 0.95$) of the predictors. The top panels report the median model error over 500 iterations (similar results for the mean error), with the model error computed by$(\hat{\beta} - \beta)^T S_x(\hat{\beta} - \beta)$, where $\hat{\beta}$ are the estimated parameters and S_x the sample covariance. The bottom panels report the true positives, that is the number of correctly found non-zero coefficients. Here three corresponds to the case of all non-zero coefficients being correctly detected.

Our results show that: lasso does not perform well when the predictors are highly correlated; the adaptive methods tend to outperform their non-adaptive versions; the adaptive LAD method outperforms all others methods for all error distributions. This results confirmed by (Lambert-Lacroix and Zwald, 2011).

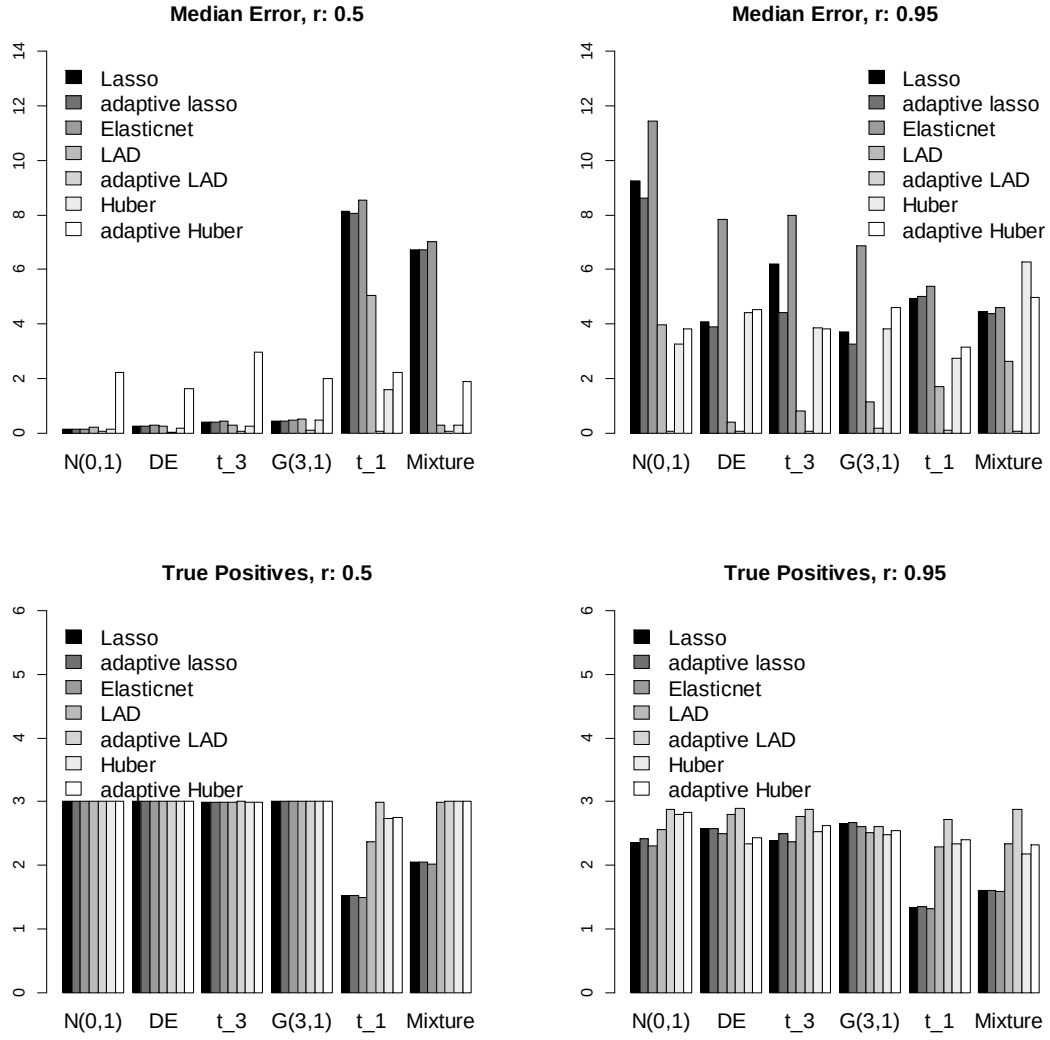

Figure 2.1: Comparison of regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for example 1 and the bottom panels the average true positives when $p = 15$ *and* $n = 100$ *.*

2.3.2 Example 2: high- dimensional

We consider a similar setting to simulation 2.3.1 but with different sample size and number of predictors. In particular, we consider a high- dimensional example with $p = 100$ and $n = 50$. Given the setup of the simulation, this a very sparse problem in which most of the coefficients are zero. Figure 2.2 presents the results of the simulation. The top panels report the median model error over 500 replications, with

the model error computed in the same way as in Figure 2.1. The bottom panels report the true positive that is the number of correctly classified non-zero coefficients.

The results support existing knowledge about the performance of the methods: lasso does not perform well when the predictors are highly correlated, the adaptive methods tend to outperform their non-adaptive versions, particularly for the adaptive LAD lasso method, and the robust methods generally outperform the non-robust ones as departures from normality increase. This is particularly evident for the cases of the mixture model and t_1 simulation, which have a severe departure from normality.

For the results in Figure 2.1 and Figure 2.2, we fixed the value of the penalty parameter λ such that exactly three non-zero coefficients are selected. The choice of the penalty parameter is in general the crucial question when applying regularized methods, particularly in a high-dimensional setting. This is not the main focus of this chapter, as long as a consistent approach is chosen for all the models compared. However, in the context of non-normal data, there is also a question about the possible sensitivity of the penalty parameter to outliers and departures from normality.

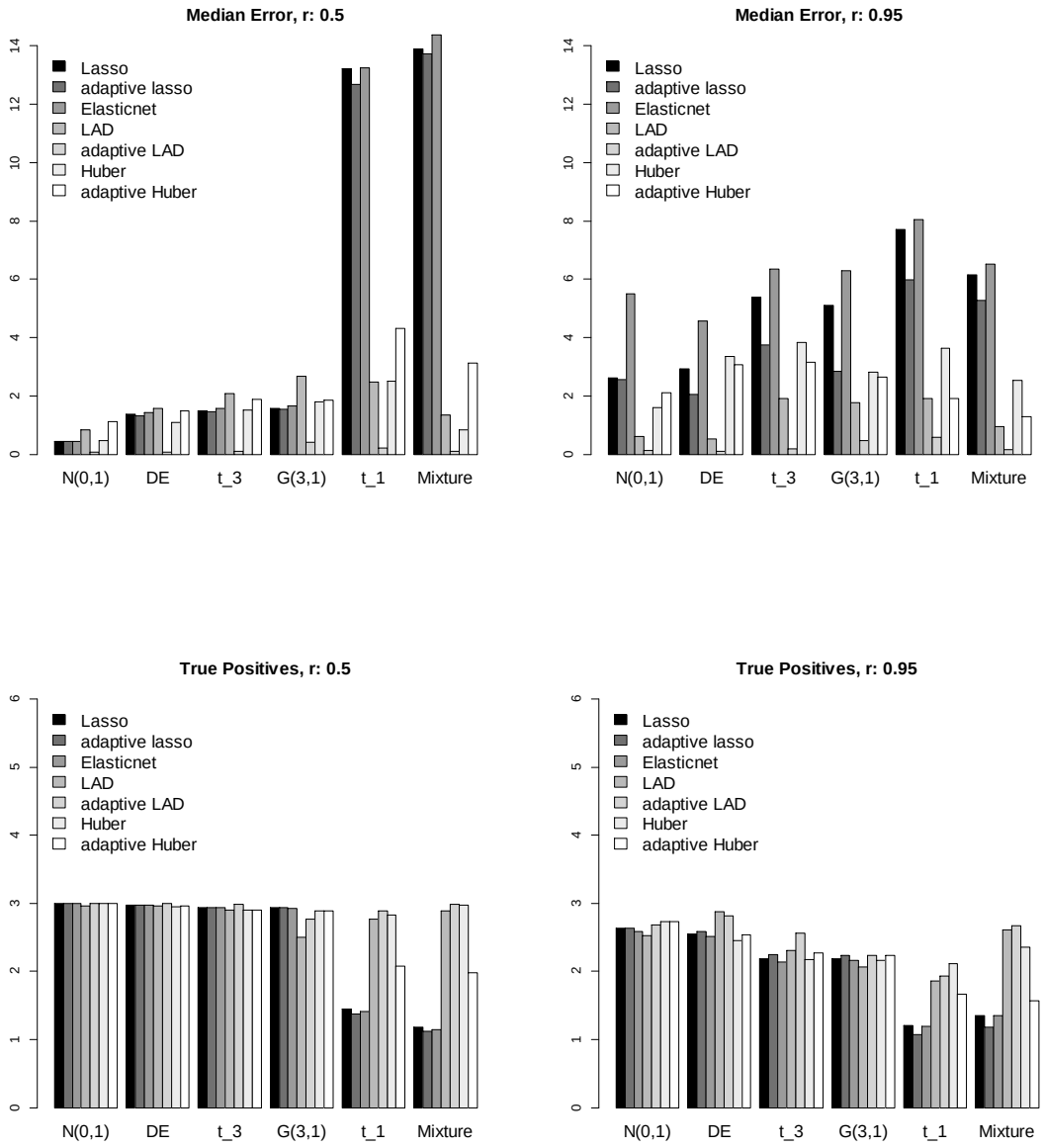

Figure 2.2: Comparison of regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 500 replications for example 2 and the bottom panels the average true positives when $p = 100$ *and* $n = 50$.

2.4 Comparison of Bayesian regularized quantile regression methods with classical methods on simulated data

In this section, four examples are considered. In each example, we use a classical simulation setting, as in section 2.3. We compare the Bayesian lasso quantile regression and Bayesian elastic net quantile regression (Li et al., 2010) with LAD (Li and Zhu, 2008) and elastic net, respectively. For each Bayesian case, we use the R-code provided by Li et al. (2010) and we run a Gibbs sampling procedure, using 13000 iterations with the first 3000 iterations as burn-in.

2.4.1 Example 3: low- dimensional with sparse coefficients

The data for example 3 is the same as example 1, that is we consider a low-dimensional data set with $p = 15$ and $n = 100$. Figure 2.3 reports the median model error over 40 iterations for both the case of low correlation ($r = 0.5$) and that of high correlation ($r = 0.95$) of the predictors.

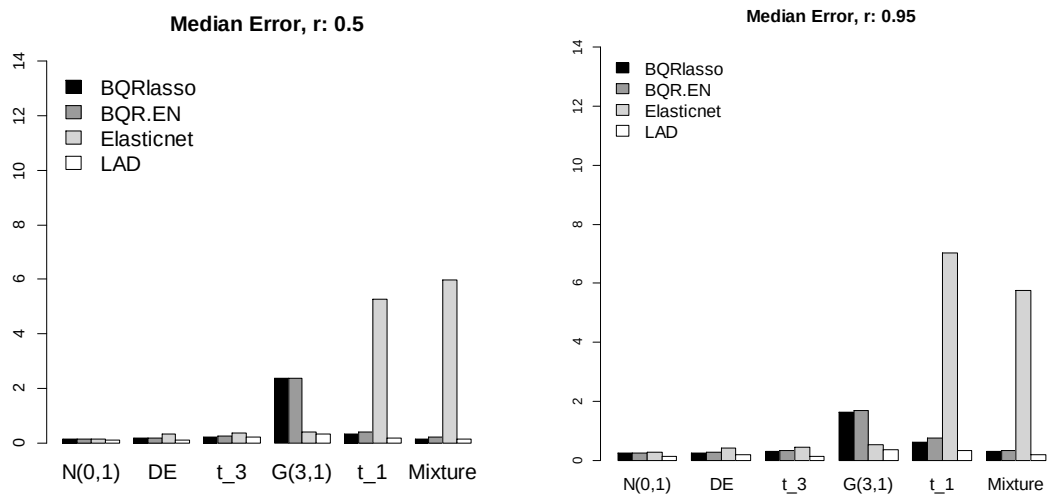

Figure 2.3: Comparison of Bayesian quantile regression methods with frequentist methods, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 3 when $p = 15$ and $n = 100$.

From Figure 2.3, we observe that the performance of the elastic net model is not satisfactory as its median model errors increase as the departure from normality increases. The LAD approach tends to perform similarly to Bayesian quantile regression methods.

2.4.2 Example 4: high – dimensional with sparse coefficients

The data for example 4 is the same as example 2, where we consider a high-dimensional data set. Figure 2.4 reports the median model error over 40 iterations for the case $p = 100$ and $n = 50$.

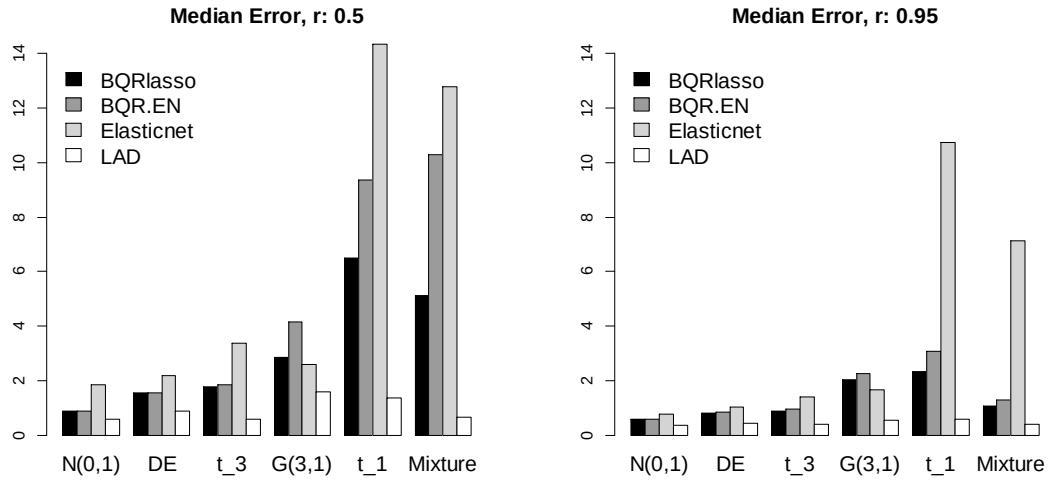

Figure 2.4:Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 4 when $p = 100$ *and* $n = 50$.

From Figure 2.4, we observe an overall good performance of the LAD estimator: the median model errors are small even when the departure from normality increases. The performances of the two Bayesian methods are similar and generally inferior to LAD. Furthermore, the results show how the elastic net is the worst performing method especially in the case of mixture normal and t_1 error distributions.

2.4.3 Example 5: simulation with non-sparse coefficients

In order to investigate the poor performance of Bayesian methods in example 4, we set up a new simulation where we have $\beta_j = 0.1$ for all j, that is a non sparse situation. Since the Bayesian methods do not give exact zero coefficients, we expect Bayesian methods to perform well in this case. Figure 2.5 reports the median model error over 40 replications for the case $p = 50$ and $n = 100$ and Figure 2.5 for the case $p = 100$ and $n = 50$.

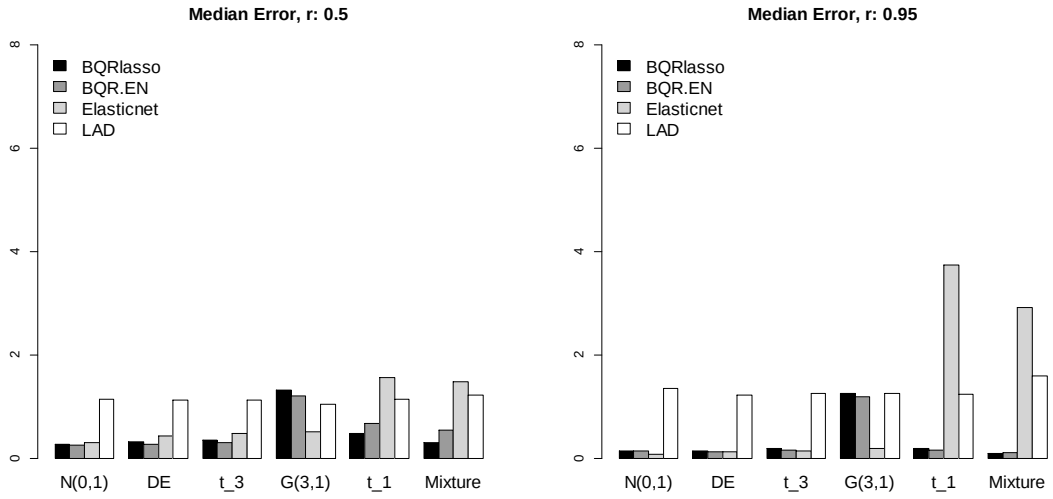

Figure 2.5:Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 5 when $p = 50$ and $n = 100$.

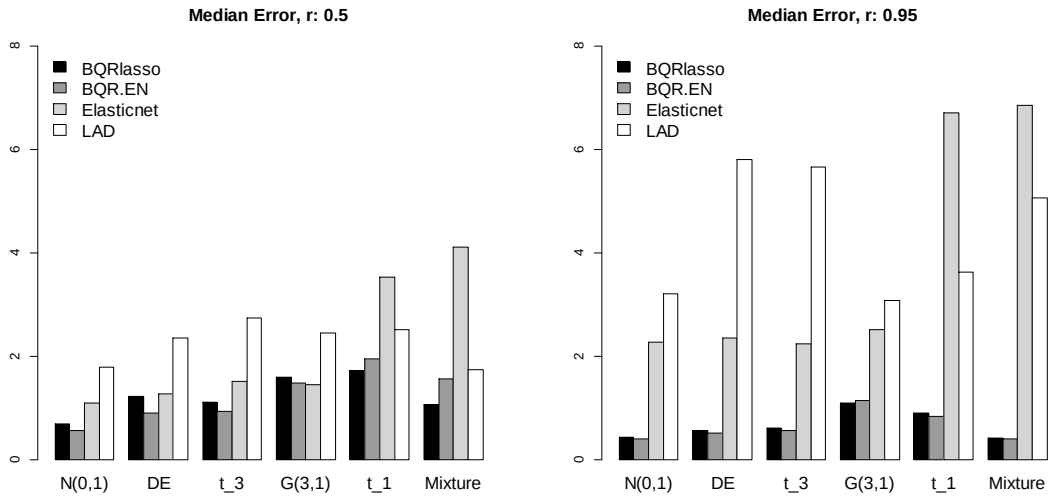

Figure 2.6*: Comparison of Bayesian regression lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for example 5 when* $p = 100$ *and* $n = 50$.

From results in Figure 2.5 and 2.6, our simulation study confirms that the performances of the two Bayesian methods are similar, and the Bayesian methods perform generally better than the non-Bayesian methods in the case non sparse coefficients. Furthermore, the results show how elastic net is the worst performing method in case of departure from normality especially when the predictors highly correlated.

2.5 Concluding remarks

Many approaches are developed in statistics that rely on the assumption of normality. These approaches are not suited to data that show clear departures from normality. This is often the case when data are contaminated, resulting in the presence of outliers. In this chapter, we have considered recently developed Bayesian quantile regression methods and robust methods, such as the Huber or LAD methods. In a high dimensional setting, when $p \geq n$, a regularized penalty on the regression coefficients is also considered. In a simulation study, we show how robust methods are superior to

the non-robust counterparts, particularly for cases where there is a large departure from normality. Adaptive versions of robust and traditional regression methods have been developed by carefully setting a weight on the β coefficients and these have shown a very good performance. The performances of the Bayesian lasso quantile regression and Bayesian elastic net quantile regression are similar, and the Bayesian methods perform generally better than the non-Bayesian methods in the case of non-sparse coefficients,but worse in the sparsity setting. This may be because Bayesian regularized methods do not return exact zero coefficients. In order to overcome this limitation, Alhamzawi and Yu (2013) proposed a variable selection method for Bayesian regularized quantile regression. In general the Bayesian approaches are more time-consuming than non-Bayesian methods but enable us to make use of all available information from data and get the distribution of the parameter estimates. In chapters 3 and 4, we will use a Bayesian estimation framework for modelling binary and censored response data respectively.

Chapter 3

Binary Quantile Regression with Group Lasso

In chapter 2 we considered the application of different regularized robust regression methods for the case when the response variable is continuous. In this chapter we consider regularized and robust regression models when the response variable is binary. Applications of regression models for binary response are very common and models specific to these problems are widely used. Quantile regression for binary response data has recently attracted attention and regularized quantile regression methods have been proposed for high dimensional problems. If the predictors have a natural group structure, a group lasso penalty has been found to be useful in regularized methods. In this chapter, we present a Bayesian Gibbs sampling procedure to estimate the parameters of a binary quantile regression model under a group lasso penalty.

3.1 Introduction

As we discussed in chapters 1and 2, quantile regression is very useful when the data do not satisfy the normal distributional assumptions underlying traditional methods or when the data are subject to some form of contamination. One line of research has extended the original quantile regression model to the case where the response is binary, as an alternative to traditional mean-based models, such as logistic and probit regression models. The methods were originally developed in the frequentist estimation setting by Manski(1975, 1985) and were subsequently extended also to the Bayesian counterpart (Yu and Moyeed, 2001; Benoit and Poel, 2012; Miguéis et al., 2012) as a

means to avoid large-sample based asymptotic results for inference and at the same time take regression parameter uncertainty into account.

Given the merits of the regularized methods described in chapter 2, regularized methods for binary response variables have also been developed. In particular, (Bae and Mallick, 2004; Genkin and Lewis, 2007; Gramacy and Polson, 2012) developed Bayesian logistic regression models under a lasso or ridge penalty, (Meier et al. (2008) developed classical logistic regression model under a group lasso penalty, and Krishnapuram et al. (2005) developed a sparse multinomial logistic regression model.

The references above refer to the estimation of mean-based regression models. A small line of research has explored the link between the robust quantile regression models and the regularized models for high-dimensional data (see chapters 1 and 2, for more details). In particular, Ji et al. (2012) have developed a quantile regression model under an L_1 penalty and for a binary response. In this chapter, we extend the work of Ji et al. (2012) on binary quantile regression models with the use of a group lasso penalty. Our model is derived in the framework of probit binary regression and offers an alternative to the mean-based logistic regression model with group lasso penalty (Meier et al., 2008), when the response is binary, the predictors have a natural group structure and quantile estimation is of interest. In section 3.2 we describe the model; in section 3.3 we describe the estimation of the parameters in a Bayesian setting; in section 3.4 we discuss how the model is used for prediction, in sections 3.5 and 3.6, we compare the method with an existing mean-based logistic regression model under group lasso penalty and binary quantile regression with lasso penalty on simulated and real data. Finally, in section 3.7, we draw some conclusions.

3.2 Binary quantile group lasso

Similar to a probit regression model, binary quantile regression models can be viewed as linear quantile regression models with a latent continuous response variable, e.g. (Ji et al., 2012). In particular, let y be the binary response variable, taking values 0 and 1, let x be the vector of p predictors, β the vector of unknown regression coefficients and $(x_i, y_i), i = 1, \ldots, n$ a sample of n observations on x and y. Given a quantile θ, $0 < \theta < 1$, we consider the model:

$$y_i^\star = x_i^T \beta_\theta + u_i \,, i = 1, \ldots, n \text{ and } y_i = h(y_i^\star),$$

where u_i are the errors, satisfying $P(u_i \leq 0 | x_i) = \theta$ and h is a link function. For binary response data, the link function is given by $h(y^\star) = I(y^\star > 0)$, with I the indicator function. In real applications, y is the observed binary response and the interest is to predict y from knowledge of x. $y^\star$ is unobserved and used mainly for modelling purposes. Some examples of $y_i^\star$ include the actual birth weight of babies in a study where the aim is to investigate the factors behind the birth of premature babies, the credit risk of a customer in a study where the aim is to discriminate between good and bad customers (Kordas, 2002) or the willingness to participate to work in a study where the factors behind the decision to work or not are investigated (Kordas, 2006).

The attractive property of this latent model is that there is a correspondence between the quantiles of y and the quantiles of $y_i^\star$, which are directly modelled. In particular, using the equivariance properties of quantile functions (Kordas, 2006), it holds that

$$Q_{y|x}(\theta) = Q_{h(y^\star|x)}(\theta) = h(Q_{y^\star|x}(\theta)) \,,$$

with $Q_{y|x}(\theta)$ denoting the θ conditional quantile of Y given x. From this, since $Q_{y_i^\star|x}(\theta) = x^T\beta_\theta$ under a linear quantile regression model, it follows that

$$Q_{y|x}(\theta) = h(x^T\beta_\theta) = I(x^T\beta_\theta > 0).$$

So the estimation of the parameters β_θ leads to the knowledge about the θ quantile of y. In the next section, we describe how to estimate β_θ under a group lasso penalty.

3.3 Bayesian parameter estimation

In a binary quantile regression model, the parameter β_θ is found by the following minimization problem (Manski, 1985):

$$\min_{\|\beta\|=1} \sum_{i=1}^{n} \rho_\theta(y_i - h(x_i^T\beta)), \tag{3.1}$$

where ρ_θ is the check function defined by

$$\rho_\theta(t) = \begin{cases} \theta t & if \quad t \geq 0, \\ -(1-\theta)t & if \quad t < 0 \end{cases}$$

or equivalently $\rho_\theta(t) = \frac{|t|+(2\theta-1)t}{2}$. The restriction on $\|\beta\| = 1$is motivated by the fact that the scale of the parameter is not identifiable, being $y_i^\star$ a latent variable.

Yu and Moyeed (2001) have shown how minimizing (3.1) is equivalent to maximising the likelihood function, under the assumption that the error comes from an asymmetric Laplace distribution with density given by $f_\theta(\mu) = \theta(1-\theta)\exp(-\rho_\theta(u))$. That is, minimising (3.1) is equivalent to maximising the likelihood

$$f(y|x,\beta,\theta) = \theta^n(1-\theta)^n exp(-\sum_{i=1}^{n} p_\theta(y_i - h(x_i^T\beta))). \tag{3.2}$$

This fact has created a straightforward working model for Bayesian inference quantile regression.

When the predictors have a natural groupe structure, the methodology above can be extended to the use of a group lasso penalty. In particular, suppose that the predictors are grouped into G groups and β_g is the vector of coefficients of the g^{th} group of predictors, which we denote with x_{ig} for the observation i. Let $\beta = (\beta_1^T, \ldots, \beta_G^T)^T$ and $x_i = (x_{i1}^T, \ldots, x_{iG}^T)^T$, $i = 1, \ldots, n$. Under a group lasso constraint, the minimization in (3.1) becomes

$$\min_{\|\beta\|=1} \sum_{i=1}^{n} \rho_\theta (y_i - h(x_i^T \beta)) + \lambda \sum_{g=1}^{G} \|\beta_g\|_{H_g}, \tag{3.3}$$

where λ is a non-negative regularization parameter, controlling the sparsity of the solution, and $\|\beta_g\|_{H_g} = (\beta_g^T H_g \beta_g)^{\frac{1}{2}}$ with $H_g = d_g I_{d_g}$ and d_g the dimension of the vector β_g. The choice of d_g in H_g has been suggested by Yuan and Lin (2006) to ensure that the penalty term is of the order of the variables in the group. Under an appropriate choice of prior distribution, the minimization problem in (3.3) can be shown to be equivalent to a maximum a posteriori solution. In particular, a Laplace prior on β_g is chosen, that is

$$\pi(\beta_g|\lambda) = C_{d_g} \sqrt{\det(H_g)}\, \lambda^{d_g} \exp(-\lambda \|\beta_g\|_{H_g}), \tag{3.4}$$

where $C_{d_g} = 2^{-\frac{(d_g+1)}{2}} (2\pi)^{-\frac{(d_g-1)}{2}} / \Gamma((d_g+1)/2)$ and Γ is the gamma function. Then, using the same asymmetric Laplace distribution for the residuals u, the minimization in (3.3) is equivalent to the maximum of the posterior distribution

$$f(\beta|y,x,\lambda,\theta) \propto exp\left(-\sum_{i=1}^{n}\rho_\theta\,(y_i\ - h(x_i^T\beta)) - \lambda\sum_{g=1}^{G}\|\beta_g\|_{H_g}\right), \tag{3.5}$$

under the constraint that $\|\beta\| = 1$.

3.3.1. Gibbs sampling procedure

We extend the Gibbs sampling procedure of Ji et al. (2012) to the case of a group lasso penalty. As a first step we rewrite the prior of β_g using the equality (Andrews and Mallows, 1974)

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2}exp\left(-\frac{a}{2}s\right)ds,$$

which holds for any $a \geq 0$. In particular, we take $a = \lambda$ and $z = \|\beta_g\|_{H_g} = (\beta_g^T H_g \beta_g)^{\frac{1}{2}}$. Then the prior in (3.4) can be rewritten as

$$\pi(\beta_g|\lambda) = C_{d_g}\sqrt{\det(H_g)}\ \lambda^{d_g}\exp\left(-\lambda\|\beta_g\|_{H_g}\right)$$

$$= C_{d_g}\sqrt{\det(H_g)}\ 2\lambda^{d_g-1}(\tfrac{\lambda}{2})\exp\left(-\lambda\|\beta_g\|_{H_g}\right)$$

$$= 2C_{d_g}\lambda^{d_g-1}\sqrt{\det(H_g)}\ \int_0^\infty \frac{1}{\sqrt{2\pi s_g}}exp\left\{-\frac{1}{2}\beta_g^T(s_g H_g^{-1})^{-1}\beta_g\right\}(\tfrac{\lambda^2}{2})\exp\left(\frac{-\lambda^2}{2}s_g\right)ds_g$$

$$= 2\,2^{-\frac{d_g+1}{2}}(2\pi)^{-\frac{d_g-1}{2}}/\Gamma((d_g+1)/2)\lambda^{d_g-1}\sqrt{\det(H_g)}\ \times$$

$$\int_0^\infty \frac{1}{\sqrt{2\pi s_g}}exp\left\{-\frac{1}{2}\beta_g^T(s_g H_g^{-1})^{-1}\beta_g\right\}(\tfrac{\lambda^2}{2})\exp\left(\frac{-\lambda^2}{2}s_g\right)ds_g$$

$$=2^{-\frac{d_g+1}{2}}(2\pi)^{-\frac{d_g-1}{2}}/\Gamma((d_g+1)/2)\lambda^{d_g+1}\sqrt{\det(H_g)}\ \times$$

$$\int_0^{\infty} \frac{1}{\sqrt{2\pi s_g}} exp\left\{-\frac{1}{2}\beta_g^T (s_g H_g^{-1})^{-1}\beta_g\right\} \exp\left(\frac{-\lambda^2}{2} s_g\right) ds_g$$

$$= \left(\frac{\lambda^2}{2}\right)^{\frac{d_g+1}{2}} (2\pi)^{-\frac{d_g-1}{2}} / \Gamma((d_g+1)/2) \sqrt{\det(H_g)} \times$$

$$\int_0^{\infty} \frac{1}{\sqrt{2\pi s_g}} exp\left\{-\frac{1}{2}\beta_g^T (s_g H_g^{-1})^{-1}\beta_g\right\} \exp\left(\frac{-\lambda^2}{2} s_g\right) ds_g$$

$$= \left(\frac{\lambda^2}{2}\right)^{\frac{d_g+1}{2}} (2\pi)^{-\frac{d_g}{2}} / \Gamma((d_g+1)/2) \sqrt{\det(H_g)} \times$$

$$\int_0^{\infty} (s_g)^{\frac{d_g-1}{2}} (s_g)^{-\frac{d_g}{2}} exp\left\{-\frac{1}{2}\beta_g^T (s_g H_g^{-1})^{-1}\beta_g\right\} \exp\left(\frac{-\lambda^2}{2} s_g\right) ds_g .$$

By using the properties of the determinant, for an $k \times k$ matrix, $\det(A^{-1}) = \frac{1}{\det(A)}$ and $\det(cA) = c^k \det(A)$,

$$\pi(\beta_g|\lambda) = \frac{(\frac{\lambda^2}{2})^{(d_g+1)/2}}{\Gamma(\frac{d_g+1}{2})} \int_0^{\infty} \frac{exp\left\{-\frac{1}{2}\beta_g^T (s_g H_g^{-1})^{-1}\beta_g\right\} s_g^{(d_g-1)/2} \exp\left(\frac{-\lambda^2}{2} s_g\right) ds_g}{\sqrt{\det(2\pi s_g H_g^{-1})}} . \quad (3.6)$$

As a second step, we use the fact that an asymmetric Laplace distributed random variable can be written as a mixture of a $N(0,1)$ distributed random variable and an exponentially distributed random variable with rate parameter $\theta(1-\theta)$ (Alhamzawi and Yu, 2013; Kozumi and Kobayashi, 2011; Lum and Gelfand, 2012). This allows rewriting the likelihood (3.2) as:

$$f(y|x,\beta,\theta) \propto exp(-\sum_{i=1}^n \rho_\theta(u_i)) = exp\left(-\sum_{i=1}^n \frac{|u_i|+(2\theta-1)u_i}{2}\right)$$

$$= \prod_{i=1}^n \int_0^{\infty} \frac{1}{\sqrt{4\pi v_i}} exp\left(-\frac{(u_i-\xi v_i)^2}{4v_i} - \zeta v_i\right) dv_i$$

with $u_i = y_i - h(x_i^T\beta)$, $\xi = (1-2\theta)$ and $\zeta = \theta(1-\theta)$.

From this, we can derive the following conditional distributions:

$$f(\beta|y,x,\lambda,\theta) \propto f(y|\beta,x,\lambda,\theta)\,\pi(\beta|\lambda)$$

$$f(\beta|y,x,\lambda,\theta) \propto \prod_{i=1}^{n}\int_0^{\infty}\frac{1}{\sqrt{4\pi v_i}}exp\left\{-\frac{(u_i-\xi v_i)^2}{4v_i}-\zeta v_i\right\}dv_i \times$$

$$\prod_{g=1}^{G} C_{d_g}\sqrt{\det(H_g)}\;\lambda^{d_g}\exp(-\lambda\left\|\beta_g\right\|_{H_g})$$

$$f(\beta|y,x,v,\lambda,\theta) \propto exp\left(-\sum_{i=1}^{n}\frac{(u_i-\xi v_i)^2}{4v_i}-\lambda\sum_{g=1}^{G}\left\|\beta_g\right\|_{H_g}\right)$$

$$f(\beta|y,x,v,\lambda,\theta) \propto exp\left\{-\sum_{i=1}^{n}\frac{(y_i-h(x_i^T\beta)-\xi v_i)^2}{4v_i}-\lambda\sum_{g=1}^{G}\left\|\beta_g\right\|_{H_g}\right\}.$$

The full conditional distribution of β_g given y,x,v,β_{-g},λ and θ is:

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto f(y|\beta,x,v,\lambda,\theta)\pi(\beta_g|\lambda)$$

$$\propto exp\left\{-\sum_{i=1}^{n}\frac{(y_i-h(\sum_{k=1}^{G}x_{ik}^T\beta_k)-\xi v_i)^2}{4v_i}-\lambda\left\|\beta_g\right\|_{H_g}\right\}.$$

If we write $\tilde{y}_{ig}=y_i-h(\sum_{k=1,k\neq g}^{G}x_{ik}^T\beta_k)-\xi v_i$, then using (3.6), we can write the conditional distribution of β_g as

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto exp\left\{-\sum_{i=1}^{n}\frac{(\tilde{y}_{ig}-h(x_{ig}^T\beta_g))^2}{4v_i}\right\}exp\left\{-\frac{1}{2}\beta_g^T(s_g{H_g}^{-1})^{-1}\beta_g\right\}.$$

We notice that this posterior distribution of β_g does not dependent on the regularized parameter λ directly. Casella (2001) proposed a Monte Carlo EM algorithm that complements a Gibbs sampler and provides marginal maximum likelihood estimates of the regularized parameter λ. For the Bayesian group Lasso, each iteration of the algorithm involves running the Gibbs sampler using a λ value estimated from the sample of the previous iteration.

Finally we put Gamma prior on λ^2 ,$\lambda^2 \sim (\lambda^2)^{b_1-1}\exp(-b_2\lambda^2)$, where b_1 and $b_2 \geq 0$ are constants.

The derivations above lead to the following Gibbs sampling procedure for the quantile θ:

1. Sample $y_i^\star$ from a truncated Normal distribution:

$$y_i^\star|y_i, x_i, \beta, v_i \sim \begin{cases} N(x_i^T\beta + \xi v_i, 2v_i)I(y_i^\star > 0), & if\ y_i = 1, \\ N(x_i^T\beta + \xi v_i, 2v_i)I(y_i^\star \leq 0), & if\ y_i = 0. \end{cases}$$

2. Sample v_i^{-1} , given $y_i^\star$, x_i and β , from an inverse Gaussian distribution with mean and shape parameters given by, respectively,

$$\mu = \sqrt{\frac{1}{(y_i^\star - x_i^T\beta)^2}} \text{ and } \eta = \frac{1}{2}.$$

To derive the distribution of v_i^{-1}, we consider the full conditional distribution of v_i given $y_i^\star, x, v_{-i}, \beta, \lambda$ and θ. This is given by:

$f(v_i|y_i^\star, x, v_{-i}, \beta, \lambda, \theta) \propto f(y|\beta, x, v, \lambda, \theta)\pi(v_i)$

$$\propto \frac{1}{\sqrt{v_i}} exp\left\{-\frac{(u_i - \xi v_i)^2}{4v_i} - \zeta v_i\right\}$$

$$\propto \frac{1}{\sqrt{v_i}} exp\left\{-\frac{({u_i}^2 - 2u_i\xi v_i + \xi^2 {v_i}^2)}{4v_i} - \zeta v_i\right\}$$

$$\propto \frac{1}{\sqrt{v_i}} exp\left\{-\frac{({u_i}^2 + (1-2\theta)^2 {v_i}^2)}{4v_i} - \zeta v_i\right\}$$

$$\propto \frac{1}{\sqrt{v_i}} exp\left\{-\frac{({u_i}^2 + {v_i}^2)}{4v_i}\right\}.$$

Making a variable transformation

$$f({v_i}^{-1}|y_i^\star, x, {v_{-i}}^{-1}, \beta, \lambda, \theta) = f({v_i}^{-1}|y_i^\star, x, {v_{-i}}^{-1}, \beta, \lambda, \theta)\left(\frac{1}{v_i^2}\right)$$

$$f({v_i}^{-1}|y_i^\star, x, {v_{-i}}^{-1}, \beta, \lambda, \theta) \propto {v_i}^{-\frac{3}{2}} exp\left(-\frac{{u_i}^2 v_i}{4} - \frac{1}{4v_i}\right)$$

$$f({v_i}^{-1}|y_i^\star, x, {v_{-i}}^{-1}, \beta, \lambda, \theta) \propto {v_i}^{-\frac{3}{2}} exp\left(-\frac{(v_i - |u_i|^{-1})^2}{4{u_i}^{-2} v_i}\right).$$

This can be recognized as an inverse Gaussian (IG) distribution given by (Chhikara and Folks, 1989) with p. d. f

$$f(x|\eta, \mu) = \sqrt{\frac{\eta}{2\pi}}\ x^{-3/2} \exp\left\{\frac{-\eta(x-\mu)^2}{2(\mu)^2 x}\right\}, \qquad x, \mu, \eta > 0$$

and with parameters $\eta = \frac{1}{2}$ and $\mu = |u_i|^{-1} = \sqrt{\frac{1}{(y_i^\star - x_i^T \beta)^2}}$.

3. Sample s_g, given β_g and λ, from an inverse Gaussian distribution with mean and shape parameters given by, respectively,

$$\mu = \sqrt{\frac{\lambda^2}{\beta_g^T H_g \beta_g}} \text{ and } \eta = \lambda^2.$$

The full conditional distribution of s_g given $y_i^\star, x, v_i, \beta, s_{-k}, \lambda$ and θ can be found by:

$$f\left(s_g \middle| y_i^\star, x, v_i, \beta, s_{-k}, \lambda, \theta\right) \propto f\left(\beta_g \middle| s_g\right) \pi\left(s_g \middle| \lambda\right)$$

$$\propto (s_g)^{-\frac{d_g}{2}} exp\left(-\frac{1}{2}\beta_g^T (s_g H_g^{-1})^{-1} \beta_g\right) s_g^{(d_g - 1)/2} \exp\left(\frac{-\lambda^2}{2} s_g\right)$$

$$\propto s_g^{-1/2} \exp\left(-\frac{1}{2} s_g [\lambda^2 s_g + \beta_g^T H_g \beta_g s_g^{-1}]\right).$$

That is, the full conditional distribution of s_g is again a generalized inverse Gaussian distribution with mean and shape parameters given by

$$\mu = \sqrt{\frac{\lambda^2}{\beta_g^T H_g \beta_g}} \text{ and } \eta = \lambda^2.$$

4. Sample β_g, given $y_i^\star, x_{-g}, \beta_{-g}, s_g, v$ from a multivariate normal distribution with mean and covariance given by

$\mu_g = \Sigma_g x_g V\left(y_i^\star - (1-2\theta)v - x_{-g}^T \beta_{-g}\right)$ and $\Sigma_g = (x_g V x_g^T + s_g^{-1} H_g)^{-1}$,

respectively, where $V = diag\left(\frac{1}{2v_i}\right)$, $i = 1,\ldots,n$, and x_g is the $d_g \times n$ matrix of observations for group g.

The full conditional distribution of β_g can be found by:

$$f\left(\beta_g \middle| y^\star, x, v, \beta_{-g}, \lambda, \theta, s_g\right) \propto exp\left\{-\sum_{i=1}^n \frac{(\tilde{y}_{ig}^\star - x_{ig}^T \beta_g)^2}{4v_i}\right\} exp\left\{-\frac{1}{2}\beta_g^T (s_g^{-1} H_g)\beta_g\right\}$$

$$\propto exp\left\{-\sum_{i=1}^n \frac{(\tilde{y}_{ig}^{\star\,2} - 2\tilde{y}_{ig}^\star x_{ig}^T \beta_g + \left(x_{ig}^T \beta_g\right)^2)}{4v_i}\right\} exp\left\{-\frac{1}{2}\beta_g^T (s_g^{-1} H_g)\beta_g\right\}$$

$$\propto exp\left\{-\sum_{i=1}^n \frac{(\tilde{y}_{ig}^{\star\,2} - 2\tilde{y}_{ig}^\star x_{ig}^T \beta_g + \left(x_{ig}^T \beta_g\right)^T \left(x_{ig}^T \beta_g\right))}{4v_i}\right\} exp\left\{-\frac{1}{2}\beta_g^T (s_g^{-1} H_g)\beta_g\right\}$$

$$\propto exp\left\{-\sum_{i=1}^n \frac{(\tilde{y}_{ig}^{\star\,2} - 2\tilde{y}_{ig}^\star x_{ig}^T \beta_g + \beta_g^T \left(x_{ig} x_{ig}^T\right)\beta_g)}{4v_i}\right\} exp\left\{-\frac{1}{2}\beta_g^T (s_g^{-1} H_g)\beta_g\right\}$$

$$\propto exp\left\{-\frac{1}{2}\left[\beta_g^T (s_g^{-1} H_g + 1/2 \sum_{i=1}^n {v_i}^{-1} x_{ig} x_{ig}^T)\,\beta_g - \sum_{i=1}^n {v_i}^{-1} \tilde{y}_{ig}^\star x_{ig}^T \beta_g\right]\right\}.$$

Let $\tilde{\Sigma}_g^{-1} = s_g^{-1} H_g + 1/2 \sum_{i=1}^n {v_i}^{-1} x_{ig} x_{ig}^T$ and $\tilde{\mu}_g = \tilde{\Sigma}_g \sum_{i=1}^n {v_i}^{-1} \tilde{y}_{ig}^\star x_{ig}^T$.

The full conditional of β_g is then $N(\tilde{\mu}_g, \tilde{\Sigma}_g)$, where $\tilde{y}_{ig}^\star = y_i^\star - \sum_{k=1,k\neq g}^G x_{ik}^T \beta_k - \xi v_i$.

5. Sample λ^2, given s_g, from a Gamma distribution with shape and rate parameters given by, respectively,

$$\alpha = \frac{p+G}{2} + b_1 \text{ and } \beta = \sum_{g=1}^G \frac{s_g}{2} + b_2$$

with b_1 and b_2 two non-negative constants which we set equal to 0.1.

The full conditional distribution of λ^2 is found by:

$$f\left(\lambda^2 \middle| y, x, v, \beta, \theta, s_g\right) \propto \prod_{g=1}^G \pi\left(s_g \middle| \lambda^2\right) \pi(\lambda^2)$$

$$\propto \prod_{g=1}^{G}(\lambda^2)^{(d_g+1)/2}\exp\left(\frac{-\lambda^2}{2}s_g\right)(\lambda^2)^{b_1-1}\exp(-b_2\lambda^2)$$

$$\propto (\lambda^2)^{\frac{p+G}{2}+b_1-1}\exp\left(-\lambda^2\left(\sum_{g=1}^{G}\frac{s_g}{2}+b_2\right)\right).$$

Thus the full conditional of λ^2 is a Gamma distribution with parameters

$\alpha = \frac{p+G}{2} + b_1$ and $\beta = \sum_{g=1}^{G}\frac{s_g}{2} + b_2$, where b_1 and $b_2 \geq 0$ are constants.

3.4 Class prediction

The estimation of the regression coefficients indicates the most influential variables for the prediction of the binary outcome y. As with any regression problem with binary response, the main interest is in the prediction of y from a new instance x for which the binary outcome, or class, is unknown. In this section, we describe how the method that we propose is used to this purpose. The classification of an instance x is based on the estimated probability $P(y = 1|x)$. For our model:

$$P(y = 1|x,\beta,\theta) = P(y_i^\star > 0|x,\beta,\theta) = P(x^T\beta_\theta + u > 0|x,\beta,\theta)$$

$$= P(u > -x^T\beta_\theta|x,\beta,\theta) = 1 - P(u < -x^T\beta_\theta|x,\beta,\theta)$$

$$=1 - \Phi_{ALD}(-x^T\beta_\theta|x,\beta,\theta),$$

where Φ_{ALD} is the cdf of an asymmetric Laplace distribution. Using the estimated β_θ from the binary quantile regression model in the formula above, we get a natural estimate of the posterior probability of x belonging to class 1. Since $P(u < 0|x) = \theta$, it follows that (Kordas, 2006):

$$x^T\beta_\theta \gtreqless 0 \iff P(y = 1|x,\beta,\theta) \gtreqless 1 - \theta\,.$$

So there is a direct link before the estimated β_θ and the probability that $P(y = 1|x) = 1 - \theta$. In general, we can expect the error to have a median around 0, which motivates the choice of $\theta = 0.5$.

In (Kordas, 2006), a second approach is also considered, where $P(y = 1|x, \beta)$ is computed as an average over different quantiles. In particular, it holds that (Kordas, 2006)

$$P(y = 1|x, \beta) = \int_0^1 I(x_i^T \beta_\theta > 0) d_\theta \, .$$

This probability can be estimated using a grid of values $\theta_1, \ldots, \theta_m$ and then taking

$$P(y = 1|x, \beta) \approx \frac{1}{m} \sum_{k=1}^{m} P\left(y = 1 \middle| x, \hat{\beta}_{\theta_k}\right),$$

with $\hat{\beta}_{\theta_k}$ the estimate of β for quantile θ_k.

As a final step in predicting y, we set a threshold t and classify a new object x to class 1 if

$$p(y = 1|x) > t.$$

The threshold t is normally chosen according to the relative misclassification costs for class 0 and 1 and corresponds to the case $t = 0.5$ for equal misclassification costs (Hand and Vinciotti, 2003).

3.5 Simulation study

In this section, we investigate the performance of our method with a simulation study. As typical for these applications, we simulate the data from

$$y_i^\star = x_i^T \beta_\theta + u_i, \quad i = 1, \ldots, n \text{ and } y_i = h(y_i^\star),$$

with β chosen to have a group structure and with different choices of the error distribution. In particular, similar to (Yu et al., 2013) and Li et al. (2010), we consider the following distributions for the error:

■Normal: $N(0;\ 1)$

■Normal: $N(0;\ 9)$

■ A mixture of two normal distributions: $0.1N(0, 10000) + 0.9N(0, 1)$

■A t distribution with 1 degree of freedom (Cauchy): t_1

■ A t distribution with 3 degree of freedom: t_3

■ Laplace distribution with location 0 and scale 10: $Laplace(0, 10)$

■ A mixture of two Laplace distributions: $0.1Laplace\ (0, 1) +\ 0.9Laplace(0, 5)$

■Skewed (skew): $\frac{1}{5}N\left(-\frac{22}{25}, 1\right) + \frac{1}{5}N(-\frac{49}{125}\ , (\frac{3}{2})^2) + \frac{3}{5}N(\frac{49}{250}, (\frac{5}{9})^2)$

■Kurtotic: (kur): $\frac{2}{3}N(0,1) + \frac{1}{3}N(0, (\frac{1}{10})^2)$

■Bimodal (bim): $\frac{1}{2}N(-1, (\frac{2}{3})^2) + \frac{1}{2}N(1, (\frac{2}{3})^2)$

■Bimodal, with separate modes (bim-sep): $\frac{1}{2}N(-\frac{3}{2}, (\frac{1}{2})^2) + \frac{1}{2}N(\frac{3}{2}, (\frac{1}{2})^2)$

■Skewed Bimodal (skew-bim): $\frac{3}{4}N\left(-\frac{43}{100}, 1\right) + \frac{1}{4}N(\frac{107}{100}, (\frac{1}{3})^2)$

■Trimodal (tri): $\frac{9}{20}N(-\frac{6}{5}, (\frac{3}{5})^2) + \frac{9}{20}N(\frac{6}{5}, (\frac{3}{5})^2) + \frac{1}{10}N(0, (\frac{1}{4})^2)$.

These distributions were chosen to have a median close to or equal to zero. Figure 3.1 shows a plot of the density functions for some of the cases considered. For the simulation, we set the sample size $n = 50$.

For the β vector, we consider the case of a large number of predictors, i.e. $p \gg n$. Similar to Li et al. (2010), we create a group structure by simulating 10 groups, each consisting of 10 covariates. The 100 variables are assumed to follow a multivariate normal distribution $N(0; \Sigma)$, with Σ having a diagonal block structure. Each block corresponds to one group and is defined by the matrix $r^{|i-k|}$, $i = i = 1,. \ . \ .,10$, $k = 1,. \ . \ .,10$. For the correlation r, we experiment both with $r = 0.95$ (well-defined group structure) and $r = 0.5$. For the β values we consider two cases:

(1) The values for the first three groups are given by

$(0.5, 1, 1.5, 2, 2.5, 2, 2, 2, 2, 2), (2, 2, 1, 1, 1, 1, 3, 3, 3, 3), (1, 1, 1, 2, 2, 2, 3, 3, 3, 3),$

and they are set to zero for all other groups.

(2) $\beta_j = 0.85$ for all j.

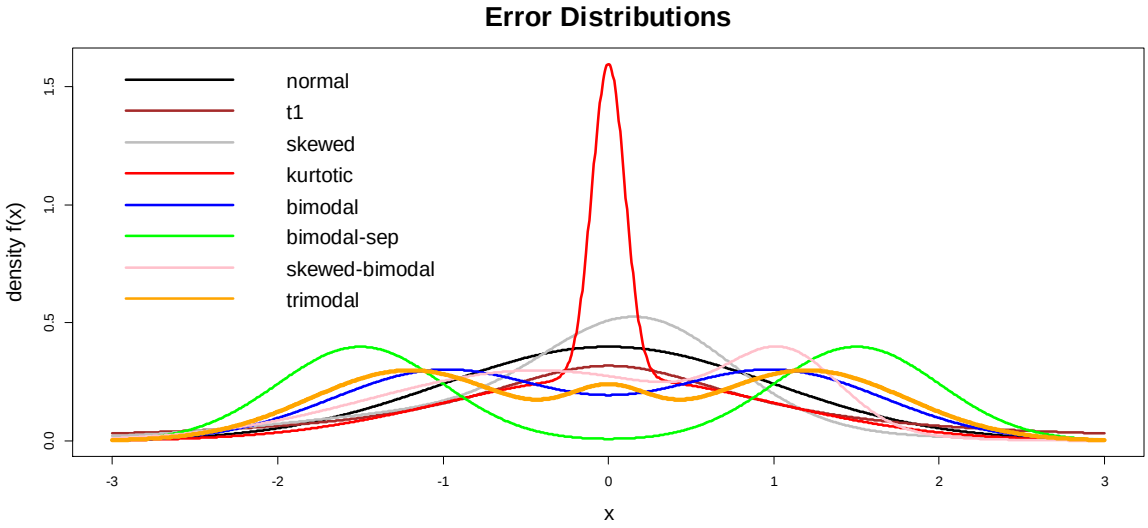

Figure 3.1: Some density functions of the errors considered in the simulation study.

For the Bayesian quantile methods and their Gibbs sampling procedures, we use 13000 iterations with the first 3000 iterations kept as burn-in. For checking convergence, we present trace plots for some βs. Figure 3.2 shows the trace plots for simulated data under case 1, $r = 0.5$ and $\theta = 0.5$. The plots suggest that the constructed chain mix quickly and has good convergence. Furthermore, in the quantile methods, we use two methods to make class predictions, as described in Section 3.4: in the first case, we use the median ($\theta = 0.5$); in the second case we take an average of three quantiles, which we take as $\theta = 0.25, 0.5, 0.75$. We compare our method, Bayesian binary quantile regression with group lasso penalty ($BBQ.grplasso$), with a frequentist mean-based logistic regression model under a lasso penalty (R package $glmnet$) (Friedman et al., 2010), a frequentist mean-based logistic regression model under a group lasso penalty (R package $grpreg$) (Breheny and Huang, 2014) and a Bayesian binary quantile regression with a lasso penalty (R package bayesQR) (Benoit et al., 2013). For $grpreg$ and $glmnet$, the penalty parameter λ is selected using 5-fold cross-validation.

Table 3.1, 3.2, 3.3 and 3.4 reports the Area Under the Curve (AUC) values for the different methods and the different error distributions, with the AUC values averaged over 40 iterations and computed on a test set of the same size as the training set. In Table 1 and 2, we consider the first scenario for the β values and we set $r=0.5$ and r=0.95 respectively, whereas in Table 3 and 4 we consider the case of all βs equal to 0.85 and $r=0.5$ and $r=0.95$ respectively. No significant differences were found between the two approaches for prediction used for the Bayesian methods. Values in bold in Table 3.1, 3.2, 3.3 and 3.4 show how the $BBQ.grplasso$ proposed in this chapter statistically significant outperforms the other methods in all cases considered in Table 3.1, 3.4 and some cases in Table 3.3. Furthermore, the results show how $grpreg$ is the worst performing method in all cases, surprisingly performing worse than $glmnet$, which is of a same nature but does not exploit the group structure of the predictors. The main competitor to $BBQ.grplasso$ seems to be $bayesQR$ which in fact differs with the proposed method only in the use of the lasso penalty in contrast to the group lasso penalty.

The results in Table 3.1 are found by implemented R functions. The running time of R functions depend upon the computer used and the size of the data set (p and n). As an example, we consider the data set in the simulation study in section 3.5 with $p = 100$ and $n = 50$. We use a computer with 2.4 GHz processor and 6 gigabytes of RAM. Computation of AUC values in Table 3.1 for one quantile takes 6 minutes for $BBQ.grplasso$, 4 minutes for $bayesQR$, 0.5 minutes for $grpreg$ and 0.5 minutes for $glmnet$. According to the results in Table 3.1 we conclude that the time-consuming of the proposed method is much larger than the other methods.

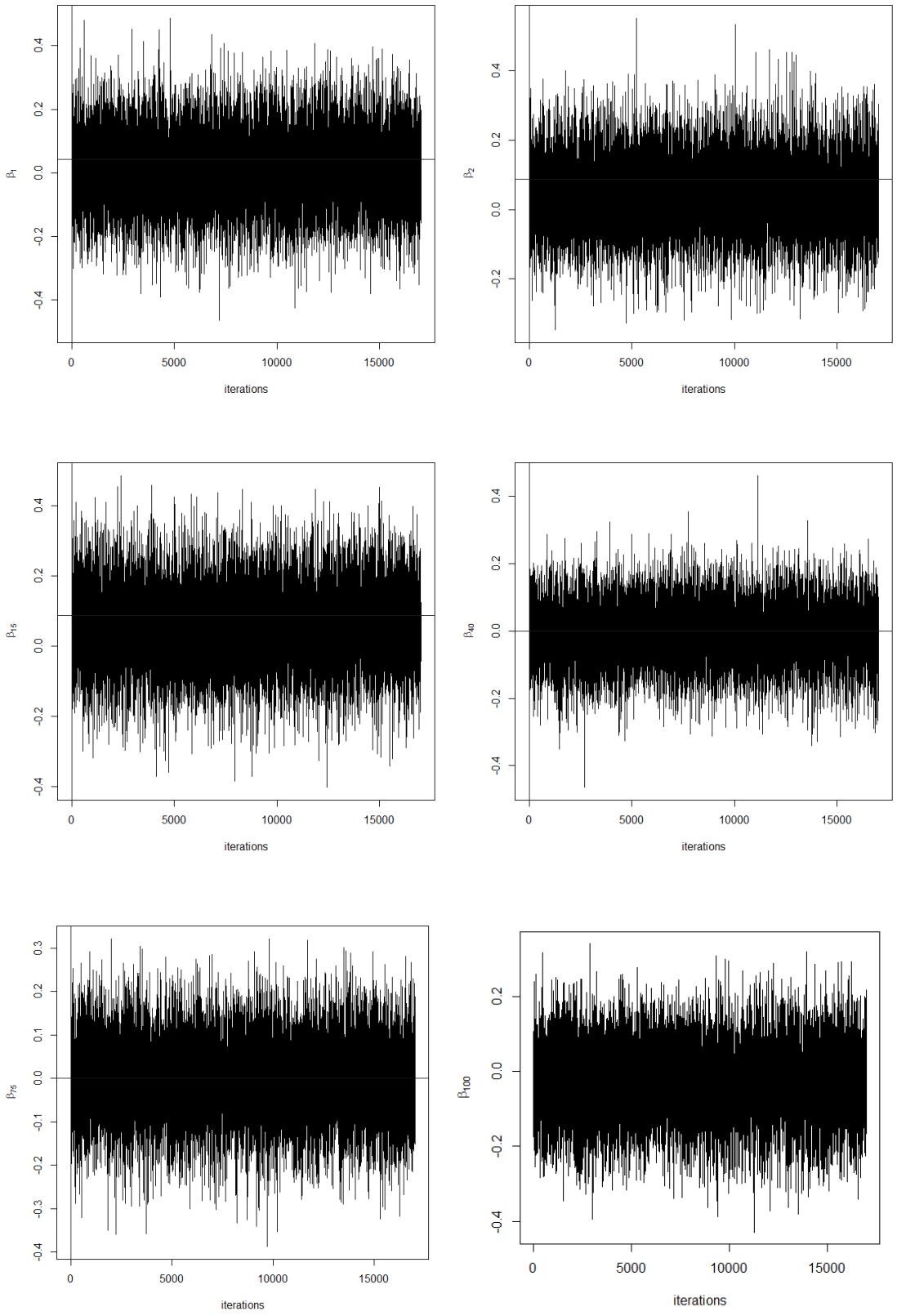

Figure 3.2: Trace plots for some selected $\hat{\beta}_j's$ *at* $r=0.5$ *and quantile 0.5 for simulation case1. The horizontal line refers to* $\frac{\beta_j}{\|\beta_j\|}$.

Table 3.1: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.5$ *and* β *values as in case (1).* $BBQ.grplasso$*: Bayesian binary quantile regression model proposed in this chapter (based on* $\theta = 0.5$ *(median) and an average of the* $\theta = 0.25, 0.5, 0.75$ *quantiles);* $grpreg$ *: frequentist mean-based logistic regression model with group lasso penalty,* $glmnet$*: frequentist mean-based logistic regression model under a group lasso penalty;* $bayesQR$*: Bayesian binary quantile regression with a lasso penalty (based on* $\theta = 0.5$ *(median) and an average of the* $\theta = 0.25, 0.5, 0.75$ *quantiles). Best mean indicated in bold.*

	BBQ. grplasso ($\theta = 0.5$)	BBQ. grplasso ($\theta = 0.25, 0.5, 0.75$)	grpreg	glmnet	BayesQR ($\theta = 0.5$)	bayesQR ($\theta = 0.25, 0.5, 0.75$)
$N(0,1)$	**0.879** (0.055)	**0.88** (0.053)	0.773 (0.103)	0.804 (0.078)	0.83 (0.066)	0.838 (0.068)
$N(0,3)$	**0.89** (0.055)	**0.89** (0.054)	0.777 (0.106)	0.842 (0.084)	0.841 (0.06)	0.842 (0.068)
$Normal\ M.$	**0.785** (0.068)	**0.785** (0.067)	0.664 (0.114)	0.728 (0.07)	0.741 (0.069)	0.751 (0.068)
t_1	**0.838** (0.06)	**0.838** (0.06)	0.725 (0.116)	0.765 (0.116)	0.787 (0.066)	0.797 (0.068)
t_3	**0.892** (0.048)	**0.891** (0.048)	0.768 (0.117)	0.793 (0.097)	0.835 (0.049)	0.843 (0.042)
$Laplace$	**0.766** (0.089)	**0.764** (0.089)	0.64 (0.125)	0.718 (0.113)	0.722 (0.087)	0.727 (0.089)
$Laplace\ M.$	**0.834** (0.058)	**0.833** (0.06)	0.696 (0.112)	0.761 (0.095)	0.791 (0.062)	0.792 (0.068)
skew	**0.885** (0.043)	**0.886** (0.043)	0.792 (0.087)	0.811 (0.086)	0.832 (0.049)	0.837 (0.051)
kur	**0.89** (0.049)	**0.888** (0.05)	0.775 (0.112)	0.816 (0.082)	0.843 (0.058)	0.846 (0.052)
bim	**0.898** (0.05)	**0.898** (0.05)	0.782 (0.098)	0.789 (0.117)	0.853 (0.066)	0.857 (0.065)
bim − sep	**0.881** (0.053)	**0.881** (0.053)	0.784 (0.096)	0.819 (0.078)	0.826 (0.066)	0.829 (0.072)
ske − bim	**0.885** (0.054)	**0.886** (0.055)	0.797 (0.099)	0.814 (0.09)	0.838 (0.067)	0.848 (0.066)
tri	**0.879** (0.053)	**0.879** (0.054)	0.774 (0.117)	0.822 (0.074)	0.82 (0.064)	0.829 (0.061)
Computational time of one replication (minutes)	6	18	0.5	0.5	4	12

Table 3.2: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.95$ and β values as in case (1). $BBQ.grplasso$: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); $grpreg$: frequentist mean-based logistic regression model with group lasso penalty, $glmnet$: frequentist mean-based logistic regression model under a group lasso penalty; $bayesQR$: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles).

	BBQ. grplasso ($\theta = 0.5$)	BBQ. grplasso ($\theta = 0.25, 0.5, 0.75$)	grpreg	glmnet	BayesQR ($\theta = 0.5$)	bayesQR ($\theta = 0.25, 0.5, 0.75$)
$N(0,1)$	0.968 (0.022)	0.967 (0.023)	0.789 (0.111)	0.967 (0.023)	0.929 (0.042)	0.942 (0.044)
$N(0,3)$	0.965 (0.024)	0.966 (0.023)	0.788 (0.083)	0.964 (0.026)	0.937 (0.033)	0.948 (0.036)
$Normal\ M.$	0.926 (0.034)	0.927 (0.033)	0.722 (0.118)	0.924 (0.044)	0.89 (0.055)	0.895 (0.046)
t_1	0.944 (0.036)	0.945 (0.036)	0.754 (0.103)	0.949 (0.033)	0.91 (0.032)	0.92 (0.036)
t_3	0.967 (0.024)	0.967 (0.024)	0.778 (0.113)	0.964 (0.025)	0.934 (0.036)	0.946 (0.032)
$Laplace$	0.89 (0.05)	0.889 (0.05)	0.71 (0.1)	0.887 (0.057)	0.862 (0.049)	0.872 (0.055)
$Laplace\ M.$	0.954 (0.026)	0.954 (0.026)	0.757 (0.109)	0.951 (0.029)	0.922 (0.038)	0.926 (0.039)
skew	0.957 (0.027)	0.946 (0.032)	0.758 (0.13)	0.963 (0.029)	0.933 (0.032)	0.951 (0.03)
kur	0.971 (0.022)	0.971 (0.022)	0.763 (0.105)	0.967 (0.023)	0.935 (0.039)	0.955 (0.032)
bim	0.965 (0.026)	0.965 (0.025)	0.774 (0.121)	0.961 (0.028)	0.924 (0.042)	0.944 (0.041)
bim − sep	0.95 (0.039)	0.941 (0.046)	0.757 (0.104)	0.967 (0.026)	0.928 (0.039)	0.95 (0.035)
ske − bim	0.97 (0.022)	0.97 (0.022)	0.764 (0.107)	0.967 (0.024)	0.934 (0.039)	0.954 (0.032)
tri	0.955 (0.025)	0.947 (0.03)	0.766 (0.096)	0.97 (0.02)	0.931 (0.033)	0.943 (0.033)

Table 3.3: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.5$ *and* β *values as in case (2).* $BBQ.grplasso$*: Bayesian binary quantile regression model proposed in this chapter (based on* $\theta = 0.5$ *(median) and an average of the* $\theta = 0.25, 0.5, 0.75$ *quantiles);* $grpreg$ *: frequentist mean-based logistic regression model with group lasso penalty,* $glmnet$*: frequentist mean-based logistic regression model under a group lasso penalty;* $bayesQR$*: Bayesian binary quantile regression with a lasso penalty (based on* $\theta = 0.5$ *(median) and an average of the* $\theta = 0.25, 0.5, 0.75$ *quantiles). Best mean indicated in bold.*

	BBQ. grplasso ($\theta = 0.5$)	BBQ. grplasso ($\theta = 0.25, 0.5, 0.75$)	grpreg	glmnet	BayesQR ($\theta = 0.5$)	bayesQR ($\theta = 0.25, 0.5, 0.75$)
$N(0,1)$	**0.887** (0.048**)**	**0.888** (0.049)	0.595 (0.094)	0.674 (0.105)	0.86 (0.049)	0.862 (0.05)
$N(0,3)$	**0.854** (0.066)	**0.853** (0.065)	0.598 (0.102)	0.655 (0.114)	0.828 (0.056)	0.823 (0.06)
$Normal\ M.$	0.723 (0.101)	0.722 (0.101)	0.54 (0.084)	0.579 (0.104)	0.706 (0.091)	0.709 (0.096)
t_1	0.824 (0.071)	0.825 (0.071)	0.594 (0.101)	0.627 (0.112)	0.796 (0.072)	0.807 (0.07)
t_3	**0.87** (0.054)	**0.868** (0.055)	0.585 (0.103)	0.642 (0.119)	0.842 (0.059)	0.847 (0.062)
$Laplace$	**0.718** (0.075)	**0.718** (0.073)	0.557 (0.084)	0.58 (0.098)	0.689 (0.07)	0.69 (0.073)
$Laplace\ M.$	0.79 (0.083)	0.79 (0.082)	0.572 (0.101)	0.618 (0.099)	0.761 (0.082)	0.765 (0.083)
skew	**0.89** (0.058)	**0.889** (0.059)	0.599 (0.104)	0.667 (0.113)	0.861 (0.059)	0.863 (0.064)
kur	0.875 (0.061)	0.875 (0.06)	0.624 (0.1)	0.681 (0.105)	0.854 (0.064)	0.858 (0.066)
bim	**0.865** (0.058)	**0.864** (0.058)	0.598 (0.103)	0.656 (0.106)	0.83 (0.06)	0.841 (0.058)
bim − sep	**0.859** (0.061)	**0.858** (0.06)	0.585 (0.105)	0.652 (0.112)	0.83 (0.058)	0.834 (0.064)
ske − bim	**0.876** (0.062)	**0.877** (0.061)	0.611 (0.103)	0.673 (0.104)	0.848 (0.067)	0.851 (0.069)
tri	**0.878** (0.058)	**0.877** (0.057)	0.593 (0.087)	0.643 (0.115)	0.845 (0.063)	0.854 (0.062)

Table 3.4: AUC values, averaged over 40 replications (with standard deviations in brackets) for the case: $n = 50, p = 100, r = 0.95$ and β values as in case (2). $BBQ.grplasso$: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); $grpreg$: frequentist mean-based logistic regression model with group lasso penalty, $glmnet$: frequentist mean-based logistic regression model under a group lasso penalty; $bayesQR$: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles). Best mean indicated in bold.

	BBQ. grplasso ($\theta = 0.5$)	BBQ. grplasso ($\theta = 0.25, 0.5, 0.75$)	grpreg	glmnet	BayesQR ($\theta = 0.5$)	bayesQR ($\theta = 0.25, 0.5, 0.75$)
$N(0,1)$	**0.962** (0.026)	**0.962** (0.027)	0.606 (0.104)	0.895 (0.046)	0.928 (0.042)	0.943 (0.046)
$N(0,3)$	**0.953** (0.037)	**0.953** (0.036)	0.616 (0.106)	0.889 (0.066)	0.91 (0.053)	0.928 (0.057)
$Normal\ M.$	**0.908** (0.047)	**0.907** (0.048)	0.566 (0.088)	0.845 (0.058)	0.868 (0.056)	0.884 (0.047)
t_1	**0.943** (0.033)	**0.943** (0.033)	0.572 (0.096)	0.878 (0.07)	0.911 (0.048)	0.923 (0.044)
t_3	**0.966** (0.026)	**0.966** (0.027)	0.6 (0.104)	0.895 (0.084)	0.928 (0.037)	0.947 (0.039)
$Laplace$	**0.872** (0.048)	**0.872** (0.048)	0.57 (0.08)	0.784 (0.1)	0.834 (0.047)	0.848 (0.047)
$Laplace\ M.$	**0.927** (0.049)	**0.927** (0.048)	0.591 (0.094)	0.862 (0.061)	0.888 (0.05)	0.905 (0.052)
skew	**0.96** (0.025)	**0.958** (0.026)	0.602 (0.104)	0.888 (0.063)	0.925 (0.038)	0.944 (0.039)
kur	**0.954** (0.03)	**0.955** (0.031)	0.646 (0.1)	0.876 (0.087)	0.917 (0.043)	0.941 (0.037)
bim	**0.963** (0.033)	**0.962** (0.034)	0.561 (0.086)	0.901 (0.067)	0.924 (0.045)	0.94 (0.048)
bim − sep	**0.967** (0.029)	**0.966** (0.028)	0.609 (0.119)	0.899 (0.068)	0.935 (0.036)	0.948 (0.041)
ske − bim	**0.969** (0.019)	**0.968** (0.02)	0.592 (0.107)	0.912 (0.048)	0.935 (0.033)	0.951 (0.025)
tri	**0.96** (0.036)	**0.959** (0.036)	0.601 (0.101)	0.89 (0.071)	0.928 (0.047)	0.94 (0.042)

Figure 3.3 confirms the results of the tables by showing the average ROC curve of the methods considered for two cases of error distributions. The figures show how the $BBQ.grplasso$ outperforms the other methods for all classification thresholds and has $bayesQR$ as its main competitor.

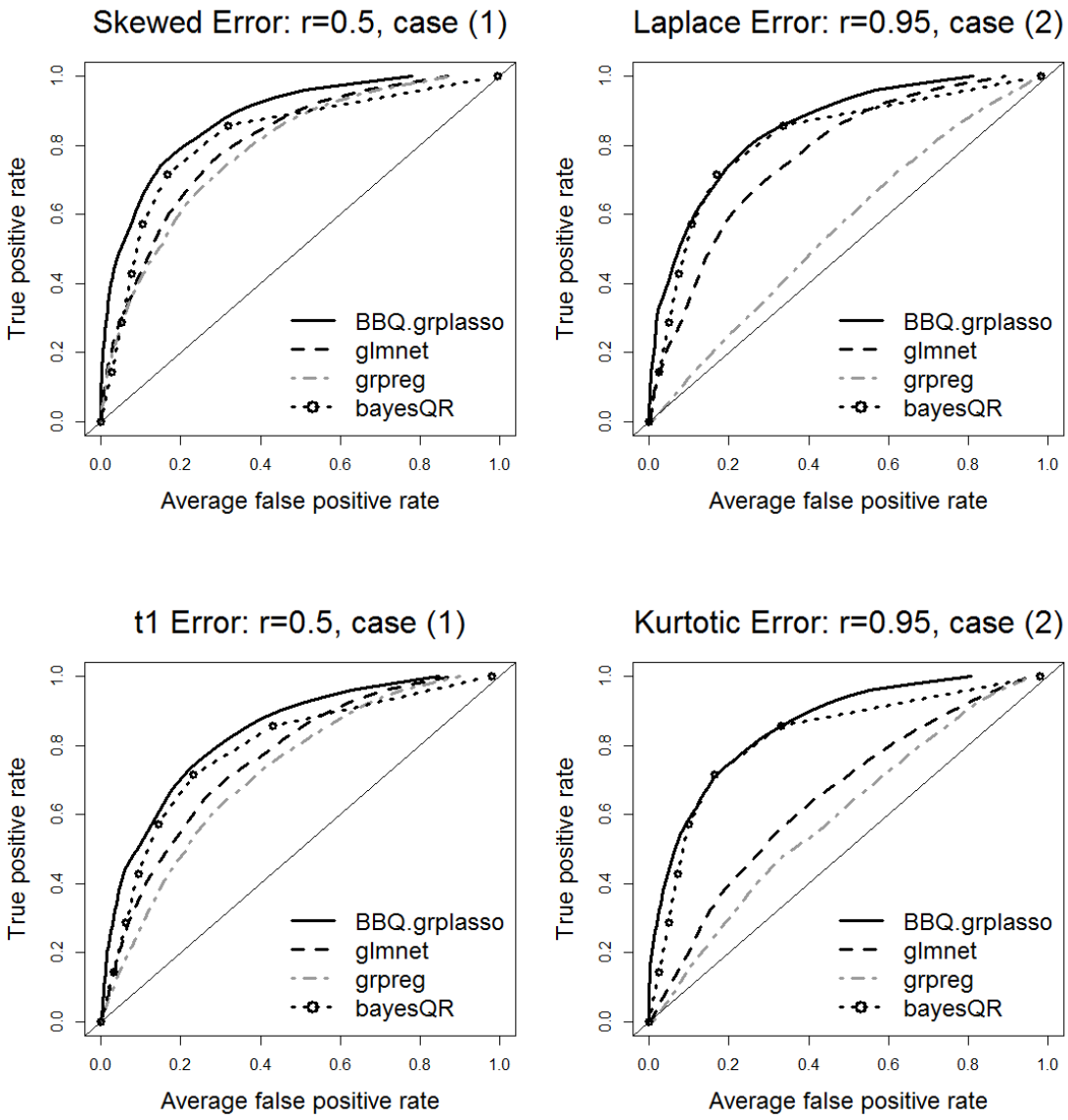

Figure 3.3: Average ROC curves (over 40 replications) of Bayesian binary quantile regression with group lasso ($BBQ.grplasso, \theta = 0.5$), compared with $grpreg$, $glmnet$ and $bayesQR$, under a Skewed (left top panel), a Laplace (right top panel), t_1 (left bottom panel) and Kurtotic (right bottom panel) error distribution.

3.6 Real application

In this section, we investigate the performance of the new method on five real applications:

■ **Birth weight dataset:** This dataset is available in the R package $grpreg$ and was used in (Yuan and Lin, 2006). The data record the birth weights of 189 babies, together with eight predictors. Among the predictors, two are continuous (mother's age and

weight) and six are categorical (mother's race, smoking status during pregnancy, number of previous premature labours, history of hypertension, presence of uterine irritability, number of physician visits). Through the use of orthogonal polynomials and dummy variables, the data is converted into 17 predictors. The goal of this study is to identify the risk factors associated with giving birth to a low birth weight baby (defined as weighing less than 2500g).

■ **Colon dataset:** This dataset is available in the R package $gglasso$ and was used by Yang and Zou (2013). The data report the expression level of 20 genes from 62 colon tissue samples, of which 40 are cancerous and 22 normal. In [37], the 20 expression profiles are expanded using 5 basis B-splines, creating a dataset with 100 predictors and a group structure.

■ **Labor force participation dataset:** This dataset is available in the R package AER and was used in (Liu et al., 2013). The data come from the Panel Study of Income Dynamics (PSID) in 1976 and contain 753 observations on women's labour supply and 18 variables. We used this dataset to demonstrate the performance of our method. The response variable, wife's participation in work, is a binary variable and some predictors are correlated so that binary group lasso model can be used. The aim of this analysis is to assess if there is a relation between several social factors (wife's age, husband's wage and wife's father education etc.) and wife's participation in work, how strong this relation is and what the influence of the factors are. Our grouping structure is similar to Liu et al. (2013) and the details are shown in Table 3.5.

Table 3.5: Variables in the labor force participation dataset

Group name	Variable name	Description of variable
G1	WE	Wife's educational attainment ,in years
	WW	Wife's average hourly earnings, in dollars
	RPWG	Wife's wage reported
	FAMINC	Family income
	MTR	Marginal tax rate of wife
	AX	Actual years of wife's previous labor market experience
G2	KL6	Number of children less than 6 years old in household
	K618	Number of children between ages 6 and 18 in household
G3	HE	Husband's educational attainment, in years
	HW	Husband's wage, in dollars
G4	WMED	Wife's mother's educational attainment, in years
	WFED	Wife's father's educational attainment, in years
G5	UN	Unemployment rate, in percentage points.
	CIT	Dummy variable = 1 if live in large city , else 0
G6	WA	Wife's age
	HA	Husband's age
G7	HHRS	Husband's hours worked

■ **Splice site detection dataset:** This dataset is available in the R package *grplasso* and is a random sample of the data used by Gene and Burge (2004) and Meier et al. (2008). It contains information on 200 true human donor splice sites and 200 false splice sites. For each site, the data report the last three bases of the exon and the third to sixth bases of the intron. Thus, the data contain 7 categorical predictors, with values A, C, G and T. These are converted into dummy variables, creating a natural group structure. We used this dataset to explain the performance of our method. As the response variable is a binary (true or false splice sites) and the predictors are categorical and converted to dummy variables, binary group lasso is appropriate method.

■ **Cleveland heart dataset**: This dataset is available from the UCI machine learning repository. The data report information on 297 patients, 160 of whom have been

diagnosed with heart disease and the remaining 137 have not been diagnosed with heart disease. The goal of the study is to predict heart disease from 13 predictors, related to patients' characteristics (age, sex, etc) and clinical information (blood pressure, cholesterol level, etc). Four of the predictors are categorical and have been converted into dummy variables, creating a group structure. The grouping structure and descriptions of attributes are shown in Table 3.6. The data obtained from the website http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

Table 3.6: Variables in the heart disease dataset

Group name	Variable name	Description of variable
G1	age	Age in Year
G2	sex	Sex (value 1: Male; value 0 : Female)
G3	cp	Chest Pain Type (value 1 – 4) (Converted into 3 dummy variables)
G4	trestbps	Trest Blood Pressure (mm Hg on admission to the hospital)
G5	cho	Serum Cholesterol (mg/dl)
G6	fbs	Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
G7	restecg	resting electrographic results (value 0 – 2) (Converted into 2 dummy variables)
G8	thalach	maximum heart rate achieved
G9	exang	exercise induced angina (value 1: yes; value 0: no)
G10	oldpeak	ST depression induced by exercise relative to rest
G11	slope	the slope of the peak exercise ST segment(value 1 – 3) (Converted into 2 dummy variables)
G12	ca	number of major vessels colored by fluoroscopy (value 0 – 3)
G13	thal	Thal (value 3: normal; value 6: fixed defect; value7:reversible defect) (Converted into 2 dummy variables)
Response	num	Class label representing four type of Heart Disease {0,1,2,3,4}

Tables 3.7 reports the AUC values of 5-fold cross validation ROC curves, averaged over 5 iterations. As before, we compare the binary quantile regression method presented in this chapter, $BBQ.grplasso$, with $grpreg$, $glmnet$ and $BayesQR$. The results show how $BBQ.grplasso$ is generally superior to $BayesQR$ on all datasets, it outperforms the other methods in the Colon dataset, but has comparable performances with the frequentist methods for the remaining datasets. Combined with the simulation study, this is probably a reflection of high levels of sparsity in the underlying model.

Table 3.7: AUC values, averaged over 5 replications (with standard deviations in brackets) on real data. $BBQ.grplasso$: Bayesian binary quantile regression model proposed in this chapter (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles); $grpreg$: frequentist mean-based logistic regression model with group lasso penalty, $glmnet$: frequentist mean-based logistic regression model under a group lasso penalty; $bayesQR$: Bayesian binary quantile regression with a lasso penalty (based on $\theta = 0.5$ (median) and an average of the $\theta = 0.25, 0.5, 0.75$ quantiles).

Data set	BBQ. grplasso ($\theta = 0.5$)	BBQ. grplasso ($\theta = 0.25, 0.5, 0.75$)	grpreg	glmnet	BayesQR ($\theta = 0.5$)	bayesQR ($\theta = 0.25, 0.5, 0.75$)
Birth	0.582 (0.042)	0.584 (0.039)	0.577 (0.041)	0.541 (0.068)	0.539 (0.041)	0.577 (0.02)
Colon	0.641 (0.037)	0.645 (0.036)	0.612 (0.091)	0.618 (0.073)	0.573 (0.056)	0.592 (0.052)
Labor	0.708 (0.016)	0.711 (0.015)	0.702 (0.024)	0.718 (0.024)	0.502 (0.009)	0.596 (0.031)
Splice	0.694 (0.016)	0.695 (0.016)	0.699 (0.018)	0.694 (0.02)	0.685 (0.015)	0.689 (0.022)
Heart	0.663 (0.015)	0.662 (0.014)	0.665 (0.014)	0.666 (0.009)	0.508 (0.02)	0.597 (0.019)

In terms of parameter estimates, since Bayesian regularized methods do not give exact zeros, we consider credible intervals to select which parameters are different from zero. The Bayesian estimates are obtained based on 16000 MCMC iteration with 3000 burn-in for birth data. Tables 3.5 and 3.6 shows the fitted coefficients for the *0.5th* quantile and *0.95th* quantile, along with their 95% credible intervals, for $BBQ.grplasso$ and the results of mean binary regression with lasso penalty ($glmnet$). The results show similar performances of the two methods on the birth dataset in both

quantiles. We conclude that for this application either the predictors in each group are not highly correlated or the regression has high levels of sparsity.

Table 3.8: 95% *credible intervals for birth dataset at* $\theta = 0.5$.

	Methods	BBQ. grplasso		*glmnet*
Group name	Variables	Lower lower 2.5%	Upper 97.5%	Mean
	Intercept	-1.890	-0.533	-0.849
G1	age1	-6.625	2.205	0
	age2	-7.096	2.520	0
	age3	-5.279	4.216	0
G2	lwt1	-8.177	0.781	0
	lwt2	-4.534	4.627	0
	lwt3	-6.763	1.853	0
G3	white	-1.018	0.274	0
	black	-0.594	1.185	0
G4	smoke	-0.215	1.136	0
G5	ptl1	0.354	2.187	0.437
	ptl2m	-2.196	1.670	0
G6	ht	-0.201	2.401	0
G7	ui	-0.387	1.410	0
G8	ftv1	-1.011	0.433	0
	ftv2	-1.104	0.608	0
	ftv3m	-0.947	1.515	0

Table 3.9: 95% *credible intervals for birth dataset at* $\theta = 0.95$.

	Methods	BBQ. grplasso		*glmnet*
Group name	Variables	Lower lower 2.5%	Upper 97.5%	Mean
	Intercept	2.881	6.679	-0.849
G1	age1	-8.029	4.325	0
	age2	-7.663	5.636	0
	age3	-6.761	6.276	0
G2	lwt1	-8.354	4.088	0
	lwt2	-6.365	6.873	0
	lwt3	-7.958	4.569	0
G3	white	-1.110	0.610	0
	black	-0.995	1.504	0
G4	smoke	-0.499	1.237	0
G5	ptl1	-0.508	1.801	0.437
	ptl2m	-2.413	2.896	0
G6	ht	-0.993	2.405	0
G7	ui	-0.683	1.652	0
G8	ftv1	-1.189	0.793	0
	ftv2	-1.296	1.128	0
	ftv3m	-1.460	2.047	0

3.7 Chapter conclusion

In this chapter, we present a novel method for binary regression problems where the predictors have a natural group structure, such as in the case of categorical variables. In contrast to existing methods for group-typed variables, we model the quantiles of the response variable, in order to account for possible departures from normality in the latent variable. In particular, we focus on class prediction and show how the probability of a new object x belonging to class 1, $p(1|x)$, is directly linked to the quantile of the latent variable, since $p(1|x) = p(y_i^\star > 0|x)$.This motivates the use of quantile-based regression for probit regression models.

We compare our method with a frequentist mean-based logistic regression model, under a lasso and a group lasso penalty, and with a Bayesian quantile-based regression model under a lasso penalty, on simulated and real data. The simulation shows a number of scenarios where the method outperforms the mean-based and quantile-based approaches. Future research will consider an extension of this method to include a variable selection prior, in a similarly to the method of Alhamzawi and Yu (2013) for Bayesian quantile regression.

Chapter 4

Tobit Quantile Regression with Group Lasso

In chapter 2 and 3 we considered regularized and robust regression methods for the case when the response variable is continuous and binary, respectively. In this chapter we consider regularized and robust regression models when the response variable is censored. Censored regression, or Tobit model, is an important regression model and has been widely used in econometrics. However, studies for variable selection problem in tobit regression model are limited in the literature. In this chapter, we propose quantile regression with group lasso for a tobit regression model. An MCMC computation method is used to update the parameters from the posterior using an Asymmetric Laplace Distribution (ALD). Simulation studies are used to compare the performance of the proposed method with tobit quantile regression with an adaptive lasso penalty.

4.1 Introduction

In econometrics censored regression models widely arise in cases where the variable of interest is only observable under some conditions. A common example is labor supply. Data in this case are available on the hours worked by employees, and a labor supply model explains the relationship between hours worked and characteristics of employees such as age, education and family status. However, we know age, education and family status for people who are unemployed but it is not possible to observe the number of hours they have worked. Many censored variables have the following characteristics: the variable is left-censored, right-censored, or both left-censored and right-censored, where the lower and/or upper limit of the dependent variable can be any number. In the

tobit model (Tobin, 1958), we have a dependent variable y that is left-censored at zero. So the dependent variable has a positive limitation, that is, only positive response values can be observed.

Many authors studied the statistical inference of censored regression models. These can be found in the literature, such as (Tobin, 1958; Powell, 1984; Pollard, 1990; Phillips 2002; Wang et al., 2007c, 2009; Barros et al. 2010). For variable selection of censored regression models, Wang et al. (2010) used the least absolute shrinkage and selection operator method (LASSO). Zhou et al. (2013) used the least absolute deviation (LAD) variable selection for the linear model with randomly censored data. However, these methods select variables individually.

As we discussed in chapter 3, the group lasso is an appropriate method when there is a group structure, for example, a categorical variable is represented by a group of dummy variables (Yuan and Lin, 2006). Recently, Liu et al. (2013) proposed the group lasso for variable selection and estimation in the tobit censored response model.

In case the data do not satisfy the normal distributional assumptions underlying traditional methods or when the data are subject to some form of contamination, Yu and Stander (2007) proposed Bayesian analysis of a tobit quantile regression model, Reich and Smith (2013) proposed Bayesian quantile regression for censored data and Ji et al., (2012) used Gibbs sampler for model selection in binary and tobit quantile regression. In this chapter, we develop a group variable selection method for the tobit censored quantile regression so we combine the work on tobit quantile regression with the work on tobit regression with lasso regularization. The rest of this chapter is organised as follows. Sections 4.2 and 4.3 introduce the modification of tobit model in quantile regression with group lasso penalty as well as presenting the Bayesian MCMC scheme

for the estimation of the parameters. The computation of predicted values is given in section 4.4. In Section 4.5, simulation scenarios are implemented to test the behaviour of the proposed method for estimation and variable selection. Section 4.6 provides an illustration of the proposed methods using the labor force participation dataset. A chapter conclusion follows in Section 4.7.

4.2 The model

Similar to a binary regression model, tobit quantile regression models can be viewed as linear quantile regression models with a latent continuous response variable, e.g. (Ji et al., 2012). In particular, let y be the response variable, let x be the vector of p predictors, β the vector of unknown regression coefficients and $(x_i, y_i), i = 1, \ldots, n$ a sample of n observations on x and y. Given a quantile θ, $0 < \theta < 1$, we consider the model:

$$y_i^\star = x_i^T \beta_\theta + u_i \,, i = 1, \ldots, n \text{ and } y_i = h(y_i^\star),$$

where u_i are the errors, $y_i^\star$ is an unobserved (''latent'') variable and h is a link function. For tobit response data, the link function is given by $h(y^\star) = max(y^\star, c)$, for a known constant c. In real applications, y (dependent variable) is censored, e.g. the number of hours worked, the amount of money that an individual spends on tobacco, given his or her characteristics. Then $y > 0$ if the individual is a smoker, and $y = 0$ if not (Henningsen, 2012). If the dependent variable is censored (e.g. zero in the above examples), parameter estimates obtained by regression methods (e.g. OLS) are biased. Consistent estimates can be obtained by the method proposed by Tobin (1958). This method is usually called ''Tobit'' model and is a special case of the censored regression model.

4.3 Bayesian parameter estimation

In tobit quantile regression model, the parameter β_θ is found by the following minimization problem:

$$\min_\beta \sum_{i=1}^{n} \rho_\theta (y_i - h(x_i^T \beta)) \tag{4.2}$$

where $\rho_\theta(.)$ is the check loss function defined by

$$\rho_\theta(t) = \begin{cases} \theta t & if \quad t \geq 0, \\ -(1-\theta)t & if \quad t < 0. \end{cases}$$

and where $h(x_i^T \beta) = max\{x_i^T \beta, c\}$, and $\theta \in (0,1)$.

As pointed out in chapter 3, Yu and Moyeed (2001) have shown how minimizing (4.2) is equivalent to maximising the likelihood function, under the assumption that the error comes from an asymmetric Laplace distribution with density given by $f_\theta(\mu) = \frac{\theta(1-\theta)}{\sigma} \exp(-\frac{1}{\sigma}\rho_\theta(u))$. That is, minimising (4.2) is equivalent to maximising the likelihood

$$f(y|x, \beta, \theta) = \frac{\theta^n (1-\theta)^n}{\sigma^n} exp\left(-\sum_{i=1}^{n} \frac{p_\theta(y_i - h(x_i^T \beta)}{\sigma}\right). \tag{4.3}$$

Under a group lasso constraint, the minimization in (4.2) becomes

$$\min_\beta \sum_{i=1}^{n} \rho_\theta (y_i - h(x_i^T \beta)) + \lambda \sum_{g=1}^{G} \left\|\beta_g\right\|_{H_g}, \tag{4.4}$$

where λ is a non-negative regularization parameter, controlling the sparsity of the solution, and $\left\|\beta_g\right\|_{H_g} = (\beta_g^T H_g \beta_g)^{\frac{1}{2}}$ with $H_g = d_g I_{d_g}$ and d_g the dimension of the vector β_g. Under an appropriate choice of prior distribution, the minimization problem

in (4.4) can be shown to be equivalent to a maximum a posterior solution. In particular, a Laplace prior on β_g is chosen, that is

$$\pi(\beta_g|\lambda) = C_{d_g}\sqrt{\det(H_g)}\ \lambda^{d_g}\exp(-\lambda\|\beta_g\|_{H_g}), \tag{4.5}$$

where $C_{d_g} = 2^{-\frac{(d_g+1)}{2}}(2\pi)^{-\frac{(d_g-1)}{2}}/\Gamma((d_g+1)/2)$ and Γ is the gamma function. Then, using the same asymmetric Laplace distribution for the residuals u, the minimization in (4.4) is equivalent to the maximum of the posterior distribution

$$f(\beta|y,x,\lambda,\theta) \propto exp\left(-\sum_{i=1}^{n}\rho_\theta\,(y_i - h(x_i^T\beta)) - \lambda\sum_{g=1}^{G}\|\beta_g\|_{H_g}\right). \tag{4.6}$$

4.3.1 Gibbs sampling procedure

Similarly to the methods described in chapter 3, we can extend the Gibbs sampling procedure of Ji et al. (2012) to the case of tobit model and a group lasso penalty. As a first step we rewrite the prior of β_g using the equality (Andrews and Mallows, 1974)

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2}exp\left(-\frac{a}{2}s\right)ds,$$

which holds for any $a \geq 0$. In particular, we take $a = \lambda$ and $z = \|\beta_g\|_{H_g} = (\beta_g^T H_g \beta_g)^{\frac{1}{2}}$. Then the prior in (4.5) can be rewritten as (for more details see section 3.3 of the chapter 3)

$$\pi(\beta_g|\lambda) = \frac{(\frac{\lambda^2}{2})^{(d_g+1)/2}}{\Gamma(\frac{d_g+1}{2})}\int_0^\infty \frac{exp\left\{-\frac{1}{2}\beta_g^T(s_g H_g^{-1})^{-1}\beta_g\right\}s_g^{(d_g-1)/2}\exp\left(\frac{-\lambda^2}{2}s_g\right)ds_g}{\sqrt{\det(2\pi s_g H_g^{-1})}} \tag{4.7}$$

As a second step, we use the fact that an asymmetric Laplace distributed random variable can be written as a mixture of a $N(0;\ 1)$ distributed random variable and an

exponentially distributed random variable with rate parameter $\theta(1-\theta)$. This allows rewriting the likelihood (4.3) as:

$$f(y|x,\beta,\theta) \propto \sigma^n exp(-\sigma \sum_{i=1}^{n} \rho_\theta(u_i)) = \sigma^n exp\left(-\sigma \sum_{i=1}^{n} \frac{|u_i|+(2\theta-1)u_i}{2}\right)$$

$$= \prod_{i=1}^{n} \int_0^\infty \frac{\sigma}{\sqrt{4\pi\sigma^{-1}v_i}} exp\left(-\frac{(u_i-\xi v_i)^2}{4\sigma^{-1}v_i} - \zeta v_i\right) dv_i$$

with $u_i = y_i - h(x_i^T\beta)$, $\xi = (1-2\theta)$ and $\zeta = \sigma\theta(1-\theta)$.

Similar to the results in chapter 3, we can derive the following conditional distributions

$$f(\beta|y,x,v,\lambda,\theta) \propto exp\left\{-\sum_{i=1}^{n} \frac{(y_i - h(x_i^T\beta)-\xi v_i)^2}{4v_i} - \lambda \sum_{g=1}^{G} \|\beta_g\|_{H_g}\right\}$$

The full conditional distribution of β_g given $y, x, v, \beta_{-g}, \lambda$ and θ is:

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto f(y|\beta,x,v,\lambda,\theta)\pi(\beta_g|\lambda)$$

$$\propto exp\left\{-\sum_{i=1}^{n} \frac{(y_i - h(\sum_{k=1}^{G} x_{ik}^T\beta_k)-\xi v_i)^2}{4v_i} - \lambda\|\beta_g\|_{H_g}\right\}.$$

If we write $\tilde{y}_{ig} = y_i - h(\sum_{k=1,k\neq g}^{G} x_{ik}^T\beta_k) - \xi v_i$, then using (4.7), we can write the conditional distribution of β_g as

$$f(\beta_g|y,x,v,\beta_{-g},\lambda,\theta) \propto exp\left\{-\sum_{i=1}^{n} \frac{(\tilde{y}_{ig} - h(x_{ig}^T\beta_g)^2}{4v_i}\right\} exp\left\{-\frac{1}{2}\beta_g^T(s_g^{-1}H_g)\beta_g\right\}$$

Finally we put a Gamma prior on λ^2 ,$\lambda^2 \sim (\lambda^2)^{b_1-1}\exp(-b_2\lambda^2)$, where b_1 and $b_2 \geq 0$ are constants.

Similarly to chapter 3, the derivations above lead to the following Gibbs sampling procedure for the quantile θ:

1. Sample $y_i^\star$ from a truncated Normal distribution:

$$y_i^\star | y_i, x_i, \beta, v_i \sim \begin{cases} \delta(y_i), & if\ y_i > c, \\ N(x_i^T\beta + \xi v_i, 2v_i) I(y_i^\star \le c), & if\ y_i = c. \end{cases}$$

Where $\delta(y_i)$ denotes a degenerate distribution with all its mass at y_i. We use the sampling scheme as described in Ji et al. (2012) to generate the $y_i^\star$.

2. Sample v_i^{-1}, given $y_i^\star$, x_i and β, from an inverse Gaussian distribution with mean and shape parameters given by, respectively,

$$\mu = \sqrt{\frac{1}{(y_i^\star - x_i^T\beta)^2}} \text{ and } \eta = \frac{1}{2}$$

3. Sample s_g, given β_g and λ, from an inverse Gaussian distribution with mean and shape parameters given by, respectively,

$$\mu = \sqrt{\frac{\lambda^2}{\beta_g^T H_g \beta_g}} \text{ and } \eta = \lambda^2.$$

4. Sample β_g, given $y_i^\star, x_{-g}, \beta_{-g}, s_g, v$ from a multivariate normal distribution with mean and covariance given by

$$\mu_g = \Sigma_g x_g V\left(y_i^\star - (1-2\theta)v - x_{-g}^T\beta_{-g}\right) \text{ and } \Sigma_g = (x_g V x_g^T + s_g^{-1}H_g)^{-1},$$

respectively, where $V = diag\left(\frac{1}{2v_i}\right)$, $i = 1, \ldots, n$, and x_g is the $d_g \times n$ matrix of observations for group g.

5. Sample λ^2, given s_g, from a Gamma distribution with shape and rate parameters given by, respectively,

$$\alpha = \frac{p+G}{2} + b_1 \text{ and } \beta = \sum_{g=1}^{G} \frac{s_g}{2} + b_2$$

with b_1 and b_2 two non-negative constants which we set equal to 0.1.

4.4 Computing predicted values

In this section we will focus on computing the predicted values of the dependent variable y in a tobit model. Let's begin with a continuous variable $y_i^\star$ and the classical tobit model (Tobin, 1958),

$$y_i^\star = x_i^T\beta + u_i\,, i = 1,\ldots,n\,,\ \ u_i\,|x_i \sim N(0,\sigma^2).$$

So $E(y_i^\star|x_i) = x_i^T\beta$. Then the probability density function of $y_i^\star$ is $y_i^\star \sim N(x_i^T\beta,\sigma^2)$ and

$$y_i = \begin{cases} y_i^\star & if \quad y_i^\star > 0, \\ 0 & if \quad y_i^\star \le 0, \end{cases} \qquad \text{or equivalently}$$

$$y_i = \begin{cases} x_i^T\beta + u_i & if \quad x_i^T\beta + u_i > 0, \\ 0 & if \quad x_i^T\beta + u_i \le 0. \end{cases}$$

In order to make a prediction of y_i from the tobit model, we consider $E(y_i)$.

$$E(y_i) = P(y_i = 0) + P(y_i > 0) * E(y_i|y_i > 0)$$

$$= P(y_i^\star \le 0) + P(y_i^\star > 0) * E(y_i^\star|y_i^\star > 0)$$

$$= P(y_i^\star > 0) * E(y_i^\star|y_i^\star > 0)$$

$$= P(y_i^\star > 0) * E[x_i^T\beta + u_i\,|y_i^\star > 0]$$

$$= P(y_i^\star > 0) * [x_i^T\beta + E(u_i|u_i > -x_i^T\beta)\,].$$

If z follows a standard normal distribution with mean 0, and variance equal to 1, then $E(z|z > c) = \frac{\phi(c)}{1-\Phi(c)}$,where c is a constant, ϕ refers to the standard normal probability density, and Φ is the normal cumulative density. Using this result:

$$E(y_i) = P(y_i^\star > 0) * (x_i^T\beta + \sigma\frac{\phi(\frac{-x_i^T\beta}{\sigma})}{1-\Phi(\frac{-x_i^T\beta}{\sigma})})$$

$$= \Phi\left(\frac{x_i^T\beta}{\sigma}\right) * (x_i^T\beta + \sigma\frac{\phi\left(\frac{x_i^T\beta}{\sigma}\right)}{\Phi\left(\frac{x_i^T\beta}{\sigma}\right)})$$

$$= \Phi\left(\frac{x_i^T\beta}{\sigma}\right) * (x_i^T\beta + \sigma\lambda(\frac{x_i^T\beta}{\sigma})),$$

where the function λ is defined as $\lambda(z) = \frac{\phi(z)}{\Phi(z)}$, and is generally referred to as inverse Mills ratio function.

We will now discuss the prediction of y_i under a tobit quantile regression model. We will therefore need to compute $E(u_i|u_i > -x_i^T\beta)$ where u_i follows an ALD with mean zero.

$\Phi_{ALD}(z) \equiv \int_{-\infty}^{z} \phi_{ALD}(u)du$, where

$$\phi_{ALD}(u) = \frac{\theta(1-\theta)}{\sigma}\begin{cases} exp\left(\frac{1-\theta}{\sigma}u\right), & if\ u \leq 0 \\ exp\left(-\frac{\theta}{\sigma}\ u\right), & if\ u > 0\ . \end{cases}$$

So Φ_{ALD} is the cumulative density function (CDF) and, ϕ_{ALD} is the density function of an asymmetric Laplace distributed random variable. In order to compute $E(u|u > c)$ we will distinguish the case of $c \leq 0$ and $c > 0$.

If $c \leq 0$, then

$E(u|u > c) = \frac{1}{1-\Phi_{ALD}(c)}[\int_c^0 u\ \phi_{ALD}(u)du + \int_0^\infty u\ \phi_{ALD}(u)du]$.

Using the definitions above of Φ_{ALD} andϕ_{ALD},

$\int_c^0 u\ \phi_{ALD}(u)du = \frac{\theta(1-\theta)}{\sigma}\int_c^0 u\ exp\left(\frac{1-\theta}{\sigma}u\right)du$

$$= \frac{\theta(1-\theta)}{\sigma}\left[\frac{e^{\frac{1-\theta}{\sigma}u}}{\frac{(1-\theta)^2}{\sigma^2}}\left(\frac{1-\theta}{\sigma}u-1\right)\right]_c^0$$

$$= \left[-\frac{\sigma\,\theta}{(1-\theta)}-\frac{\sigma\theta\, e^{\frac{1-\theta}{\sigma}c}}{(1-\theta)}\left(\frac{1-\theta}{\sigma}c-1\right)\right]$$

and $\int_0^\infty u\,\phi_{ALD}(u)du = \int_0^\infty u\,\phi_{ALD}(u)du$

$$= \frac{\theta(1-\theta)}{\sigma}\int_0^\infty u\ exp\left(-\frac{\theta}{\sigma}u\right)du$$

$$= \frac{\sigma(1-\theta)}{\theta}.$$

Therefore:

$$E(u|u>c) = \frac{1}{1-\Phi_{ALD}(c)}\left[\int_c^0 u\,\phi_{ALD}(u)du + \int_0^\infty u\,\phi_{ALD}(u)du\right]$$

$$= \frac{1}{1-\Phi_{ALD}(c)}\left[-\frac{\sigma\,\theta}{(1-\theta)}-\frac{\sigma\theta\, e^{\frac{1-\theta}{\sigma}c}}{(1-\theta)}\left(\frac{1-\theta}{\sigma}c-1\right)+\frac{\sigma(1-\theta)}{\theta}\right]$$

$$= \frac{1}{1-\Phi_{ALD}(c)}\left[\frac{\sigma(1-2\theta)}{\theta(1-\theta)}-\frac{\theta\sigma e^{\frac{1-\theta}{\sigma}c}}{(1-\theta)}\left(\frac{1-\theta}{\sigma}c-1\right)\right].$$

If $c > 0$

$$\int_c^\infty u\,\phi_{ALD}(u)du = \frac{\theta(1-\theta)}{\sigma}\int_c^\infty u\ exp\left(-\frac{\theta}{\sigma}u\right)du$$

$$= \sigma\frac{(1-\theta)}{\theta}\left[e^{-\frac{\theta}{\sigma}c}\left(\frac{\theta}{\sigma}c+1\right)\right].$$

This gives

$$E(u|u>c) = \begin{cases} \frac{1}{1-\Phi_{ALD}(c)}\left[\frac{\sigma(1-2\theta)}{\theta(1-\theta)}-\frac{\theta\sigma e^{\frac{1-\theta}{\sigma}c}}{(1-\theta)}\left(\frac{1-\theta}{\sigma}c-1\right)\right], & if\ c\le 0 \\ \frac{1}{1-\Phi_{ALD}(c)}\cdot\sigma\frac{(1-\theta)}{\theta}\left[e^{-\frac{\theta}{\sigma}c}\left(\frac{\theta}{\sigma}c+1\right)\right], & if\ c>0 \end{cases}.$$

This derivation allows us to compute $E(u_i|u_i > -x_i^T\beta)$. In particular:

If $c = -x_i^T\beta_\theta \leq 0$, then

$$E(y_i) = P(y_i^\star > 0) * [-c + \frac{1}{1-\Phi_{ALD}(c)} [\frac{\sigma(1-2\theta)}{\theta(1-\theta)} - \frac{\theta\sigma e^{\frac{1-\theta}{\sigma}(c)}}{(1-\theta)} (\frac{1-\theta}{\sigma}(c) - 1)]$$

$$= \left(1 - \Phi_{\text{ALD}}(c)\right) * [-c + \frac{1}{1-\Phi_{ALD}(c)} [\frac{\sigma(1-2\theta)}{\theta(1-\theta)} - \frac{\theta\sigma e^{\frac{1-\theta}{\sigma}(c)}}{(1-\theta)} \left(\frac{1-\theta}{\sigma}(c) - 1\right)].$$

If $c > 0$

$$E(y_i) = \left(1 - \Phi_{\text{ALD}}(c)\right) * [-c + \frac{1}{1-\Phi_{ALD}(c)} . \sigma \frac{(1-\theta)}{\theta} [e^{-\frac{\theta}{\sigma}c} \left(\frac{\theta}{\sigma}(-c) + 1\right)]].$$

It is clear from the last formula, how the final predic value is related the parameter estimate, predictor values, standard deviation σ of the parameter estimate and the value of θ .

4.5 Simulation study

In this section, we investigate the performance of our method with a simulation study. We compare our method, Bayesian tobit quantile regression with group Lasso penalty here denoted by $BTQ.grplasso$, with Bayesian tobit quantile regression with an adaptive Lasso penalty, here denoted by $BTQ.adalasso$ (Alhamzawi, 2013). For simulating the data we consider the model

$$y_i^\star = x_i^T\beta_\theta + u_i, i = 1,. \ . \ ., n \text{ and } y_i = h(y_i^\star),$$

with β chosen to have a group structure and with different choices of the error distribution. In particular, similar to (Yu et al., 2013), we consider the following cases for the error:

■Normal: $N(0;\ 1)$

■Kurtotic: (kur): $\frac{2}{3}N(0,1) + \frac{1}{3}N(0, (\frac{1}{10})^2)$.

For the β vector and its group structure, in the first case, we consider the case of a small number of predictors. Similar to the simulation in chapter 3, we simulate 5 groups, each consisting of 3 covariates. The 15 variables are assumed to follow a multivariate normal distribution $N(0, \Sigma)$, with Σ having a diagonal block structure. Each block corresponds to one group and is defined by the matrix $r^{|i-k|}$, $i = i = 1,. \ . \ .,3,\ k = 1,. \ . \ .,3$. For the correlation r, we experiment both with $r = 0.95$ (well-defined group structure) and $r = 0.5$. The β values for five groups are given by

$(0.5, 1, 1.5), (2,2.5, 2), (2, 2, 2), (2,2,2\ ), (0,0,0).$

In the second case, we consider the case of a large number of predictors having a group structure. We simulate 10 groups, each consisting of 10 covariates. The 100 variables are assumed to follow a multivariate normal distribution $N(0;\ \Sigma)$, with Σ having a diagonal block structure. Each block corresponds to one group and is defined by the matrix $r^{|i-k|}$, $i = i = 1,. \ . \ .,10, k = 1,. \ . \ .,10$. For the correlation r,we experiment both with $r = 0.95$ (well-defined group structure) and $r = 0.5$. The β values for the first three groups are given by

$(0.5, 1, 1.5, 2,2.5, 2, 2, 2, 2, 2), (2, 2, 1, 1, 1,1, 3, 3, 3, 3), (1,1, 1, 2, 2, 2, 3, 3, 3, 3),$

and they are set to zero for all other groups.

For each case, we use the Gibbs sampling procedure, using 17000 iterations with the first 1000 iterations as burn-in. We only report the results for $\theta = 0.5$ (median) quantile. Similar results are obtained for other quantiles. Figures 4.1, 4.2 and 4.3 report

the relative bias, variance of the estimated parameters and median error respectively, averaged over 100 iterations for $BTQ.grplasso$ and $BTQ.adalasso$.

The relative average bias of an estimated coefficient is defined by

Bias $(\hat{\beta}_j) = \frac{1}{100}\sum_{i=1}^{100}(\hat{\beta}_j^i - \beta_j)$,

where $\hat{\beta}_j^i$ is the tobit quantile regression coefficient estimate for the ith repetition and β_j is the true value of the jth coefficient. The variance of the parameter estimate is computed by $V(\hat{\beta}_j) = \frac{1}{99}\sum_{i=1}^{100}(\hat{\beta}_j^i - \bar{\beta}_j)^2$ where $\bar{\beta}_j = \frac{1}{100}\sum_{i=1}^{100}(\hat{\beta}_j^i)$. The median error also computed for the model similar in the chapter 2.

From result in Figure 4.1, our simulation study confirms that the performances of $BTQ.grplasso$ and BTQ.adalasso are similar in case normality and when the predictors are low correlated. Furthermore, the results show how BTQ.adalasso is worst performing method in case of departure from normality especially when the predictors are highly correlated.

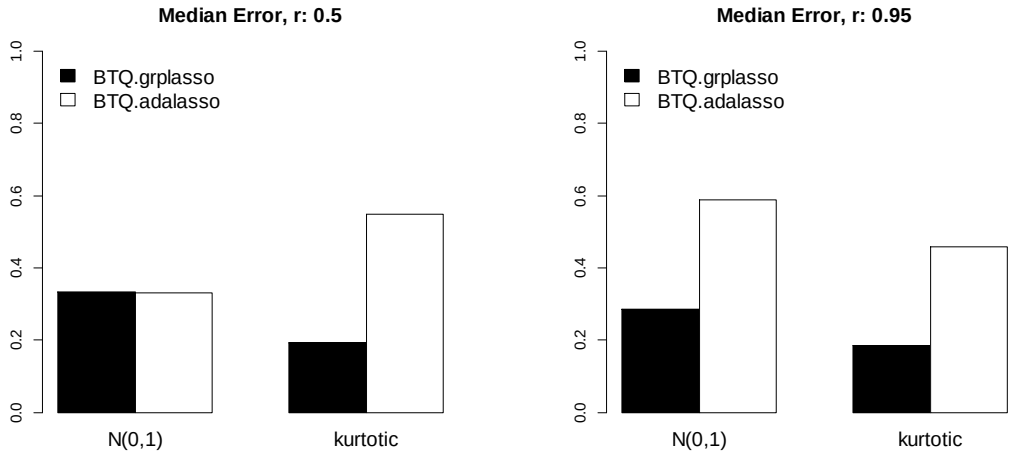

Figure 4.1: Comparison of Bayesian tobit quantile regression with group lasso ($BTQ.grplasso$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso$)under normal and kurtotic error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for the simulation study when $p = 15$, n=100 and $\theta = 0.5$.

Figures 4.2 and 4.3 clearly show that $BTQ.grplasso$ behave better than BTQ.adalasso on simulated data especially when we have high correlation within the groups. The results indicate that assuming the group structure for the predictors improves the parameter estimates of the tobit regression model in terms of their and variance.

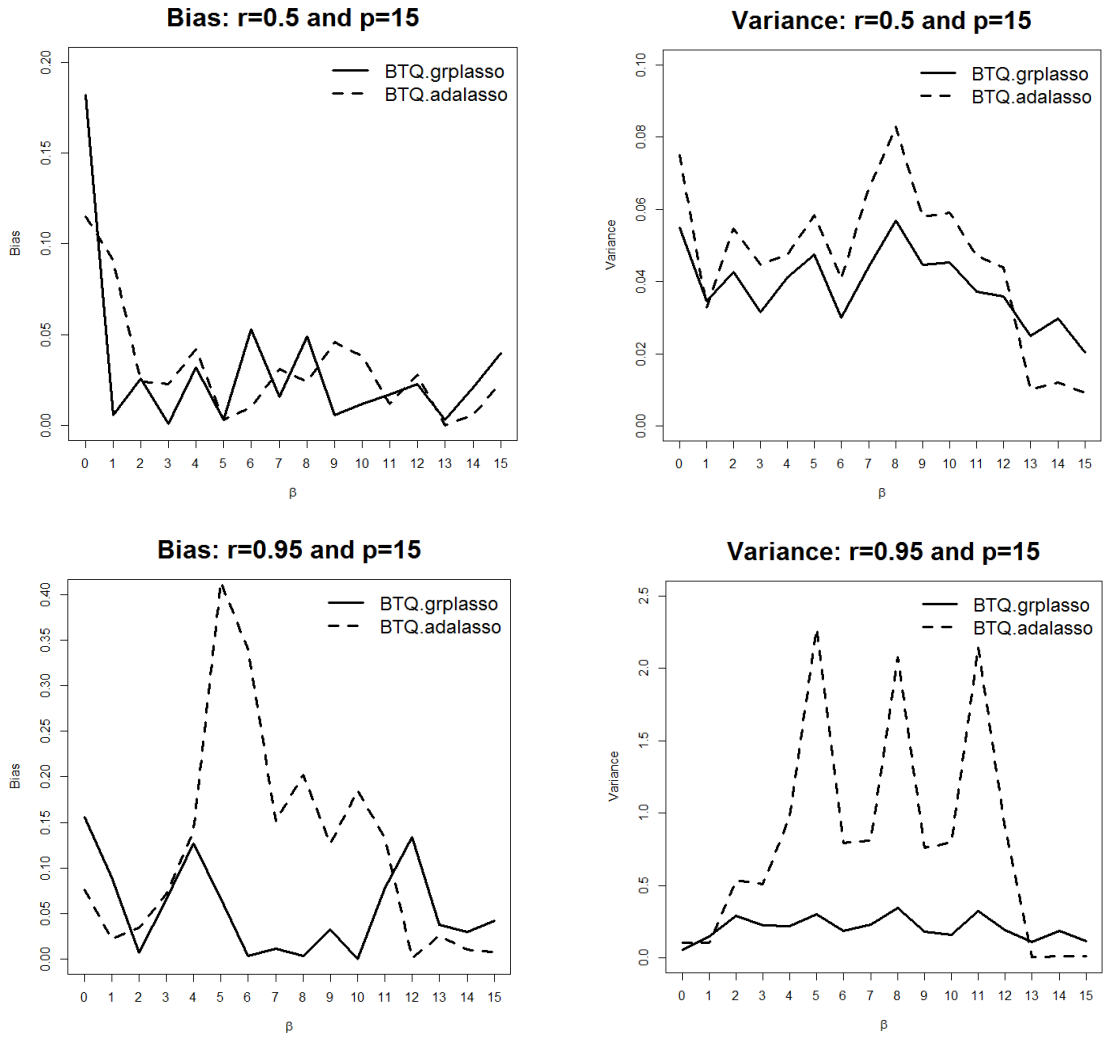

Figure 4.2: Bias and variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a normal distribution for the error, $p = 15$, n=100 and $\theta = 0.5$.

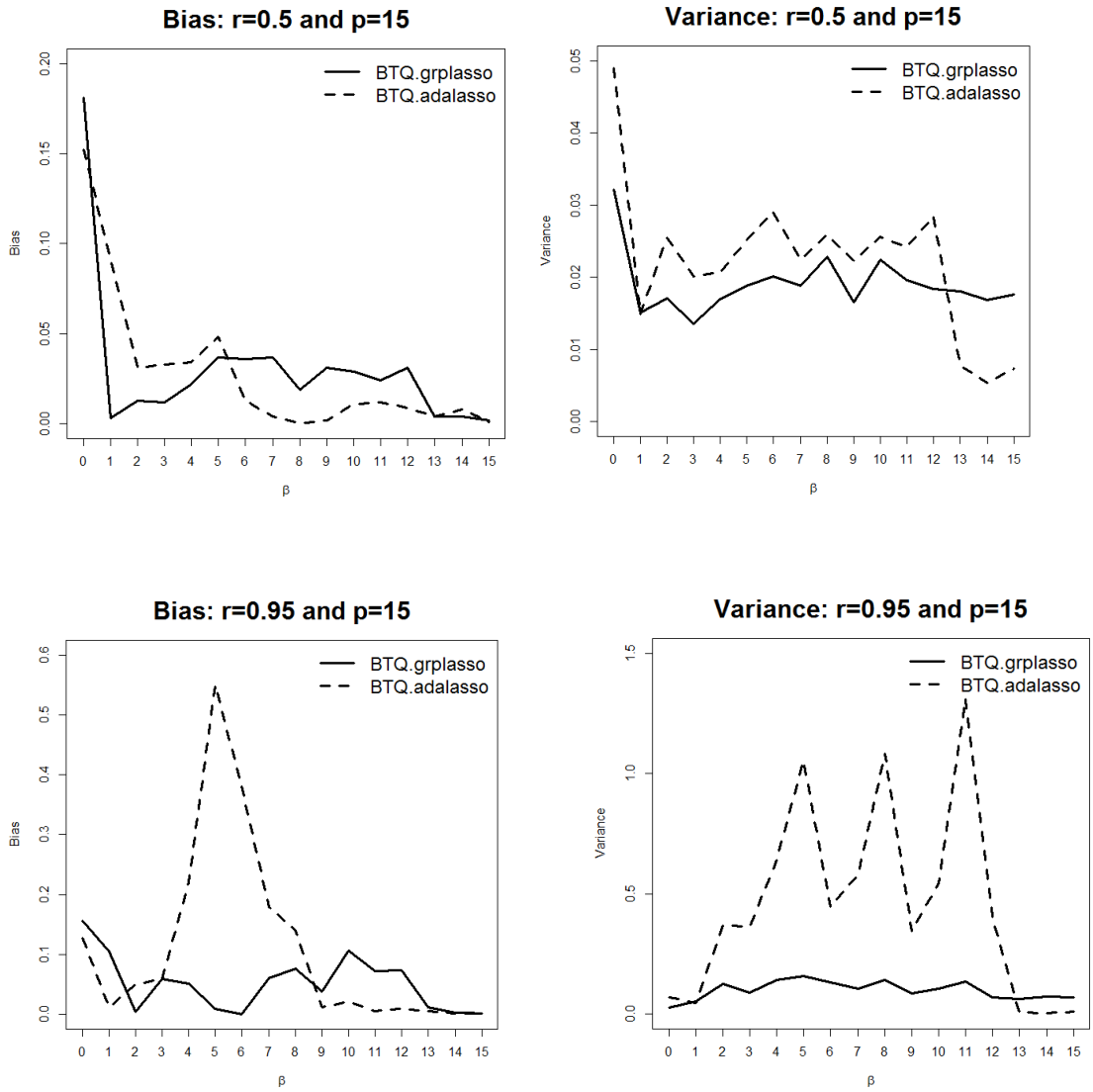

Figure 4.3: Bias and variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso (BTQ. grplasso, solid line) and Bayesian tobit quantile regression with an adaptive lasso penalty (BTQ. adalasso, dashed line) under a Kurtotic distribution for the error, $p = 15$, *n=100 and* $\theta = 0.5$.

In the next figures (4.4 - 4.12), we consider the case of a large number of predictors ($p = 100$) having a group structure. We only report the results for $\beta_0, \ldots, \beta_{50}$. Similar results are obtained for the other $\beta's$. The figures clearly show that $BTQ.grplasso$ behave better than $BTQ.adalasso$, especially when we have high correlation within the groups, in terms of the median error bias and variance of the estimated parameters.

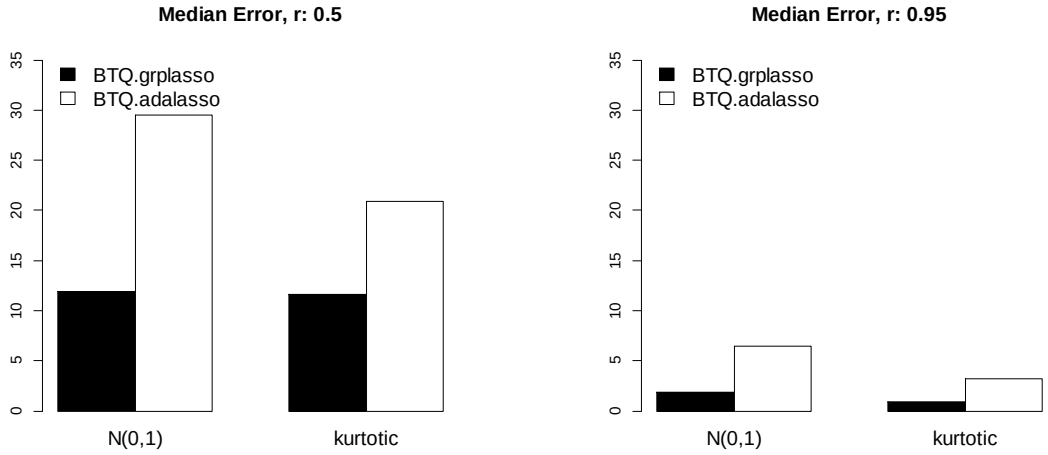

Figure 4.4: Comparison of Bayesian tobit quantile regression with group lasso ($BTQ.grplasso$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso$)under normal and kurtotic error distributions, for low (left) and high (right) correlated predictors. The plot shows the median model error over 40 replications for the simulation study when $p = 100$, *n=100 and* $\theta = 0.5$.

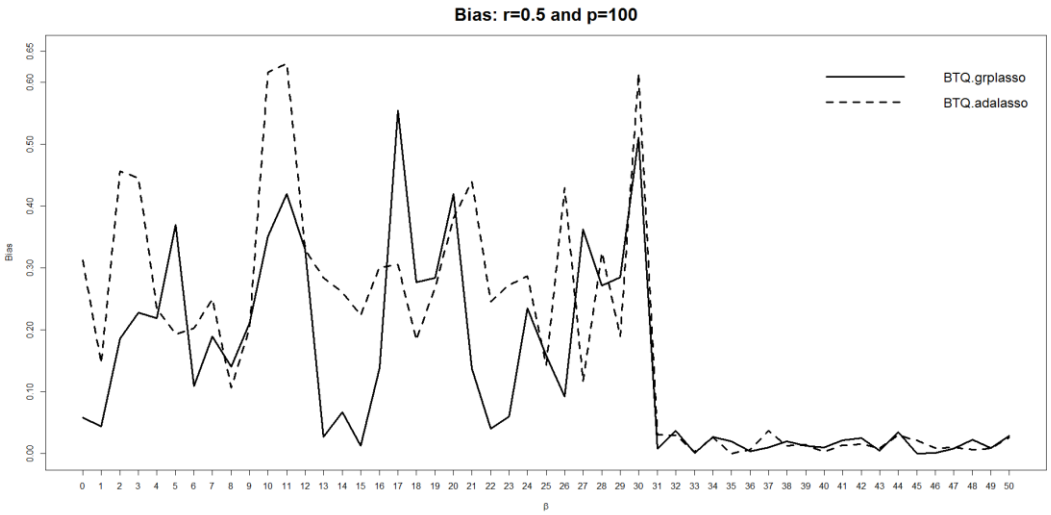

Figure 4.5: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a normal distribution for the error, $r = 0.5,\ p = 100$, *n=100 and* $\theta = 0.5$.

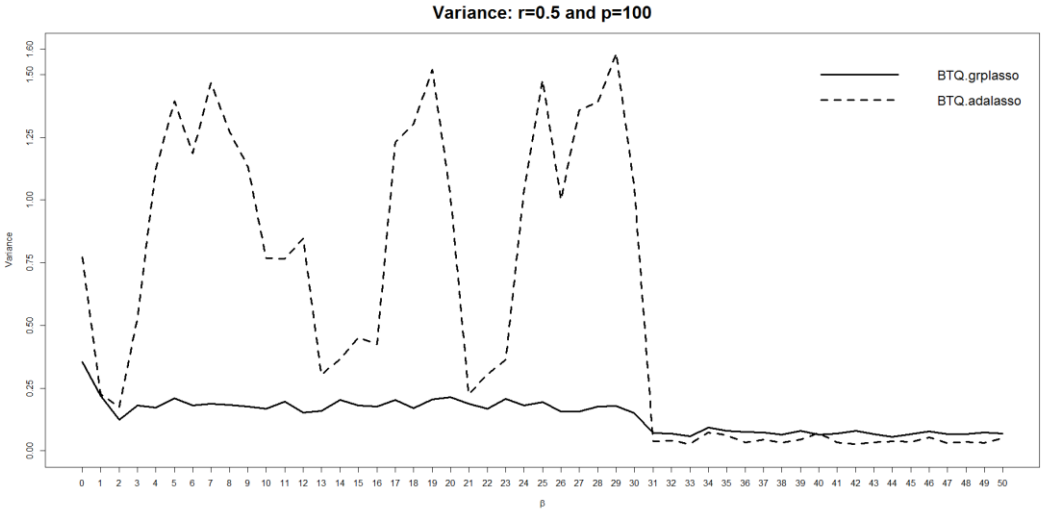

Figure 4.6: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a normal distribution for the error, $r= 0.5, = 100$ *, n=100 and* $\theta = 0.5$.

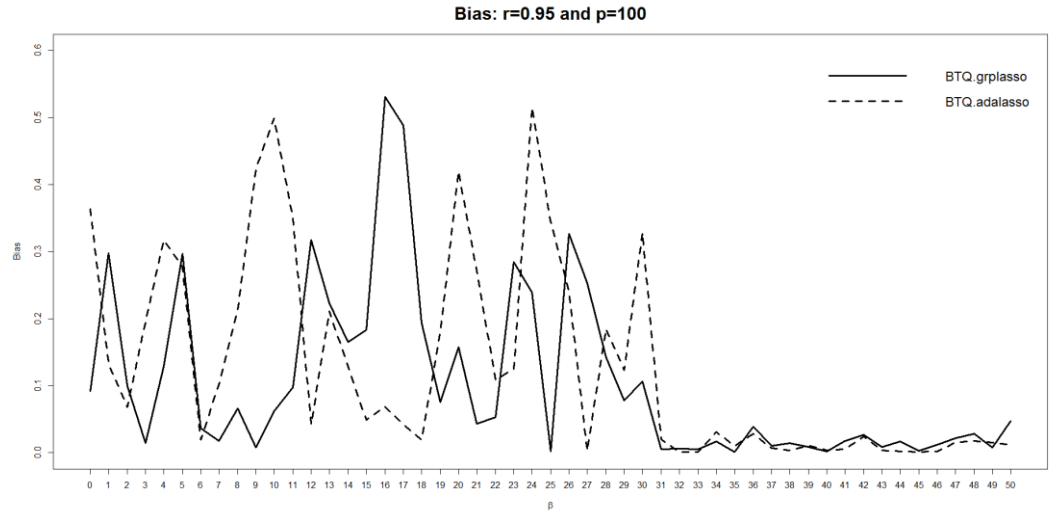

Figure 4.7: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a normal distribution for the error, $r = 0.95, p = 100$ *, n=100 and* $\theta = 0.5$.

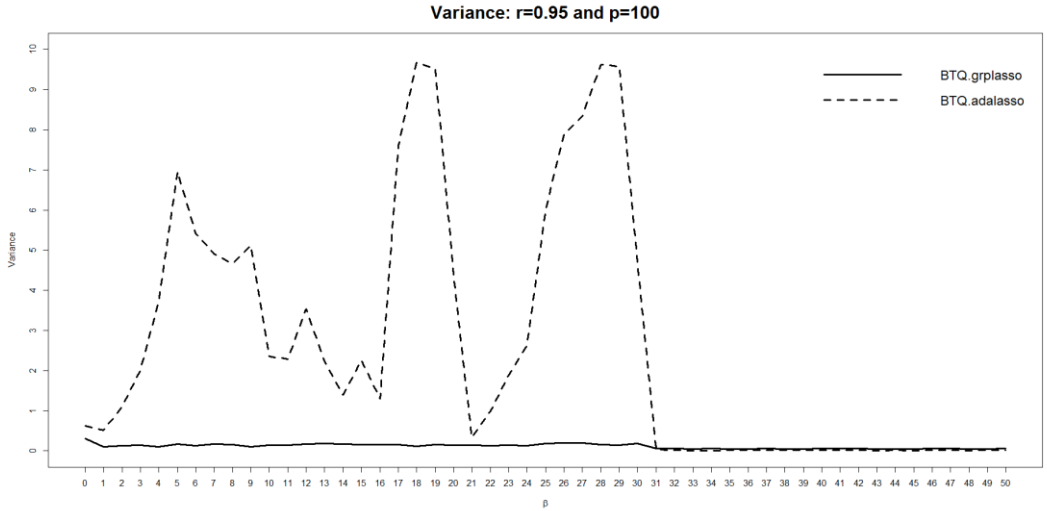

Figure 4.8: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a normal distribution for the error, $\mathrm{r} = 0.95$ *,*$p = 100$ *, n=100 and* $\theta = 0.5$.

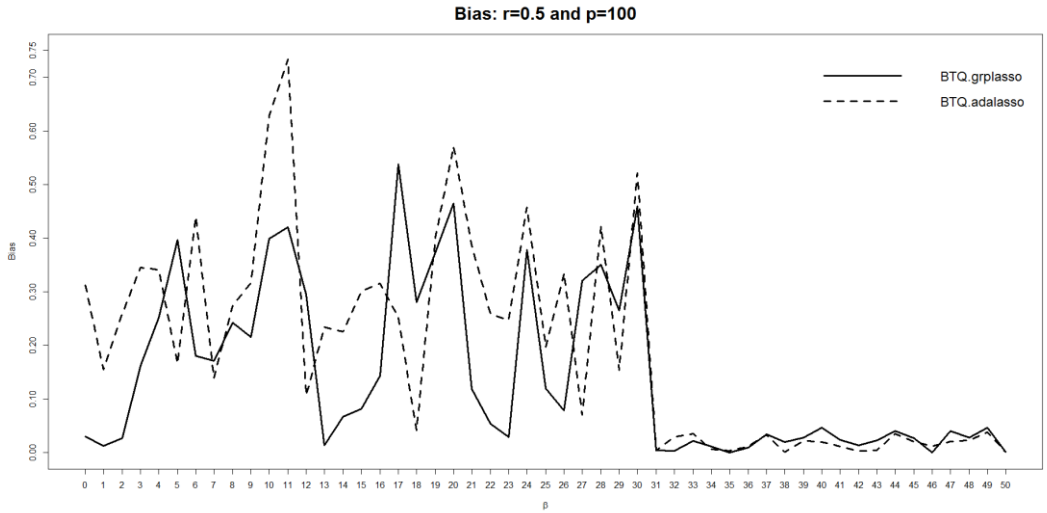

Figure 4.9: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a Kurtotic distribution for the error, $r = 0.5$ *,* $p = 100$ *, n=100 and* $\theta = 0.5$.

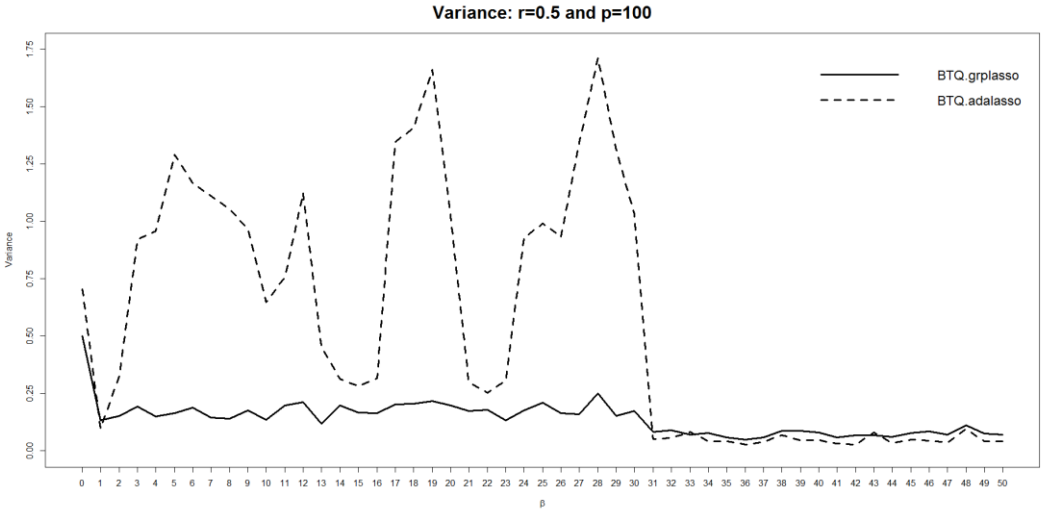

Figure 4.10: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a Kurtotic distribution for the error, $r = 0.5$, $p = 100$, n=100 and $\theta = 0.5$.

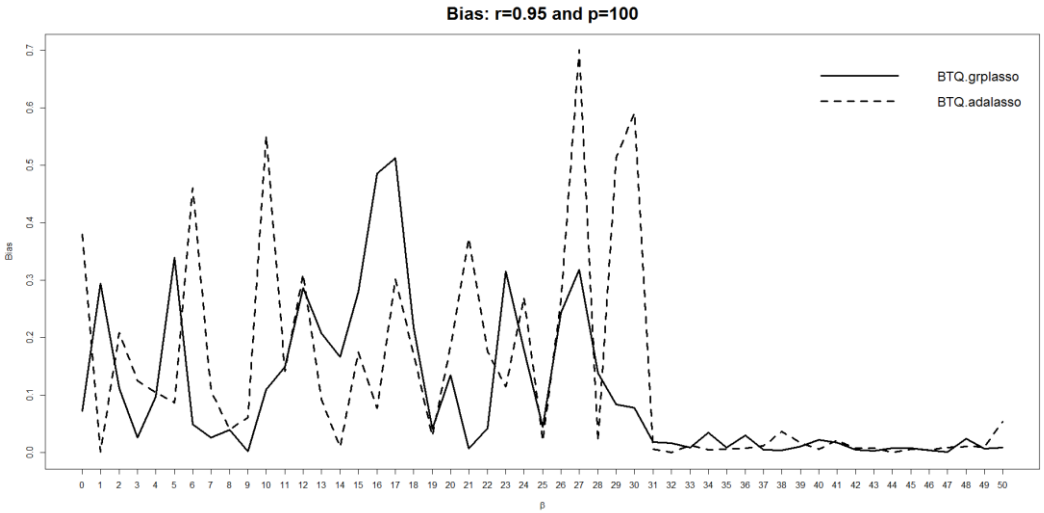

Figure 4.11: Bias (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a Kurtotic distribution for the error, $r = 0.95, p = 100$, n=100 and $\theta = 0.5$.

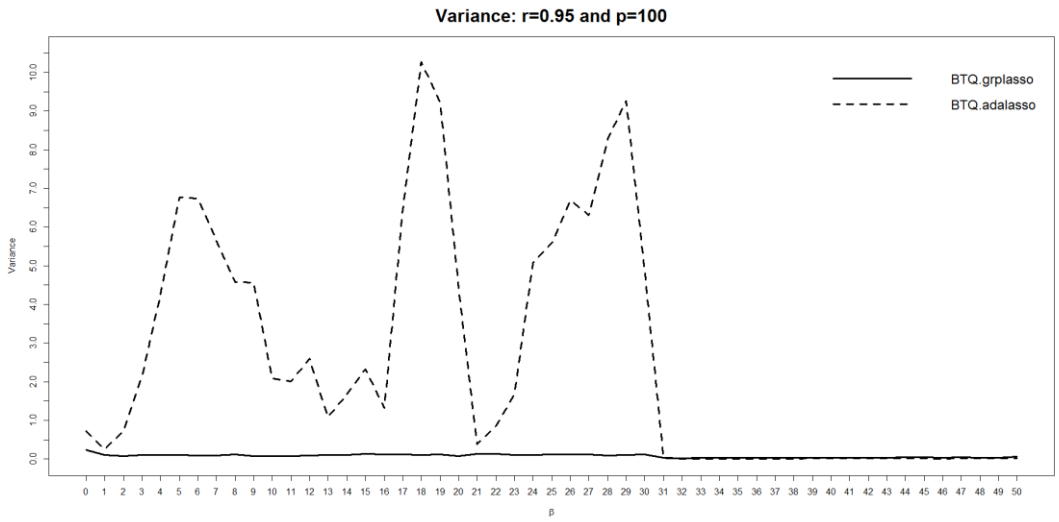

Figure 4.12: Variance (averaged over 100 replications) of the regression coefficients for Bayesian tobit quantile regression with group lasso ($BTQ.grplasso, solid\ line$) and Bayesian tobit quantile regression with an adaptive lasso penalty ($BTQ.adalasso, dashed\ line$) under a Kurtotic distribution for the error, $r=0.95$, $p=100$, $n=100$ and $\theta=0.5$.

4.6 Real application

In this section, we used our method to estimate the parameters of the labor force participation dataset described in chapter 3**.** Now because the wife's annual hours of work, which is a positive variable, we consider as response variable. The aim of this analysis is to assess if there is a relation between several social factors (wife's age, husband's wage and wife's father education etc.) and wife's annual hours of work, how strong this relation is and what the influence of the factors are. The grouping structure of the predictors is given in Table 3.5. We apply our method ($BTQ.grplasso$) and $BTQ.adalasso$ method to this dataset. Since Bayesian regularized methods do not give exact zero, we consider credible intervals to select which parameters are different from zero. The Bayesian estimates are obtained based on 17000 MCMC iterations with 1000 burn-in. Tables 4.2 and 4.3 show the fitted coefficients for the $0.5th$ quantile and $0.95th$ quantile, respectively along with their 99% credible intervals for $BTQ.grplasso$and $BTQ.adalasso$.We compare the results in Tables 4.2 and 4.3 with

the results of mean tobit regression with group lasso penalty ($M.grplasso$) which is reported by Liu et al. (2013).Also the running times of the code are reported in the Tables 4.2.

The results in Table 4.2 and 4.3 show a similar performance of the three methods on the labor dataset, particularly for $BTQ.grplasso$ and $BTQ.adalasso$. For example, we can see groups G4, G5 and G7 are not selected by all three methods. Furthermore, the advantage of group lasso is to carry out group selection, meaning that within a group, coefficients will either all be zero or all nonzero, therefore we do not set this property in some cases such as in group G1, where notice all elements in group G1 are selected by the $M.grplasso$ method while only 4 elements are selected by $BTQ.grplasso$ and $BTQ.adalasso$ methods. Again, this might be due to the regression having high levels of sparsity, which gives Bayesian methods a disadvantage. This issue will be explored further in future research.

The results in Table 4.1 and Table 4.2 are found by implemented R functions. We consider the data set in the real data in section 4.6 with $p = 18$ and $n = 753$. We use a computer with 2.4 GHz processor and 6 gigabytes of RAM. Computation values of credible intervals for labor force participation dataset in Table 3.1 for one quantile takes 75 minutes for $BTQ.grplasso$, 60 minutes for $BTQ.adalasso$ and 50 minutes for $M.grplasso$. According to the results in Table 4.1 and Table 4.2 we conclude that the time-consuming of the proposed method is not much larger than the other methods.

Table 4.1: 99% *credible intervals for labor force participation dataset at* $\theta = 0.5$.

	Methods	$BTQ.grplasso$		$BTQ.adalasso$		$M.grplasso$
	Variables name	Lower 0.5%	Upper 99.5%	Lower 0.5%	Upper 99.5%	Mean
Group name	Intercept	0.161	0.402	0.000	0.000	0.727
G1	WE	0.000	0.000	0.000	0.000	−0.082
	WW	0.040	0.134	0.042	0.211	0.393
	RPWG	0.186	0.302	0.166	0.365	0.460
	FAMINC	0.000	0.000	0.000	0.000	−0.038
	MTR	-9.935	-2.920	-13.146	-1.086	−0.084
	AX	0.030	0.061	0.027	0.080	0.179
G2	KL6	-0.952	-0.333	-1.438	-0.300	0.000
	K618	0.000	0.000	0.000	0.000	0.000
G3	HE	0.000	0.000	0.000	0.000	0.000
	HW	-0.213	-0.081	-0.271	-0.059	0.000
G4	WMED	0.000	0.000	0.000	0.000	0.000
	WFFD	0.000	0.000	0.000	0.000	0.000
G5	UN	0.000	0.000	0.000	0.000	0.000
	CIT	0.000	0.000	0.000	0.000	0.000
G6	WA	-0.060	-0.004	0.000	0.000	0.000
	HA	0.000	0.000	0.000	0.000	0.000
G7	HHRS	0.000	0.000	0.000	0.000	0.000
Computational time (minutes)		75		60		50

Table 4.2: 99% *credible intervals for labor force participation dataset at* $\theta = 0.95$.

	Methods	$BTQ.grplasso$		$BTQ.adalasso$		$M.grplasso$
	Variables name	Lower 0.5%	Upper 99.5%	Lower 0.5%	Upper 99.5%	Mean
Group name	Intercept	0.197	0.376	-0.079	0.228	0.727
G1	WE	-0.113	0.085	-0.180	0.146	−0.082
	WW	-0.073	0.103	-0.087	0.251	0.393
	RPWG	-0.019	0.194	-0.070	0.288	0.460
	FAMINC	0.000	0.000	0.000	0.000	−0.038
	MTR	-11.449	-0.752	-13.278	3.974	−0.084
	AX	0.003	0.066	-0.013	0.095	0.179
G2	KL6	-0.644	0.241	-0.956	0.543	0.000
	K618	-0.159	0.118	-0.275	0.254	0.000
G3	HE	-0.058	0.086	-0.111	0.121	0.000
	HW	-0.244	-0.092	-0.256	0.009	0.000
G4	WMED	-0.053	0.056	-0.097	0.109	0.000
	WFFD	-0.062	0.041	-0.105	0.079	0.000
G5	UN	-0.079	0.024	-0.121	0.065	0.000
	CIT	-0.461	0.287	-0.759	0.556	0.000
G6	WA	-0.067	0.016	-0.088	0.044	0.000
	HA	-0.041	0.035	-0.070	0.058	0.000
G7	HHRS	-0.001	0.000	-0.001	0.000	0.000
Computational time (minutes)		75		60		50

4.7 Conclusion

The goal of this chapter is to propose a method for tobit regression problems where the predictors have a natural group structure, such as in the case of categorical variables. In contrast to existing methods for group-typed variables, we model the quantiles of the response variable, in order to account for possible departures from normality in the latent variable. This motivates the use of quantile-based regression for tobit regression models.

We compare the method with Bayesian tobit quantile regression with adaptive lasso penalty on simulated data and real data and with mean tobit regression with group lasso penalty in real data. The simulation study shows that our method behave better than $BTQ.adalasso$ especially when we have high correlation within the groups in terms of the bias and variance of the parameter estimates. The real data shows similar performance of the three methods in most cases, in terms of group selection and parameter estimates.

Chapter 5

Conclusions and Future Research

The work in this thesis focuses on variable selection for high-dimensional data by using a combination of regularized and robust regression methods. The major contributions of the thesis and possible future research are summarised as follows.

5.1 Main Contributions

In Chapter 2, we focus on the regularized and robust regression methods for continuous response variable. We give an overview, state of related research and present a comparative simulation study for different regularized and robust regression methods when the response variable is continuous under different error distributions. Moreover, the chapter present some concluding remarks concerning to these methods. This chapter aims to help researchers to choose the correct model when their data could be contaminated with outliers.

In Chapter 3, we focus on the regularized and robust regression methods for binary response variable. A group lasso penalty for binary quantile regression models is developed. The error distribution is assumed to be an Asymmetric Laplace Distribution (ALD). We have presented a Bayesian approach for binary quantile regression combined with group lasso as a variable selection technique. The main advantages of this method are: firstly, performs estimation and variable selection simultaneously, and the procedure is robust when the data are subject to some form of contamination or outliers, secondly, the procedure can select important predictors for the different quantile of the response variable. The performance of the proposed methods was shown on both simulated and real data in comparisons with other existing methods.

In Chapter 4, we develop a regularized and robust regression method for a censored response variable under a group lasso penalty. We have presented a Bayesian approach for the estimation of parameters. The performance of the proposed methods was evaluated on simulated data in terms the bias and variance.

5.2 Recommendations for Future Research

1. In this thesis we studied the regularized and robust regression methods in the situation when the heavy-tailed errors or outliers are found in the responses or vertical outliers. We plan to study regularized and robust regression methods in the situation when the heavy-tailed errors or outliers are found in both the explanatory variables and responses or when we have bad leverage observations that are both outlying in the explanatory variables and located far from the true regression line.
2. The work presented in Chapter 3 and 4 motivate us to recommend a number of interesting future work recommendations. Two of these are:
    a. To study the Bayesian binary quantile regression with adaptive lasso penalty or with other group variable selection penalties. For examples, group MCP , group SCAD and group Bridge (Huang et al., 2012)
    b. To study the Bayesian censored quantile regression with other group variable selection penalties. For examples, group MCP , group SCAD and group Bridge (Huang et al., 2012)
3. Due to disadvantage of performs Bayesian regularized regression methods in case high dimensional sparse, we recommend using the idea of conjugate priors Bayesian quantile regression method (Alhamzawi and Yu, 2013) to other models such as Bayesian adaptive lasso, Bayesian elastic net and Bayesian group lasso .

Bibliography

Arslan, O. (2012). Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression. *Computational Statistics and Data Analysis 56*, 1952–1965.

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

Alhamzawi, R., K. Yu, and D. Benoit (2012). Bayesian adaptive lasso quantile regression. *Statistical Modelling 12* (3), 279 – 297.

Alhamzawi, R. (2013). Tobit Quantile Regression with the adaptive Lasso penalty. *The 4th International Scientist*, 11/2013.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *The 2nd International Symposium on Information Theory*, Akademia Kiad´o, Budapest, pp. 267–281.

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

Bai, X. (2010). Robust Mixtures of Regression Models. Thesis submitted to the department of statistics and college of arts and sciences of Kansas State University.

Bradic, J. and J. Fan (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society, B 73* (3), 325–349.

Bach, F. (2008). Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research 9*, 1179–1225.

Barros, M., M. Galea, M. Gonzalez, and V. Leiva (2010). Influence diagnostics in the tobit censored response model. *Statistical Methods and Applications 19*, 379–397.

Bae, K. and B. Mallick (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics 20* (18), 3423–3430.

Belloni, A. and V. Chernozhukov (2011). Post L_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics 39*, 82–130.

Benoit, D. and D. Poel (2012). Binary quantile regression: a Bayesian approach based on the asymmetric Laplace density. *Journal of Applied Econometrics 27* (7), 1174–1188.

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

Breheny, P. and J. Huang (2014) Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing 24*, (99), 1573-1375.

Casella, G. (2001). Empirical Bayes Gibbs Sampling. *Biostatistics 2*, 485–500.

Craven, P. and G. Wahba (1979). Smoothing noisy data with spline functions. *Numerische Mathematik 31*, 377–403.

Chhikara, R.S. and L. Folks (1989). The Inverse Gaussian distribution: Theory, Methodology, and Applications. Marcel Dekker, New York.

Draper, N.R. and H. Smith (1998). Applied Regression Analysis (Third Edition). New York: Wiley.

Efron, B.,T. Hastie, I. Johnstone, and R., Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–451.

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

Flexeder, C. (2010). Generalized Lasso Regularization for Regression Models. A Ph.D. thesis submitted to the University of Munchen.

Frank, I. E. and J. Friedman (1993). A Statistical view of some chemometrics regression tools, *Technometrics 35*, 109–148.

Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso. *Journal of Computational and Graphical Statistics 7*, 397–416.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 31*(1), 1–22.

Genkin, A., D. D. Lewis and D. Madigan (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics 49* (14), 291–304.

Gene, Y. and C. Burge (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, *Journal of Computational Biology 11*, 475 - 494.

Gramacy, R. and N. Polson (2012). Simulation-based regularized logistic regression. *Bayesian Analysis 7*(3):503–770.

Hand, D. and V. Vinciotti (2003). Local versus global models for classification problems: fitting models where it matters. *The American Statistician 57* (2), 124–131.

Hastie, T. Tibshirani R., and Friedman J.H. (2009). Elements of Statistical Learning. Springer-Verlag :New York.

Henningsen A. (2012). Estimating Censored Regression Models in R using the censReg Package. http://cran.at.r-project.org/web/packages/censReg/vignettes/censReg.pdf.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics , 12* (1), 55–67.

Huang, J. and T. Zhang (2010). The benefit of group sparsity. *The Annals of Statistics 38* (4), 1978–2004.

Huang, J., P. Breheny, and S. Ma (2012). A selective review of group selection in high-dimensional models. *Statistical Science 27*, 481-499.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 35*(1), 73-101.

Huber, P. (1981). Robust Statistics. New York: John Wiley and Sons.

Hubert, M. and Rousseeuw, P. J. (1997). Robust regression with both continuous and binary regressors, *Journal of Statistical Planning and Inference. 57*, 153–163.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science 27*, 481-499.

Ji, Y., N. Lin, and B. Zhang (2012). Model selection in binary and tobit quantile regression using the gibbs sampler. *Computational Statistics and Data Analysis 56* (4), 827 – 839.

Koenker, R. and G. W. Bassett (1978). Regression quantiles. *Economerrica 46*, 33–50.

Koenker, R. (2004). Quantile regression for longitudinal data. *J. Multivar. Anal.* 91,

74–89.

Kordas, G. (2002). Credit scoring using binary quantile regression. *Statistics for Industry and Technology*, 125–137.

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

Koren,Y., Bell,R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems, *Computer 42* (8), 30-37.

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

Krishnapuram, B., L. Carin, M. A. Figueiredo, and A. J. Hartemink (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 957–968.

Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electronic Journal of Statistics 5*, 1015–1053.

Lazaridis,C.D.(2008). L^q -Norm Shrinkage Regressionand Application. A thesis submitted to the University of Melbourne.

Li, Q., R. Xi, and N. Lin (2010). Bayesian regularized quantile regression. *Bayesian Analysis 5*, 1–24.

Liu, X., Wang, Z. and Wu, Y. (2013). Group variable selection and estimation in the tobit censored response model. *Computational Statistics & Data Analysis 60*, 80-89.

Li, Y. and J. Zhu (2008). L_1-norm quantile regressions. *Journal of Computational and Graphical Statistics 17*, 163–185.

Lounici, K., M. Pontil, A. Tsybakov and S. van de Geer (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics 39*, 2164–2204.

Lum, K. and A. Gelfand (2012). Spatial quantile multiple regression using the asymmetric Laplace process. *Bayesian Analysis 7* (2), 235–258.

Lustig, M., Donoho, D.L., Santos, J.M., and Pauly, J.M. (2008). Compressed Sensing MRI. *IEEE Signal Processing Magazine 25*, 72-82.

Mallows, C. (1973). Some comments on Cp. *Technometrics 15*, 661–675.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3* (3), 205–228.

Manski, C. (1985). Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator. *Journal of Econometrics 27* (3), 313–333.

Meier, L., S. van de Geer, and P. Buhlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Serie B 70* (1), 53–71.

Miguéis, L. V., D. F. Benoit, and D. Van den Poel (2012). Enhanced decision support in credit scoring using Bayesian binary quantile regression. *Journal of the Operational Research Society* (in press).

Peng, J., J. Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., J. R. Pollack, J. R., and Wang, P. (2013). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, *The Annals of Applied Statistics 4* (1), 53-77.

Phillips, R. (2002). Least absolute deviations estimation via the EM algorithm. *Statistics and Computing 12*, 281–285.

Pollard, D. (1990). Empirical Process: Theory and Application. In: *NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 2*. Institute of Mathematical Statistics, Hayward.

Powell, J. (1984). Least absolute deviations estimate for the censored regression model. *Journal of Econometrics 25*, 303–325.

Reed, C. (2011). Bayesian Parameter Estimation and Variable Selection for Quantile Regression. Ph.D. thesis submitted to the department of mathematics and school of information systems, computing and mathematics of Brunel University.

Rosset, S. (2003). Topics in Regularization and Boosting. Dissertation submitted to the department of statistics and the committee on graduate studies of Stanford University.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics 35* (3), 1012–1030.

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association 79*, 871-880.

Rousseeuw, P. J. and A. Leroy. (1987). Robust Regression and Outlier Detection. New York, John Wiley and Sons.

Rousseeuw, P.J. and Yohai, V. (1984). Robust Regression by Means of S-estimators. Robust and Nonlinear Time Series Analysis, edited by J. Franke, W. Härdle, and R.D. Martin, Lecture Notes in Statistics *26, Springer Verlag*, New York, 256-274.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 19* (2), 461–464.

Sharma, D., H. Bondell, and H. Zhang (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics 22* (2), 319–340.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics 22* (2), 231–245.

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

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine 16*, 385–395.

Tibshirani, R. J. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Annals of Applied Statistics 3*, 822–829.

Tobin, J. (1958). Estimation of relationship for limited dependent variables. *Econometrica 26*, 24-36.

Wang, H., Li, G., and Jiang, G. (2007a). Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics 25*, 347 - 355.

Wang, H., Li, G., and Tsai, C. L. (2007b). Regression coefficient and autoregressive order shrinkage and selection via lasso. *Journal of Royal Statistical Society Series B, 69* (1).63-78

Wang, Z., Wu, Y., Zhao, L. (2007c). Change-point estimation for censored regression model. *Science in China. Series A. Mathematics 50*, 63–72.

Wang, Z., Wu, Y., Zhao, L. (2009). Approxiamtion by randomly weighting method in censored regression model. *Science in China. Series A. Mathematics 52*, 561–576.

Wang, Z., Wu, Y., Zhao, L. (2010). A LASSO-type approach to variable selection and estimation for censored regression model. *Chinese Journal of Applied Probability and Statistics 26*, 66–80.

Wei, F. and J. Huang (2010). Consistent group selection in high-dimensional linear regression. *Statistics in Medicine 16*, 1369–1384.

Fox, J. and S. Weisberg (2010). Robust Regression in R.

http://socserv.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Robust-Regression.pdf

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistical Sinica 19*, 801–817.

Xu, J. and Ying, Z. (2010). Simultaneous estimation and variable selection in median regression using lasso-type penalty. *Annals of the Institute of Statistical Mathematics 62*, 487–514.

Yang, Y. and H. Zou (2013). A fast unified algorithm for solving group-lasso penalized learning problems. *Statistics and Computing.* Major revision.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression," *Annals of Statistics 15*, 642-656.

Yu, K., C. Cathy, C. Reed, and D. Dunson (2013). Bayesian variable selection in quantile regression. *Statistics and Its Interface 6*, 261–274.

Yu, K. and R. Moyeed (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* 137,260–76.

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

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

Zhou, Z., Jiang, R., Qian, W. (2013). LAD variable selection for linear models with randomly censored data. *Metrika 76,* 287-300.

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

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101,* 1418–1429.