8-2017

Adaptive Robust Methodology for Parameter Estimation and Variable Selection

Tao Yang
Clemson University, taoy@g.clemson.edu

Adaptive Robust Methodology for Parameter Estimation and Variable Selection

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Tao Yang
August 2017

Accepted by:
Dr. Colin Gallagher, Committee Chair
Dr. Christopher McMahan
Dr. Robert Lund
Dr. Xiaoqian Sun

Abstract

The dissertation consists of three distinct but related projects. We consider regression model fitting, variable selection in regression, and autocorrelation estimation in time series. In each procedure we formulate the problem in terms of minimizing an objective function which adapts to the given data.

First we propose a robust M-estimation procedure for regression. The main purpose of the proposed methodology is to develop a procedure that adapts to light/heavy tailed, symmetric/asymmetric distributions with/without outliers. We focus on studying the properties of the maximum likelihood estimator of the asymmetric exponential power distribution, a broad distribution class that holds both Normal and asymmetric Laplace distributions as special cases. The proposed methodology unifies least squares and quantile regression in a data driven manner to capture both tail decay and asymmetry of the underlying distributions. Finite sample performance of the method is exhibited via extensive Monte Carlo simulation and real data applications.

Second, we capitalize on the success of the proposed method and extend it to a variable selection procedure that selects the important predictors under a sparse setting. Quantile regression Lasso, i.e., quantile regession with L_1 norm on the regression coefficients for regularization is a robust technique to perform variable selection. However which quantile should be adopted is unclear. The proposed methodology introduces a way to choose the most "informative" quantile of interest that is used in the adaptive quantile regression Lasso. A modified BIC criterion is used to select the optimal tuning parameter. The proposed procedure selects the quantile based on the log-likelihood of the asymmetric Laplace distribution, and aims to perform the best quantile regression Lasso which is confirmed in both simulation study and a real data analysis.

Third, we focus on alleviating the underestimation issue of the sample autocorrelation in linear stationary time series. We first formulate autocorrelation estimation into a least squares prob-

lem and then apply a penalization to regulate the autocorrelation estimate. An adaptive sequence is proposed for tuning parameter and is shown to work well for stationary time series when the sample size is small and correlation is high.

Dedication

I dedicate this work to my loving parents, who always believe in me and support me in every step of the way.

Acknowledgments

First and foremost, I would like to show my deep gratitude to my advisors, Dr. Colin Gallagher and Dr. Christopher McMahan for their consistent support during the whole program. Especially, I want to thank Dr. Gallagher for his excellent mathematical insights that inspire me and thank Dr. McMahan for his extreme patience in the face of numerous obstacles. I could not have completed the work without their inspiration and advice.

I would also like to thank Dr. Robert Lund and Dr. Xiaoqian Sun for their serving on my committee, and thank to my fellow Ph.D. student Yan Liu for many discussions we had about details of research. In addition, I would like to thank the Department of Mathematics Sciences for providing financial support during my Ph.D. studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical inferences are usually based on assumptions and lots of statistical procedures rely on distributional models for the error term. Many assumed error structures, e.g. Normal distribution closely describes the distribution of many practical data sets and is mathematical convenient to analyze the pattern in real applications, which many not be exactly true. We consider a robust methodology that allows data to choose the loss function, equivalently to choose underlying distributions from a broad class of distributions for regression coefficients estimation and variable selection. Before introducing the idea of the proposed procedures, we first present a brief review on different robust methodologies.

Huber (1964) proposed M-estimation approach and later Huber (1973) extended this to a regression setting, walked toward a theory of robust estimation by introducing the following loss function:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \tag{1.1}$$

where $\rho(\cdot)$ is a non-constant function. Many regression methods can be viewed as an M-estimation procedure. E.g., $\rho(x) = x^2$ yields least squares (LS), an optimal procedure under Normal distributions but very sensitive to a slight amount of outliers in the data set. Quantile regression (QR) as an alternative robust method introduced by Bassett and Koenker (1978) corresponds to $\rho(x) = \rho_\tau(x)$, where the quantile loss function $\rho_\tau(x) = |x|\{\tau I(x \geq 0) + (1-\tau)I(x < 0)\}$. To further improve on the efficiency of QR, Zou and Yuan (2008) developed a regression method called composite

quantile regression (CQR) which takes a sum of K quantile loss functions at preset quantiles; i.e., $\rho(x) = \sum_{k=1}^{K} \rho_{\tau_k}(x)$, where $\tau_k = k/(1+K)$, for $k = 1, \cdots, K$. Sun et al. (2013) further extend the idea of CQR by assigning different weights to the quantiles used in the CQR loss function with $\rho(x) = \sum_{k=1}^{K} w_k \rho_{\tau_k}(x)$, This allows the weighted composite quantile regression (WCQR) to deal with both symmetric and asymmetric error distributions, and the optimal weights w_k's are selected by minimizing the asymptotic variance of the estimator so that it allows the data to determine which quantiles should be weighted more or less. Another way to achieve robust is to adaptively choose the power which the residuals are raised to in the loss function. Agrò (1992) proposed L_p regression which is essentially by taking $\rho(x) = -\ln f(x)$, where f is the probability density function (pdf) of the generalized error distribution (GED) (see Subbotin, 1923), in which the shape parameter p and the regression coefficients are estimated simultaneously.

These motivate us to develop a procedure that actually chooses the loss function according to two critical information given by the data, the tail decay and the skewness of the underlying error structure. We seek a procedure that allows continuous transition of loss function from LS to QR depending on these two characteristics of the underlying distribution. In particular, we propose an M-estimation procedure with $\rho(x) = -\ln f(x)$, where f is the pdf of a general distribution called asymmetric exponential power distribution (AEPD) which holds many distributions, e.g., Normal, Laplace, asymmetric Laplace as special cases. The proposed method which is essentially the maximum likelihood estimation for mis-specified models estimates regression coefficients and two additional parameters α (tail decay) and τ (asymmetry) simultaneously, where the estimates of α and τ determine the loss function for estimating regression coefficients. The details of the work are shown in Chapter 2.

We then capitalize on the success of the proposed methodology and extend it to regression coefficients estimation and variable selection simultaneously in Chapter 3. Least absolute shrinkage and selection operator (Lasso) first introduced by Tibshirani (1996) is a useful approach to select important predictors under sparse setting and adaptive Lasso developed by Zou (2006) used adaptive weights to penalize coefficients differently further improved the performance of variable selection. Then Wu and Liu (2009) developed penalized QR for variable selection, i.e., QR Lasso which replaces the LS loss in Lasso/Adaptive Lasso by the quantile check loss function to achieve robust property in variable selection. However the performance of QR Lasso procedure relies on the quantile of interest. The efficiency of QR Lasso can drop drastically for different quantiles and which quantile should be

used to achieve the best performance for a given data is what we focus on Chapter 3. Following the same vein of the proposed methodology in Chapter 2, we propose a loss function to be minus log-likelihood of the asymmetric Laplace distribution (ALD), a special case of AEPD for $\alpha = 1$, plus an adaptive L_1 norm penalization for regression coefficients estimation and variable selection simultaneously. The maximum likelihood estimator based on ALD studied by Bera et al. (2016) is essentially a penalized QR which selects the most "informative" quantile, and we further extend it to propose a penalized log-likelihood which aims to choose the most efficient QR Lasso with the quantile of interest dictated by data.

In the last chapter of this dissertation, we focus on the estimation of autocorrelation function in linear stationary time series. Evidence shows that the standard error of simple estimator such as ordinary least squares (OLS) tends to be underestimated, and thus it produces narrower confidence intervals due to ignoring the correlation within error terms (see Bence, 1995, Cochrane and Orcutt, 1949, an Hurlbert, 1984). Bence (1995) considered adjustment which depends on accurate estimation of the autocorrelation at lag one (ρ). Sample autocorrelation which is widely used in time series is a consistent estimate of population autocorrelation and asymptotic normality was established. However it is quite generally realized that sample autocorrelation is liable to bias. The bias is usually negative which is common to many other estimation techniques, e.g., two-step Cochrane-Orcutt, the Durbin estimator, or maximum-likelihood estimation (see Park and Mitchell, 1980, Beesley and Griffiths, 1982, Griffiths and Beesley, 1984, King and Giles, 1984). And the bias issue deteriorates when sample size is smaller and absolute value of ρ gets larger. In Chapter 4 we focus on applying a penalization idea to alleviate this issue. From a best linear predictor point of view, the problem can be formulated into a penalized LS, where the penalization is to adaptively "drag" the ρ estimate toward ± 1 so that underestimation issue can be alleviated.

Chapter 2

A robust regression methodology via M-estimation

2.1 Introduction

Regression is the most common and useful statistical tool which can be used to quantify the relationship between a response variable (y) and explanatory variables ($\boldsymbol{x}$). To this end, the seminal works of both Legendre in 1805 and Gauss in 1809 proposed the method of least squares (LS), which has arguably become the most popular approach to conducting a regression analysis. This popularity is likely attributable to the fact that the LS estimator can be expressed in closed form and can be shown to achieve minimum variance among all unbiased estimators, when the underlying error distribution is normal; e.g., see Rao (1945). However, this approach does not provide an optimal estimator for non-normal settings and is very sensitive to outlying observations (Koenker and Bassett, 1978). Further, experience has shown that LS regression may not be appropriate when the response variable differs from the regression function in an asymmetric manner, which is commonly encountered in medical data, among other venues. In lieu of these deficiencies, herein a general regression methodology is proposed which allows for the possibility of non-normal tail behavior and asymmetry in the conditional distribution of y given $\boldsymbol{x}$, but will still perform well for symmetric and/or normally distributed data.

One way to improve parameter estimates for non-normal data and to guard against the

influence of outlying observations is to replace the LS loss function (i.e., the squared error loss) by a loss function which can accommodate asymmetry in the error distribution and is less susceptible to the magnitude of the residuals. For example, in 1793 Laplace proposed least absolute deviations (LAD), or L_1-norm regression as an alternative to LS. This regression technique is less sensitive to outlying observations and is more appropriate, when compared to LS, for error distributions whose tails are heavier than that of the normal. More generally one can replace the LAD estimator, with an L_p norm estimator; for further discussion see Zeckhauser and Thompson (1970), Mineo (1989) and Agrò (1992). Quantile regression estimates are found by minimizing the quantile (check) loss function, and since they estimate quantiles of the conditional distribution of y given $\boldsymbol{x}$, they are appropriate for asymmetric and heavy tailed distributions (Koenker and Bassett, 1978).

Each of the aforementioned loss functions have corresponding conditional distributions of y given $\boldsymbol{x}$ for which the maximum likelihood estimator (MLE) is equivalent to the estimator which minimizes the corresponding loss: the LS estimator corresponds to the MLE when the error distribution is normal; the LAD estimator is equivalent to the MLE under Laplace errors, the L_p norm estimator corresponds to the MLE when the error terms obey the generalized error distribution (GED) (Subbotin, 1923); and the quantile regression estimator is equivalent to the MLE when the errors follow an asymmetric Laplace distribution (ALPD). Moreover, in these very specific settings the regression estimators are asymptotically most efficient. More generally, the aforementioned loss functions do provide consistent estimators, under standard regularity conditions, but the efficiency of the resulting estimator is inherently tied to the chosen loss and underlying error distribution. That is, there does not exist a universally most efficient approach to conducting a regression analysis. Although, provided a priori knowledge of the error distribution, which is typically not available, a regression methodology could be selected with efficiency in mind. For example, in a location scale regression framework, the efficiency of the quantile regression estimator depends on the quantile of interest. Moreover, under asymmetric Laplace errors the asymptotic variance of the quantile regression estimator is minimized when the analysis proceeds to use the true skewness parameter as the quantile of interest. More generally, the quantile that corresponds to minimizing the asymptotic variance of the estimator depends on the underlying error distribution, which is unknown. The salient point is that to perform a regression analysis an analyst must select a particular methodology, which, in some sense, is equivalent to specifying either the error distribution or loss function under which the regression coefficients are estimated. This work provides a more general approach which

allows the loss function to be selected in a data adaptive fashion, thus resulting in a more efficient and robust estimator.

In order to develop a robust regression procedure one could consider two competing approaches; i.e., perform the regression analysis with respect to multiple loss functions or allow the characteristics of the data to dictate the selection of the loss function. In order to improve the efficiency of quantile regression, Zou and Yuan (2008) introduced composite quantile regression (CQR), which optimizes over a sum of multiple quantile loss functions. As a robust regression procedure, CQR combines the strength of multiple quantile regressions to estimate the same "slope" coefficients across different quantiles. Kai et al. (2010) adapted CQR to the local polynomial framework and established that for many common non-normal errors this extension provided for gains in estimation efficiency when compared to its local LS counterpart. Regretfully, when implementing CQR it is still unclear how many quantiles should be used and simply increasing the number of quantiles does not necessarily improve the efficiency of the estimator; for further discussion see Kai et al. (2010). Alternately, one could consider a convex combination of loss functions; e.g., Zheng et al. (2013) extended CQR by embedding the usage of an empirically weighted average of quantile loss functions and the LS loss function, so that the LS loss tends to be weighted heavier for normally distributed data. Rather than using several quantiles, another tact would be to let the data select the quantile of interest in quantile regression. Bera et al. (2016) proposed a Z-estimator which could be used to simultaneously obtain the quantile regression estimator and the quantile of interest in a data driven fashion, and is hereafter referred to as ZQR. In particular, this estimator is obtained by minimizing an objective function which is inspired by the maximum likelihood score function under the ALPD. Proceeding in this fashion results in a penalized quantile regression framework where the penalty depends on the quantile of interest.

Motivated by the work of Bera et al. (2016), the regression methodology presented herein is developed in the same vein. In particular, a robust loss function is constructed so that the proposed estimator corresponds to the MLE when the error terms obey the asymmetric exponential power distribution (AEPD). The AEPD class of distributions was first proposed by Ayebo and Kozubowski (2003) and holds the normal, skewed normal, Laplace, ALPD, and GED as special cases, among many others. Developing the proposed regression methodology under the AEPD has several definitive advantages. First, and foremost, the proposed method selects the best loss function (e.g., LS, LAD, L_p, quantile, etc.) from a broad class in a data driven fashion. For this reason,

one could view this proposal as a method which unifies and bridges the gaps between LS, LAD, L_p norm, and quantile regression. Secondly, the proposed technique can effectively capture the tail decay and/or asymmetry of the error distribution, thus maintaining a high level of estimation efficiency in venues where other competing procedures do not. Lastly, as the AEPD holds many common distributions as special cases (e.g., normal, skewed normal, ALPD, GED, etc.), one may use model selection criteria, such as AIC or BIC, to identify the "best" model (e.g., LS fit, specific quantile regression fits, etc.), as is demonstrated in subsequent sections.

The remainder of this article is organized as follows. Section 2 presents the modeling assumptions, develops the proposed loss function based on the AEPD, and provides a stable numerical algorithm which can be used to obtain the regression parameter estimates. The consistency and asymptotic normality of the proposed estimator are established in Section 3. The results of an extensive Monte Carlo simulation study designed to assess the finite sample performance of the proposed procedure is provided in Section 4. The results of the motivating data analyses are provided in Section 5. Section 6 concludes with a summary discussion, and the regularity conditions under which the theoretical results can be established are provided in the appendix.

2.2 Methodology

2.2.1 Model Assumption

Consider a linear regression model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon, \tag{2.1}$$

where y denotes the response variable, $\boldsymbol{x}$ is a $(p+1)$-dimensional vector of covariates, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, and ϵ is the error term. Throughout the remainder of this article it is assumed that the error term is independent of the covariates (i.e., $\boldsymbol{x} \perp \epsilon$, where $\perp$ denotes statistical independence) and that the probability density function of ϵ has a unique mode at zero. Under these assumptions, the linear predictor $\boldsymbol{x}'\boldsymbol{\beta} = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p$ represents the unique mode of the conditional distribution of y given $\boldsymbol{x}$. Note, this model becomes a mean regression model when the distribution of ϵ is symmetric and has a finite first moment. The primary focus of this work is aimed at estimating the "slope" parameters, $\boldsymbol{\beta}^* = (\beta_1, \ldots, \beta_p)'$, since the intercept, β_0, provides solely for a shift between different regression functions of interest; i.e., regression functions

such as the mean and median for (2.1) have identical unknown slope parameters.

For ease of exposition, assume that the error term in (2.1) follows an AEPD, this assumption is later relaxed in subsequent sections. A random variable ϵ is said to follow an AEPD if there exist parameters $\alpha > 0$, $\mu \in \mathbb{R}$, $\sigma > 0$ and $0 < \tau < 1$ such that the probability density function of ϵ has the form

$$f(\epsilon) = \frac{\alpha\tau(1-\tau)}{\Gamma(\frac{1}{\alpha})\sigma} \exp\left[-\frac{|\epsilon-\mu|^\alpha}{\sigma^\alpha}\{I(\epsilon \geq \mu)\tau^\alpha + I(\epsilon < \mu)(1-\tau)^\alpha\}\right], \tag{2.2}$$

where μ is the location (mode) parameter, σ is the scale parameter, τ controls the skewness and α is the shape (tail decay) parameter. For notational brevity, this relationship is denoted $\epsilon \sim AEPD(\mu, \alpha, \sigma, \tau)$. The AEPD class of distributions hold many common distributions as special cases; e.g., the epsilon-skew-normal distribution, studied by Mudholkar and Hutson (2000), is obtained when $\alpha = 2$, which holds the normal distribution as a special case when $\tau = 0.5$; Specifying $\alpha = 1$ results in the ALPD which holds the Laplace distribution as a special case when $\tau = 0.5$; And the GED results from specifying $\tau = 0.5$. Moreover, as α approaches ∞, the AEPD approaches a uniform distribution with parameter $(\mu - \sigma/(1-\tau), \mu + \sigma/\tau)$. To illustrate the broad spectrum of shapes for which the AEPD density can take, Figure 2.1 depicts several AEPD densities for different combinations of α and τ, where $\mu = 0$ and σ is specified such that the variance is unity.

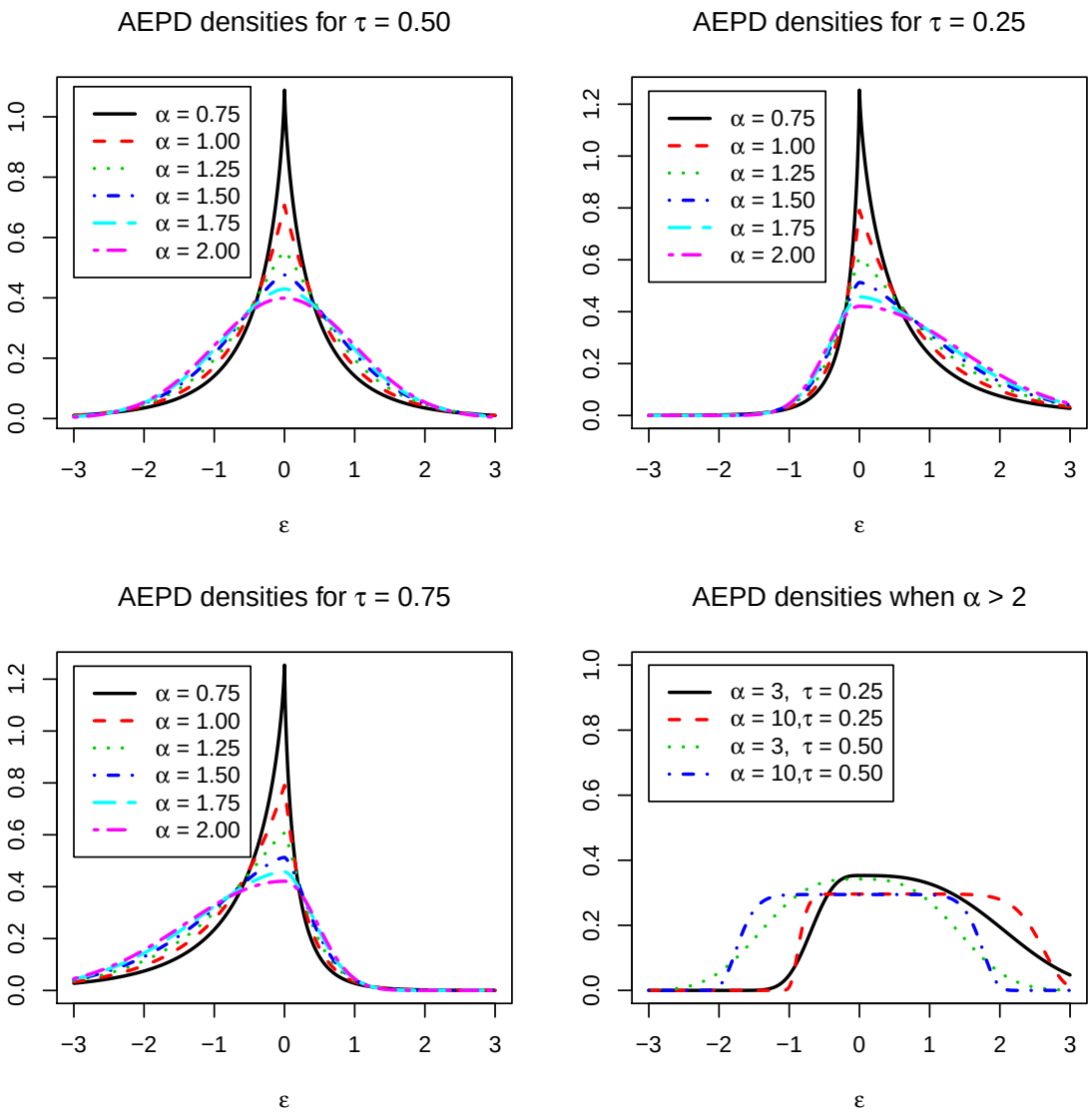
Under the aforementioned assumptions, the response variable conditionally, given the covariates, follows an AEPD; i.e., $y|\boldsymbol{x} \sim AEPD(\boldsymbol{x}'\boldsymbol{\beta}, \alpha, \sigma, \tau)$. Thus, the log-likelihood of the observed data $\{(y_i, \boldsymbol{x}_i'), i = 1, \ldots, n\}$ is given by

$$\begin{aligned}\rho(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \ln\left\{\frac{\alpha}{\Gamma(1/\alpha)}\right\} + \ln\{\tau(1-\tau)\} - \ln(\sigma) \\ - \frac{1}{n\sigma^\alpha}\sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|^\alpha \left\{I(y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta})\tau^\alpha + I(y_i < \boldsymbol{x}_i'\boldsymbol{\beta})(1-\tau)^\alpha\right\},\end{aligned} \tag{2.3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \alpha, \sigma, \tau)$ denotes the collection of model parameters and $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0', \alpha_0, \sigma_0, \tau_0)$ represents the true unknown value of $\boldsymbol{\theta}$. More generally, in the case in which the error distribution does not belong to the AEPD class, (2.3) can be viewed as a loss function, which still can be used to efficiently estimate the regression coefficients, as is demonstrated in Sections 4 and 5. In either case, let $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\alpha}, \hat{\sigma}, \hat{\tau})$ denote the value of $\boldsymbol{\theta}$ which maximizes (2.3); i.e., $\hat{\boldsymbol{\theta}}$ is the proposed estimator of $\boldsymbol{\theta}_0$.

To illustrate how the proposed approach is data adaptive, it is first noted that the process

Figure 2.1: The AEPD densities for different parameter configurations.



of estimating $\boldsymbol{\theta}_0$ via maximizing (2.3), can be viewed as a two-step process. First, for fixed values of α, σ and τ an estimate of $\boldsymbol{\beta}_0$ is obtained by minimizing the following loss function

$$\sum_{i=1}^{n} |y_i - \boldsymbol{x}_i'\boldsymbol{\beta}|^{\alpha} \left\{ I(y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta})\tau^{\alpha} + I(y_i < \boldsymbol{x}_i'\boldsymbol{\beta})(1-\tau)^{\alpha} \right\}, \tag{2.4}$$

This estimator is denoted as $\hat{\boldsymbol{\beta}}(\alpha, \tau)$. The second step estimates the remaining parameters by maximizing (2.3) after replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\alpha, \tau)$. The key feature of this approach is that every combination of α and τ corresponds to a different loss function specification in (2.4), and consequently results in obtaining a different estimate of $\boldsymbol{\beta}_0$. For example, if $\alpha = 2$ and $\tau = 0.5$, the proposed approach and LS obtain the same estimate; when $\alpha = 1$ and $\tau = \tau^*$, the resulting estimate is identical to the quantile regression estimate with the quantile of interest being τ^*. The salient point: by estimating α_0 and τ_0 the proposed procedure allows the data to determine the shape and skewness of the underlying distribution and as consequence selects the form of the loss function which is used to estimate the regression coefficients.

2.2.2 A general error distribution and the Kullback Leibler divergence

In the setting in which the error distribution does not belong to the AEPD class, one could view the model for the conditional distribution of y, given $\boldsymbol{x}$, as being misspecified. Denote the true probability density function for y, given $\boldsymbol{x}$, by $f^*(y|\boldsymbol{x})$, the assumed parametric density by $f(y|\boldsymbol{x};\boldsymbol{\theta})$, and the density of $\boldsymbol{x}$ as $h(\boldsymbol{x})$. Further, define the joint density of y and $\boldsymbol{x}$ as $g^*(y, \boldsymbol{x}) = f^*(y|\boldsymbol{x})h(\boldsymbol{x})$ and $g_{\boldsymbol{\theta}}(y, \boldsymbol{x}) = f(y|\boldsymbol{x};\boldsymbol{\theta})h(\boldsymbol{x})$ under the true and assumed model, respectively. Subsequently, the Kullback-Leibler divergence is defined by

$$D_{KL}(g^*||g_{\boldsymbol{\theta}}) = -E\left\{ \ln \frac{g_{\boldsymbol{\theta}}(y, \boldsymbol{x})}{g^*(y, \boldsymbol{x})} \right\} = -E\left\{ \ln \frac{f(y|\boldsymbol{x};\boldsymbol{\theta})}{f^*(y|\boldsymbol{x})} \right\}, \tag{2.5}$$

where the expectation is taken with respect to the true distribution g^*. Minimizing (2.5) with respect to $\boldsymbol{\theta}$, or equivalently maximizing (2.3), results in identifying the AEPD density closest to $f^*(y|\boldsymbol{x})$, i.e., $f(y|\boldsymbol{x};\hat{\boldsymbol{\theta}})$ is the "projection" of $f^*(y|\boldsymbol{x})$ onto the AEPD class. More specifically, obtaining $\hat{\boldsymbol{\theta}}$ as the maximizer of (2.3) is equivalent to finding the AEPD density closest to the true probability density with respect to the observed empirical distribution. This feature allows the proposed approach to be robust to the structure of the underlying error distribution and to maintain a high level of estimation

efficiency, by permitting the loss function (i.e., the assumed AEPD density) to adapt to the true underlying structure of the data.

2.2.3 Numerical algorithm

In order to develop a numerical algorithm for obtaining $\hat{\boldsymbol{\theta}}$, the dimension of the loss function presented in (2.3) is reduced. In particular, for fixed values of $\boldsymbol{\beta}$ and α, the values of σ and τ which maximize (2.3) can be expressed as

$$\sigma(\boldsymbol{\beta},\alpha)=\left(\frac{\alpha}{n}[e^{+}(\boldsymbol{\beta},\alpha)\tau(\boldsymbol{\beta},\alpha)^{\alpha}+e^{-}(\boldsymbol{\beta},\alpha)\{1-\tau(\boldsymbol{\beta},\alpha)\}^{\alpha}]\right)^{1/\alpha},$$

$$\tau(\boldsymbol{\beta},\alpha)=\left[1+\{e^{+}(\boldsymbol{\beta},\alpha)/e^{-}(\boldsymbol{\beta},\alpha)\}^{1/(\alpha+1)}\right]^{-1},$$

respectively, where $e^{+}(\boldsymbol{\beta},\alpha)=\sum_{i=1}^{n}|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|^{\alpha}I(y_i\geq\boldsymbol{x}_i'\boldsymbol{\beta})$ and $e^{-}(\boldsymbol{\beta},\alpha)=\sum_{i=1}^{n}|y_i-\boldsymbol{x}_i'\boldsymbol{\beta}|^{\alpha}I(y_i<\boldsymbol{x}_i'\boldsymbol{\beta})$. Replacing σ and τ in (2.3) by $\sigma(\boldsymbol{\beta},\alpha)$ and $\tau(\boldsymbol{\beta},\alpha)$, respectively, leads to the following loss function

$$Q(\boldsymbol{\beta},\alpha)=\ln\left\{\frac{\alpha}{\Gamma(1/\alpha)}\right\}-\frac{1}{\alpha}\ln\left(\frac{\alpha}{n}\right)-\frac{1}{\alpha}-\frac{1+\alpha}{\alpha}\ln\left\{e^{+}(\boldsymbol{\beta},\alpha)^{1/(\alpha+1)}+e^{-}(\boldsymbol{\beta},\alpha)^{1/(\alpha+1)}\right\}. \quad (2.6)$$

In order to maximize (2.6) with respect to $\boldsymbol{\beta}$ and α, an iterative algorithm is employed with a well specified initial value. The proposed algorithm proceeds as follows:

1. Set $j=1$ and initialize $\boldsymbol{\theta}^{(0)}=\{\boldsymbol{\beta}^{(0)},\alpha^{(0)},\sigma^{(0)},\tau^{(0)}\}$ as $\boldsymbol{\beta}^{(0)}=\boldsymbol{\beta}(\tau^{(0)})$, $\alpha^{(0)}=1$,

$$\sigma^{(0)}=n^{-1}\sum_{i=1}^{n}\rho_{\tau^{(0)}}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta}^{(0)}),$$

$$\tau^{(0)}=\arg\min_{\tau}\sum_{i=1}^{n}\rho_{\tau}\{y_i-\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)\}/\{\tau(1-\tau)\},$$

   where $\rho_{\tau}(\cdot)$ is the usual quantile check loss function and $\boldsymbol{\beta}(\tau)=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\rho_{\tau}(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})$.

2. Compute $\boldsymbol{\beta}^{(j)}=\arg\max_{\boldsymbol{\beta}}Q(\boldsymbol{\beta},\alpha^{(j-1)})$ and $\alpha^{(j)}=\arg\max_{\alpha}Q(\boldsymbol{\beta}^{(j)},\alpha)$, respectively, and set $j=j+1$.

3. Repeat step 2 until convergence.

At the point of convergence the proposed estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}', \hat{\alpha}, \hat{\sigma}, \hat{\tau})$ is determined as $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(j)}$, $\hat{\alpha} = \alpha^{(j)}$, $\hat{\sigma} = \sigma(\boldsymbol{\beta}^{(j)}, \alpha^{(j)})$, and $\hat{\tau} = \tau(\boldsymbol{\beta}^{(j)}, \alpha^{(j)})$. Note, the more complex initialization step provides the numerical algorithm with a well posed initial value and results in gains in computational efficiency. Further, the necessary optimization steps throughout the algorithm can easily be completed using standard numerical software; e.g., `quantreg`, `optim`, and `optimize` in R.

2.3 Asymptotic properties

The proposed methodology falls under the general class of M-estimators introduced by Huber (1964), and as such, standard regularity conditions ensure consistency and asymptotic normality of the resulting estimators. The specific technical conditions required are given in the appendix, along with a brief discussion.

Theorem 2.3.1. *(Consistency). Under regularity condition (A1)-(A6), provided in the appendix,* $\hat{\boldsymbol{\theta}}$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*; i.e.* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Theorem 2.3.2. *(Asymptotic Normality). Under regularity condition (A1)-(A7), provided in the appendix, and for* $\alpha_0 > 1$*. The M-estimator* $\hat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}_0$ *is asymptotically normal; i.e.,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{V}_{2\boldsymbol{\theta}_0}^{-1} \boldsymbol{V}_{1\boldsymbol{\theta}_0} \boldsymbol{V}_{2\boldsymbol{\theta}_0}^{-1}),$$

where $\boldsymbol{V}_{1\boldsymbol{\theta}_0} = E\{\psi(y, \boldsymbol{x}, \boldsymbol{\theta}_0)\psi(y, \boldsymbol{x}, \boldsymbol{\theta}_0)'\}$, $\boldsymbol{V}_{2\boldsymbol{\theta}_0} = [\partial E\{\psi(y, \boldsymbol{x}, \boldsymbol{\theta})\}/\partial \boldsymbol{\theta}']|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ *and* $\psi(y, \boldsymbol{x}, \boldsymbol{\theta}) = \partial \ln\{f(y|\boldsymbol{x}; \boldsymbol{\theta})\}/\partial \boldsymbol{\theta}$.

Note, the proofs of Theorem 2.3.1 and Theorem 2.3.2 are standard, and simply involve verifying the conditions outlined by Huber (2009); for further discussion see the appendix. It is worthwhile to note that the computation of the asymptotic covariance matrix in Theorem 2.3.2 depends on the unknown distribution of the errors, thus making a direct appeal to asymptotic based inference challenging; i.e., an additional step has to be undertaken in order to estimate the error distribution. This same challenge is commonly encountered in other existing techniques; e.g., QR. Further, based on simulation studies (results not shown), it was ascertained that standard asymptotic based inference based on the result established in Theorem 2.3.2 may not be appropriate

for relatively small sample sizes; e.g., when $n = 200$. Thus, it is suggested that bootstrapping be adopted for the purposes of conducting finite sample inference. In general, bootstrapping is an approach which can be used to obtain an improved approximation of the sampling distribution of a statistic, and is most theoretically sound for statistics which achieve asymptotic normality. Under the regularity conditions provided in the appendix, it can easily be shown that the conditions of Arcones and Giné (1992) are satisfied, thus assuring in this context that the bootstrapped estimator is consistent and possess the same limiting distribution depicted in Theorem 2.3.2. These results tend to suggest that bootstrapping will provide reliable finite sample inference through use of the standard bootstrap distribution.

In what follows, the bootstrapping procedure implemented throughout the remainder of this article is briefly described. To begin, for a given data set (i.e., $(y_i, \boldsymbol{x}_i'), i = 1, \ldots, n$) the numerical algorithm described in Section 2.2.3 is used to obtain an estimate of the regression parameters. Using the regression coefficient estimates, one then computes the residuals $e_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$, for $i = 1, \ldots, n$. A random sample of size n is then drawn from the set of residuals, with replacement, providing the bootstrapped residuals e_i^*, for $i = 1, ..., n$. The bootstrapped response is subsequently obtained via $y_i^* = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + e_i^*$, and the proposed approach is used to model this data (i.e., $(y_i^*, \boldsymbol{x}_i'), i = 1, \ldots, n$) resulting in the bootstrapped estimate $\boldsymbol{\theta}^*$. This process is repeated B times yielding B bootstrap replicates of the regression coefficients. The bootstrap replicates can then be used to construct standard error estimates in the usual fashion (Efron, 1982), and $(1-\alpha)100\%$ bootstrap confidence intervals using the empirical $(\alpha/2)100\%$th and $(1-\alpha/2)100\%$th percentiles of the bootstrap distribution.

2.4 Simulation study

In order to examine the finite sample performance of the proposed approach, the following Monte Carlo simulation study was conducted. This study considers a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{ for } i = 1, \ldots, n, \tag{2.7}$$

where $\beta_0 = 1$, $\beta_1 = 0.1$, and $x_i \sim N(0, 1)$. In order to illustrate the robustness property of the proposed estimator, several distributions of the error term ϵ_i are considered, both within and outside of

the AEPD class. In particular, the investigations discussed herein consider the settings in which the error terms are distributed $AEPD(0, 2, \sigma, 0.5)$, $AEPD(0, 1, \sigma, 0.5)$, and $AEPD(0, 1.5, \sigma, 0.25)$, with the two former specifications providing for standard normal and Laplacian errors, respectively, where the σ parameters were selected so that the variance of the error term is 1. For error distributions outside of the AEPD class, this study considers Student's t-distribution with 3 degrees of freedom; a skewed normal distribution with a slant parameter of 4 (Azzalini, 1985); a skewed t-distribution with 3 degrees of freedom and a skewing parameter of 0.5 (Fernandez and Steel, 1998); a Chi-square distribution with 3 degrees of freedom; and a log-normal distribution with location and scale parameters being set to be 0 and 0.5, respectively. These choices provide for a broad spectrum of characteristics of the error distribution which are commonly encountered in practical applications; to include symmetry, heavy tails, and positive skewness. For each of the above error distributions, $m = 500$ independent data sets were generated, each consisting of $n = 200$ observations.

The proposed methodology denoted by AME (adaptive M-estimator) was implemented to analyze each of the simulated data sets, using the techniques outlined in Section 2 and 3. In order to provide a comparison between the proposed methodology and existing techniques, several competing procedures were also implemented. In particular, each data set was analyzed using LS, LAD, and ZQR. The two former techniques are staples among standard data analysis methods, while the latter can be viewed as a generalization of quantile regression which estimates the quantile of interest along with the rest of the model parameters, thus allowing the approach to "adapt" to the data. In order to estimate standard errors and to construct confidence intervals the standard techniques were used for LS, while bootstrapping techniques with $B = 1000$ were used for the proposed approach, LAD, and ZQR. It is worthwhile to point out that all of the aforementioned methods attempt to estimate the same slope coefficient (i.e., β_1) in the data generating model above; i.e., β_1 is the slope coefficient for the mean and all quantile functions. Thus, this study focuses solely on the results that were obtained from the proposed approach and the three competing techniques for the slope parameter.

Table 2.1 provides a summary of the estimators resulting from the proposed procedure, across all considered error distributions. In particular, this summary includes the empirical bias, the relative efficiency of the estimator (i.e., the average estimated standard error of the estimator divided by the average estimated standard error of the proposed estimator), empirical coverage probabilities associated with 95% confidence intervals, and average confidence interval length. From these results, one will notice that the proposed method performs very well; i.e., our estimator exhibits little if any

Table 2.1: Simulation results summarizing the estimates of the "slope" coefficient obtained by AME, LS, LAD, and ZQR, for both the AEPD and non-AEPD error distributions. This summary includes the average estimate minus the true value (Bias), relative efficiency (Eff), estimated coverage probability (Cov) associated with 95% confidence intervals, and averaged confidence interval length (AL). Here t_3 denotes Student's t-distribution with 3 degrees of freedom; χ^2_3 denotes a Chi-square distribution with 3 degrees of freedom; $SN(4)$ denotes a skewed normal distribution with a slant parameter of 4; $ST(3, 0.5)$ for skewed t-distribution with 3 degrees of freedom and a skewing parameter of 0.5.

	N(0,1)				t_3			
	Bias	Eff	Cov	AL	Bias	Eff	Cov	AL
LS	0.0051	**0.9741**	0.960	**0.2796**	-0.0064	1.2331	0.954	0.4677
LAD	0.0037	1.2626	0.958	0.3598	-0.0036	1.0238	0.953	**0.4042**
ZQR	0.0037	1.2064	0.976	0.3765	-0.0028	1.0023	0.955	0.4174
AME	0.0053	1.0000	0.974	0.3104	-0.0023	**1.0000**	0.961	0.4238
	χ^2_3				Log-normal			
LS	0.0053	2.3995	0.943	0.6811	-0.0017	1.4453	0.929	0.1665
LAD	-0.0016	2.5165	0.959	0.7644	-0.0028	1.4602	0.965	0.1814
ZQR	-0.0043	1.1525	0.977	0.4027	-0.0049	1.1287	0.963	0.1429
AME	-0.0030	**1.0000**	0.969	**0.3426**	-0.0032	**1.0000**	0.965	**0.1276**
	$SN(4)$				$ST(3, 0.5)$			
LS	-0.0014	1.0958	0.952	0.1768	-0.0224	2.0737	0.947	0.7030
LAD	-0.0020	1.4075	0.966	0.2276	-0.0081	1.4816	0.971	0.5871
ZQR	0.0004	1.2190	0.958	0.1988	-0.0066	1.0047	0.977	0.4375
AME	-0.0012	**1.0000**	0.942	**0.1659**	-0.0075	**1.0000**	0.979	**0.4375**
	Laplace				$AEPD(0, 1.5, \sigma, 0.25)$			
LS	-0.0016	1.2904	0.942	0.2781	0.0037	1.3014	0.957	0.2780
LAD	-0.0028	**0.9875**	0.964	**0.2347**	0.0030	1.5021	0.979	0.3390
ZQR	-0.0042	0.9971	0.968	0.2434	0.0012	1.1321	0.973	0.2733
AME	-0.0044	1.0000	0.972	0.2562	0.0001	**1.0000**	0.965	**0.2510**

evidence of bias and the empirical coverage probabilities appear to attain their nominal level.

Table 2.1 also provides the same summary for the other three competing regression techniques. Unsurprisingly, the same conclusions discussed above can also be drawn for LS, LAD, and ZQR, but differences in performance are apparent. First, under normal (Laplacian) errors the most efficient procedure is LS (LAD), which can be ascertained by examining both the relative efficiency and the average confidence interval length. Note, this finding was expected since LS and LAD result in the MLE under normal and Laplacian errors, respectively. With that being said, one will also note that the estimators resulting from the proposed approach are almost as efficient as the most efficient estimator under normal and Laplacian errors, even though the proposed method is tasked to estimate two additional parameters in these settings. Second, for all other considered error distributions the proposed method provided for the most efficient estimator, with the exception of the setting in which the errors obey a Student's t-distribution. In some cases the efficiency gains are

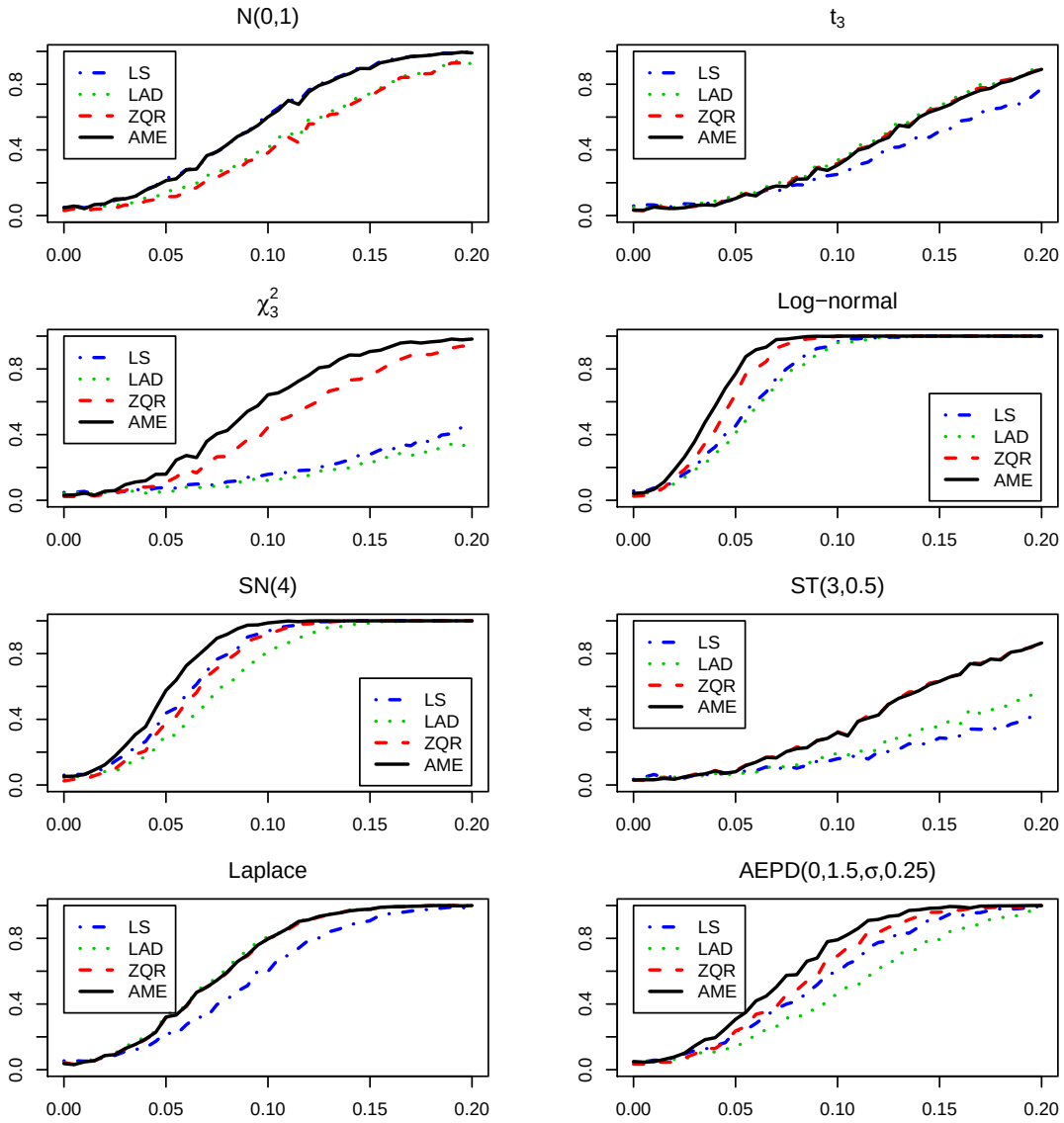
substantial; e.g., under Chi-square errors the proposed estimator is twice as efficient when compared to the the LS and LAD estimators. In general, the proposed approach performed better in terms of estimation efficiency than LS, LAD, and ZQR when the error distribution was asymmetric. Moreover, the proposed approach surprisingly outperformed ZQR, which is the most comparable existing technique, in all considered settings. In summary, this simulation study illustrates that the proposed methodology provides reliable estimates across a broad spectrum of potential error distributions, and can provide for more efficient estimates when compared to existing regression methods. Moreover, these gains in estimation efficiency are more dramatic for asymmetric error distributions.

2.4.1 Power of the hypothesis test

In order to investigate other inferential characteristics of the proposed approach, a power analysis was conducted to assess the performance of the proposed methodology when utilized to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, at the $\alpha = 0.05$ significance level. Data for this study was generated in the exact same fashion as was described above with a few minor exceptions; i.e., here the slope coefficient is taken to be $\beta_1 \in \{0, 0.005, \ldots, 0.2\}$, then for each error distribution and value of β_1, $m = 1000$ independent data sets are generated each consisting of $n = 500$ observations. Our approach along with LS, LAD, and ZQR were applied to each of the data sets and the results from these analyses were used to create 95% confidence intervals, as was described in the previous section. Decisions between the null and alternative hypothesis were made based on the confidence intervals in the usual fashion. These results were then used to construct power curves for each of the regression techniques, under each of the considered error distributions.

Figure 2.2 provides the empirical power curves for all four regression techniques across all considered error distributions. Again, as one should expect, in the case of Gaussian and Laplacian errors the methods with the most power are LS and LAD, respectively, but the power curve for the proposed approach is practically identical. In contrast, when the error distribution is not normal or Laplace, LS and LAD can suffer from a dramatic loss in power (e.g., Chi-square or skewed t errors) a feature which the proposed approach does not possess. In fact, for skewed distributions one will note that the proposed approach has the most power to detect departures from the null, under all considered configurations. In summary, the findings from this study reinforce the main findings discussed above; i.e., the proposed methodology provides for efficient estimation and reliable inference across a broad spectrum of error distributions.

Figure 2.2: Empirical power curves obtained under AME, LS, LAD, and ZQR. Here t_3 denotes Student's t-distribution with 3 degrees of freedom; χ^2_3 denotes a Chi-square distribution with 3 degrees of freedom; $SN(4)$ denotes a skewed normal distribution with a slant parameter of 4; $ST(3, 0.5)$ for skewed t-distribution with 3 degrees of freedom and a skewing parameter of 0.5.



2.5 Data applications

In this section the proposed M-estimator is used to analyze two data sets. These applications further illustrate the useful properties of the proposed regression methodology.

2.5.1 Blood pressure data

The National Health and Nutrition Examination Survey (NHANES) is a Center for Disease Control and Prevention program which was initiated to assess the general health of the populous in the United States. As a part of this study, data is collected from participants via questionnaires and various physical exams, to include laboratory testing. This information is subsequently made publicly available so that researchers may address/explore future medical, environmental, and public health issues that the United States, and more generally the world, may face. One such issue involves the significant number of adults who are affected by high blood pressure. In fact, the World Health Organization (World Health Organization; 2016) estimates that 22% of adults overs the age of 18 have abnormally high blood pressure, equating to approximately 1.2 billion afflicted individuals world wide. Individuals with chronic high blood pressure may develop further sequelae to include aneurysms, coronary artery disease, heart failure, strokes, dementia, kidney failure, etc. (Chobanian et al., 2003). Thus, developing a sound understanding of the relationship that exists between blood pressure and other risk factors is essential to public health.

To this end, the analysis considered herein examines blood pressure data collected on the participants of the NHANES study during the years of 2009-2010, and attempts to relate this response (diastolic blood pressure) to several different risk factors. In particular, the risk factors selected for this study include a binary variable (Food) indicating whether the participant had eaten within the last 30 minutes (with 1 indicating that they had, and 0 otherwise), the average number of cigarettes smoked per day during the past 30 days (Cigarette), the average number of alcoholic drinks consumed per day during the past 12 months (Alcohol), and the participants age (Age). This analysis assumes that a first order linear model is appropriate, and uses the proposed approach as well as LS, LAD, and ZQR to complete model fitting. These techniques were implemented in the exact same fashion as was described in Section 3.3. Table 2.2 reports the estimated regression coefficients as well as the corresponding standard errors obtained from this analysis.

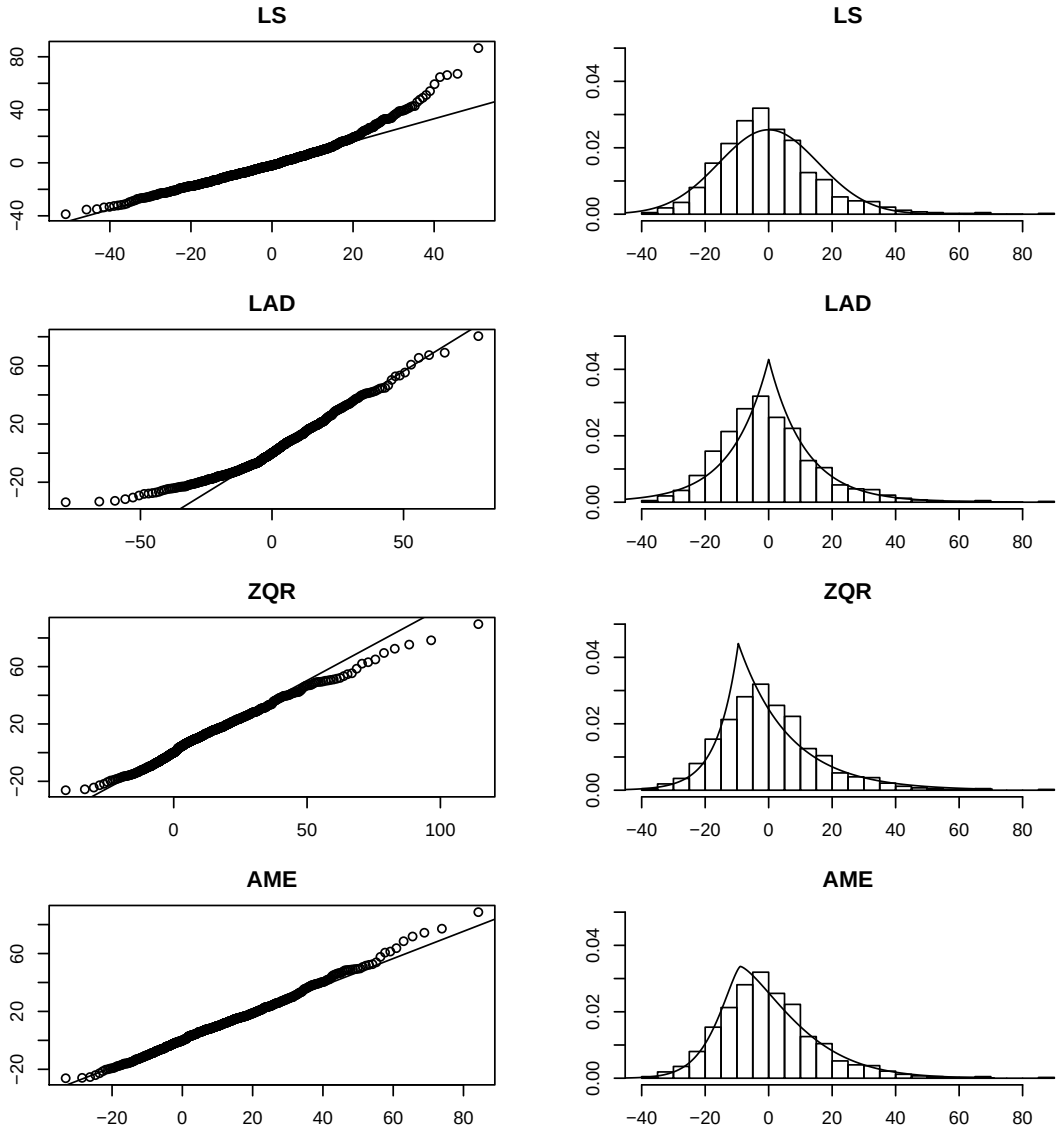
From the results presented in Table 2.2, one will note that the findings between the four

Table 2.2: Blood pressure data analysis: Estimated regression coefficients, their estimated standard errors in parenthesis, and the values of the model selection criteria AIC and BIC, resulting from AME, LS, LAD, and ZQR.

Estimate(SE)					
Method	Food	Alcohol	Cigarette	Age	(AIC,BIC)
LS	-0.481(1.279)	0.452(0.157)	-0.001(0.069)	0.473(0.041)	(7066.61, 7095.06)
LAD	0.170(1.251)	0.430(0.170)	-0.024(0.057)	0.385(0.044)	(7028.93, 7057.37)
ZQR	1.429(1.451)	0.429(0.182)	0.000(0.051)	0.286(0.051)	(6985.03, 7018.22)
AME	0.675(1.170)	0.405(0.160)	0.004(0.047)	0.316(0.039)	(6962.68, 7000.61)

regression methodologies are similar, but differences are apparent. In particular, this analysis finds that age and alcohol are significantly (positively) related to diastolic blood pressure, with the other two covariates being insignificant. The effect estimate associated with alcohol consumption is in agreement across all of the techniques, but the same cannot be said for the age effect. In particular, the proposed method actually renders an age effect estimate that is statistically different (or essentially) than the effect estimate which was obtained by LS. In contrast, the age effect estimates obtained by the proposed approach and ZQR are generally in agreement, this is likely attributable to the fact that both of these techniques are designed to adapt to the asymmetry of the data, which is present in this analysis; e.g., the proposed approach estimated the shape parameter to be $\hat{\alpha} \approx 1.4$ and the skewness parameter to be $\hat{\tau} \approx 0.3$, indicating that the error distribution has heavy tails and is right skewed. To further investigate this, Figure 2.3 provides QQ-plots and histograms of the residuals obtained under the four regression methodologies. From Figure 2.3 one would note that LS, LAD, and likely even ZQR would fail basic diagnostic checks. Further, when comparing the standard error estimates of the age effect one will also note that the proposed approach renders a smaller value when compared to LS, LAD and ZQR, which is not surprising given the results discussed in Section 4. Ultimately, in terms of choosing a "best" model fit in this scenario one could make use of the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to select between model fits, noting that the proposed approach holds the other 3 as special cases. Table 2.2 provides the values of these model selection criteria for all of the regression techniques, and one will note that both techniques unanimously select the model fit via the proposed approach. In summary, whether based on standard diagnostic procedures, estimator efficiency, or model selection criteria, the proposed approach appears to be the favorable technique for this application.

Figure 2.3: QQ-plots and histogram of the residuals under AME, LS, LAD, and ZQR for the blood pressure dataset.



2.5.2 Miscarriage data

The Collaborative Perinatal Project (CPP) was a longitudinal study, conducted from 1957 to 1974, which was aimed at assessing multiple aspects of maternal and child health (Hardy, 2003). Even though this study was conducted half a century ago, the information collected still constitutes an important resource for biomedical research in many areas of perinatology and pediatrics. For example, in 2007 a nested case-control study which examined whether circulating levels of chemokines were related to miscarriage risk was conducted using ($n = 745$) stored serum samples collected as a part of the original CPP study, for further details see Whitcomb et al. (2007). In particular, this study focused on monocyte chemotactic protein-1 (MCP1), which is a cytokine that is located on chromosome 17 in the human genome and is believed to have a pregnancy regulatory function, for further details see Wood (1997). The cases (controls) in the study were participants who had (not) experienced a spontaneous miscarriage, and cases were matched with controls based on gestational age. In addition to the measured MCP1 levels, several other variables were collected on each participant; i.e., age, race (with 1 denoting Africa American, and 0 otherwise)), smoke (with 1 denoting that the participant had smoked before and 0 otherwise) and miscarriage status (with 1 denoting that miscarriage had been experienced, and 0 otherwise).

In this analysis, the measured MCP1 level is considered to be the response variable and all other variables are treated as covariates. A full linear model consisting of all first order terms, as well as all pairwise interactions, is assumed and best subset model selection is implemented using BIC as the criteria. The proposed approach was used to fit all possible models, including models where (α, τ) were set to be $(2, 0.5)$, $(1, 0.5)$, and $(1, \tau)$, which are equivalent to implementing LS, LAD, and ZQR, respectively. Model fitting was conducted in the exact same fashion as was described in Section 3.3. This process identified an intercept only model for LS, a model consisting of race as the only covariate for ZQR, and a first order model consisting of both age and race as covariates for LAD and AME, with BIC values of 431.28, -395.51, -822.02, and -937.55 for LS, LAD, ZQR, and AME, respectively. From these results, one will note that important relationships will potentially be missed when the appropriate regression methodology is not used. In particular, this study illustrates that the best model chosen under these existing procedures may differ from the best model chosen under the proposed approach. When one considers the model selection process described above, it would be natural to place more faith in the results that were obtained under the proposed approach. This

assertion is based on two primary facts. First, the proposed procedure holds the other competing procedures as special cases, thus the analysis described above should be viewed as a much more in depth model selection process which evaluates whether it is more reasonable to view α and τ as being fixed (with known value) or as unknown parameters. Second, of the considered techniques the proposed procedure has the ability to more aptly adapt to the underlying error structure, and is therefore able to render a more reliable analysis.

2.6 Conclusions

This work has developed a general robust regression methodology, which was inspired by the asymmetric exponential power distribution. In particular, the proposed methodology is robust with respect to the underlying error structure, thus rendering reliable estimation and inference across a broad spectrum of error distributions, even in the case of heavy tails and/or asymmetry. This is made possible by the fact that the loss function is chosen in a data adaptive fashion, during the estimation process, thus capturing the shape and skewness of the underlying distribution of the errors. The asymptotic properties of the proposed estimator are established. Through an extensive Monte Carlo simulation study, the proposed approach was shown to perform as well if not better than several existing techniques. In particular, these studies show that the proposed method generally performs better than these existing techniques when the error distribution is heavy-tailed and/or skewed. The strengths of the proposed method were further exhibited through the analysis of two motivating data sets. To further disseminate this work, code (written in R) which implements the proposed methodology has been prepared and is available upon request.

Acknowledgments

This work was partially supported by Grant R01 AI121351 from the National Institutes of Health.

Appendix

The regularity conditions under which consistency and asymptotic normality of the proposed M-estimator and the bootstrapped estimator can be established are provided below.

A1: $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ is an i.i.d. sequence of random variables.

A2: The conditional cumulative distribution function of $y|\boldsymbol{x}$ is absolutely continuous and has a positive density denoted by $f^*(\cdot|\cdot)$.

A3: The parameter space $\Theta \subset \Xi \equiv \{\boldsymbol{\theta}|\alpha > 0, \sigma > 0, \tau \in (0,1), \beta_i \in \mathbb{R}, \forall i\}$, and Θ is a compact set.

A4: There exists a unique $\boldsymbol{\theta}_0 \in \Theta$ such that $E\{\ln f(y|\boldsymbol{x}, \boldsymbol{\theta})\}$ is maximized and has nonsingular second derivative at $\boldsymbol{\theta}_0$.

A5: $\|n^{-1}\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{\theta}})\| = o_p(n^{-1/2})$.

A6: There exists a δ such that $0 < \delta < \underline{\alpha} - 1$, $E\left(|\epsilon|^{2\bar{\alpha}+4\delta}\right) < \infty$, and $E\left(|x_j|^{2\bar{\alpha}+4\delta}\right) < \infty$, for $j = 1, ..., p$, where $\underline{\alpha}$ and $\bar{\alpha}$ denote the infimum and supremum of the set of all α in Θ, respectively.

A7: $\boldsymbol{V}_{1\boldsymbol{\theta}_0}$ and $\boldsymbol{V}_{2\boldsymbol{\theta}_0}$ exist and are finite, $\boldsymbol{V}_{1\boldsymbol{\theta}_0}$ is positive definite, and $\boldsymbol{V}_{2\boldsymbol{\theta}_0}$ is invertible.

Conditions A1-A5 and A7 are common in the literature, see Huber (1967), Huber and Ronchetti (2009), Koenker (2005), and Bera et al. (2016). Condition A2 restricts the conditional distribution of the dependent variable and Condition A7 is assumed so that the asymptotic variance of the estimator exists and is finite. Condition A4 ensures identifiability and existence of a unique solution. Condition A5 is used to ensure that the derivative of the loss function evaluated at $\hat{\boldsymbol{\theta}}$ is "nearly-zero". Condition A6 restricts the absolute moments on the conditional distribution of $y|\boldsymbol{x}$ and each covariate x_j in order to ensure the asymptotic behavior of the proposed estimator.

The consistency and asymptotic normality of the proposed estimator and the bootstrapped estimator are established by verifying the conditions in Huber and Ronchetti (2009) (page 127 for consistency and Theorem 6.6 as well as its corollary for normality) and Arcones and Giné (1992) (page 42), respectively. The arguments to verify these assumptions are similar to those in Zhu (2009) and Bera et al. (2016), and the details of these arguments are available from the corresponding author.

Chapter 3

Adaptive Penalized Quantile Regression for Variable Selection

3.1 Introduction

Suppose that $(\boldsymbol{x}_i', y_i)_{i=1}^n$ represents an independent and identically distributed (iid) sample which is assumed to follow a joint distribution $g(y, \boldsymbol{x})$, with y being a response variable and $\boldsymbol{x}$ being a p-dimensional covariate vector. In many settings, the conditional mean of y given $\boldsymbol{x}$ is assumed to be $\beta_0 + \boldsymbol{x}'\boldsymbol{\beta}$. In general, some of the covariates are important and others are not; in some studies e.g. genomics/genetics it is the case that most are unimportant. In these situations, variable selection techniques can be used to attempt to identify important variables, which is equivalent to identify which regression coefficients are non-zero. In general, two evaluation metrics are concerned: the mean squared error (MSE) or the mean absolute error (MAE) of the regression coefficients estimates to measure the accuracy of the estimates and the average number of correctly specified and mis-specified unimportant covariates to assess the variable selection performance. Both metrics are related to the accuracy of prediction that takes both the bias and the variability of the estimates into account. The following literature review gives a brief introduction of several methods which have been proposed to perform regression coefficients estimation and variable selection simultaneously.

In practice, the true density $g(y, \boldsymbol{x})$ is rarely known, so a parametric model $f_{\boldsymbol{\theta}}(y, \boldsymbol{x})$ is assumed, where $\boldsymbol{\theta}$ denotes the collection of unknown parameters of the model. Consider nested

models, where part of $\boldsymbol{\theta}$ is allowed to be any sub-vector of a p dimensional "slope" coefficient vector with coefficients corresponds to covariates being potentially related to the response, then a working model class is constructed which consists of a set of competing models that is denoted by $\{\mathscr{M}_m : m = 1, \cdots, 2^p\}$. If the true model belongs to the assumed model class, from the perspective of Kullback-Leibler (KL) divergence defined in (2.5), Akaike (1974) shows that up to a constant, the estimated KL divergence for model $\mathscr{M}_m$ can be asymptotically expanded as $-l_n(\hat{\boldsymbol{\theta}}_m) + \sum_{j=1}^{p} I(\hat{\theta}_{m,j} \neq 0)$, where $l_n(\hat{\boldsymbol{\theta}}_m)$ is the log-likelihood function for model $\mathscr{M}_m$ evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_m$ for $m = 1, \cdots, 2^p$, and Schwarz (1978) proposes the BIC criterion from the Bayesian principle which yields $-l_n(\hat{\boldsymbol{\theta}}_m) + \ln(n)/2 \sum_{j=1}^{p} I(\hat{\theta}_{m,j} \neq 0)$ as the criterion for model selection. Both AIC and BIC choose a model that minimizes the penalized likelihood; i.e.,

$$\arg\min_{\boldsymbol{\theta}} -l_n(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_0, \tag{3.1}$$

for different values of λ, where $||\cdot||_0$ is called the L_0 norm that counts the number of non-zero components. Many traditional methods can amount to penalized least squares (LS) methods with different choices of λ when the Normal likelihood in (3.1) is used. This includes Mallow's C_p statistics in Mallows (1973) corresponds to $\lambda = 1$ and the adjusted R^2 is approximately equivalent to $\lambda = 1/2$. (See details in Fan and Lv, 2010). (3.1) can be solved when p is rather small via best subsets, but as p increases this problem quickly becomes unsolvable. In case when the number of covariates p is comparable to or larger than the sample size n, the computation of L_0 regularization is infeasible.

Tibshirani (1996) proposed the least absolute shrinkage and selection operator (Lasso) that is defined as follows,

$$\arg\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{x}_i' \boldsymbol{\beta})^2 \qquad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{3.2}$$

here t is a tunning parameter. If t is greater than or equal to the L_1 norm of the OLS estimator, then (3.2) is equivalent to the unconstrained problem, which yields the OLS estimator. For smaller values of t, Lasso shrinks the estimated coefficients toward the origin, and typically sets some of the coefficients to be identically equal to zero. The optimization problem described in (3.2) has its dual

form which is given by

$$\underset{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}(y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta})^2+\lambda\sum_{j=1}^{p}|\beta_j|. \tag{3.3}$$

Problems (3.2) and (3.3) are equivalent; that is, for a given $\lambda \geq 0$, there exists a $t \geq 0$ such that the two problems share the same solution, and vice versa, see Osborne et al. (2000) for derivation. Efron et al. (2004) proposes the least angle regression (LARS), a fast and efficient algorithm to produce the entire Lasso solution path. In fact, LARS is as computationally efficient as computing the OLS estimates. Zou (2006) proposed adaptive Lasso which is an adaptive version of Lasso, naturally assigns different weights to the coefficients in the L_1 penalty, that is defined as

$$\underset{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\{y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta}\}^2+\lambda\sum_{j=1}^{p}w_j|\beta_j|, \tag{3.4}$$

where w_j's are the adaptive weights. Zou (2006) shows that Lasso estimate (3.3) can be inconsistent in some scenarios, and a necessary condition for adaptive Lasso being consistent is derived. The adaptive Lasso enjoys the oracle properties (see Fan and Li, 2001), i.e., performs as well as if the true underlying distribution is given. The adaptive Lasso outperforms Lasso by making use of a natural idea that more penalty should be put on the unimportant covariates. One convenient and effective way to choose the weights is by taking $w_j = 1/|\hat{\beta}_{OLS,j}|$, and $\hat{\beta}_{OLS,j}$ denotes the OLS estimate. The smaller $|\hat{\beta}_{OLS,j}|$ is, the stronger penalty $1/|\hat{\beta}_{OLS,j}|$ will be put on coefficient β_j.

Lasso/Adaptive Lasso are expected to perform well under Normal errors since they can be regarded as penialized log-likelihood of Normal distributions. However, in practice, it is common to see heavy-tailed errors and contaminated distributions with outliers, in which case it is well known that OLS may fail to provide reliable estimates, due to its squared power on residuals that penalizes large residuals too much. LAD regression estimates the conditional median function, which has been shown to be robust with respect to the underlying distributions. In the seminal paper of Koenker and Bassett (1978), the idea of LAD is generalized, and quantile regression (QR) is introduced to estimate the conditional quantile function of the response. Li and Zhu (2008) considered variable selection through L_1 regularization in linear quantile regression models, i.e., QR Lasso, with the loss

function defined as follows,

$$\underset{\beta_0\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i-\beta_0-\boldsymbol{x}_i'\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j|. \tag{3.5}$$

QR Lasso and the corresponding adaptive version are successful tools to perform variable selection with robust property. However guidance on the selection of the quantile used has not yet been fully addressed. Zou and Yuan (2008) developed composite quantile regression (CQR) which takes a sum of quantile loss functions at preset quantiles. By making use of the strength of multiple QR to estimate the same slope coefficients across different quantiles, the CQR estimator is shown to lose less than 30% asymptotic relative efficiency with respect to LS for a collection of densities when the number of quantiles approaches ∞. The idea of CQR is further extended in Bradic et al. (2011) by taking a weighted linear combination of quantile loss functions; i.e., the weighted composite quantile regression (WCQR), which can deal with both symmetric and asymmetric distributions. In Bradic et al. (2011) the WCQR plus L_1 regularization as loss function is studied, that is defined as

$$\underset{\beta_1,\cdots,\beta_q\in\mathbb{R},\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\ \frac{1}{n}\sum_{k=1}^{q}\sum_{i=1}^{n}w_k\rho_{\tau_k}\{y_i-\beta_k-\boldsymbol{x}_i'\boldsymbol{\beta}\}+\lambda\sum_{j=1}^{p}w_j|\beta_j|, \tag{3.6}$$

where the weights $w_k's$ satisfy $\sum_{i=1}^{q}w_k=1$ and are selected by minimizing the asymptotic variance of the estimator. This provides an adaptive procedure that allows the data to determine how the weights are distributed to those preset quantiles.

However how many number of quantities should be used in the loss function remains an open question and increasing the number of quantiles does not necessarily improve the efficiency of WCQR estimator (see Sun et al., 2013). This motivates us to look for a new approach that selects the optimal quantile in a continuous manner. Since QR can be viewed as the maximum likelihood estimator (MLE) of the asymmetric Laplace distribution (ALD) when the parameter τ in ALD is fixed, estimating the MLE of ALD would allow to estimate the most "informative" quantile for QR. Thus in this work, we perform variable selection by focusing on adaptively selecting the optimal quantile of interest in a different manner than methods described above, which we propose a penalized likelihood method by simply making use of the ALD. The proposed method allows the data to dictate which quantile of interest should be used by estimating an additional parameter which assesses the skewness of the underlying distribution. So essentially the proposed method

selects the most "informative" quantile used in QR Lasso for variable selection, inherits the robust property of QR, and complements QR by improving the accuracy of regression coefficients estimates across various error structures.

The remainder of this Chapter is organized as follows. Section 3.2 presents the modeling assumptions, introduces ALD with adaptive L_1 regularization, provides details of a stable numerical algorithm to obtain the regression parameter estimates, and presents a method for selecting the tuning parameter. Then results of an extensive Monte Carlo simulation study are presented in Section 3.3. We apply the proposed procedure to a motivating data set in Section 3.4. Section 3.5 concludes with a summary discussion.

3.2 Methodology

3.2.1 Model Assumption

Consider the linear regression model

$$y_i = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \cdots, n, \tag{3.7}$$

$(y_i, \boldsymbol{x}_i')$'s are *iid* random variables, with y_i being the response variable, $\boldsymbol{x}_i$ a p dimensional vector of covariates, β_0 is the intercept parameter, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is a p dimensional vector of regression coefficients, and ϵ_i is the error term. Herein, the $\boldsymbol{x}_i$ and ϵ_i are assumed to be independent, thus equivalently, the conditional quantiles of y_i given $\boldsymbol{x}_i$ have the same "slope" coefficient $\boldsymbol{\beta}$ and only differ in the intercept. Next we introduce the ALD based on which the adaptive robust variable selection procedure is proposed.

3.2.2 Asymmetric Laplace distribution and Penalized Quantile Regression

The conditional distribution of $y_i|\boldsymbol{x}_i$ follows $ALD(\boldsymbol{x}'\boldsymbol{\beta}, \sigma, \tau)$ if, the probability density is given as

$$f(y_i|\boldsymbol{x}_i) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{\frac{\rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma}\right\}.$$

The well known Laplace (double exponential) distribution is hold as a special case when $\tau = 0.5$. The log-likelihood of the observed data $\{(y_i, \boldsymbol{x}_i'), i = 1, \ldots, n\}$ is given by

$$\frac{1}{n}\sum_{i=1}^{n} \ln f(y_i|\boldsymbol{x}_i; \sigma, \tau, \beta_0, \boldsymbol{\beta}) = \log\{\tau(1-\tau)\} - \log(\sigma) - \frac{\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta})}{\sigma}. \tag{3.8}$$

For fixed τ and $(\beta_0, \boldsymbol{\beta}')$, the maximizer $\hat{\sigma}$ of (3.8) can be solved as $\hat{\sigma} = \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta})/n$. Plug in $\hat{\sigma}$ back into (3.8) and the log-likelihood function is proportional to

$$Q(\beta_0, \boldsymbol{\beta}, \tau) = \frac{\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta})}{\tau(1-\tau)}. \tag{3.9}$$

MLEs $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}')$ and $\hat{\tau}$ can be solved by minimizing (3.9). Note that for any fixed τ, $\{\hat{\beta}_0(\tau), \hat{\boldsymbol{\beta}}'(\tau)\}$ is the QR estimate with quantile of interest being τ. Then $\hat{\tau}$ can be obtained by minimizing $Q(\hat{\beta}_0(\tau), \hat{\boldsymbol{\beta}}(\tau), \tau)$.

The minimizer $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}', \hat{\tau})$ of (3.9) is studied in Bera et al. (2016), which they denote $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}', \hat{\tau})$ as ZQR and showed the asymptotic properties of $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}', \hat{\tau})$ for error distributions that are not only within ALD class, but also for model mis-specifications, in which case τ_0 can be interpreted as a skewness measurement for the underlying distribution. Our proposed method is developed in the same vein, such that by the model assumption (3.7), we can make use of the log-likelihood of the ALD to estimate the same "slope" coefficients $\boldsymbol{\beta}$ by simultaneously estimating one additional parameter τ, together with an adaptive Lasso penalty on the "slope" coefficients, to perform variable selection with improved performance. In particular, we consider the following optimization problem,

$$\min_{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p, \tau \in (0,1)} \frac{1}{n} \frac{\sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta})}{\tau(1-\tau)} \qquad \text{subject to } \sum_{j=1}^{p} w_j|\beta_j| \le t, \tag{3.10}$$

where w_j is adaptive weight for β_j. Compared to QR Lasso, the difference in our proposed methodology is that we have one more parameter τ to estimate. For any $\boldsymbol{\beta}$ that is in the feasible region $\sum_{j=1}^{p} w_j|\beta_j| \le t$, the minimizer $\hat{\tau}$ can be solved analytically, that is,

$$\hat{\tau} = \left\{1 + \sqrt{\frac{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^+}{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^-}}\right\}^{-1},$$

where $x^+ = \max\{x, 0\}$ and $x^- = \min\{-x, 0\}$. So it is more interpretable from the analytical expression of $\hat{\tau}$ that $\hat{\tau}$ measures the skewness of the underlying distribution. If the underlying

distribution is symmetric then on average $\hat{\tau} \approx 0.5$ and $\hat{\tau} > 0.5$ for left skewed and $\hat{\tau} < 0.5$ for right skewed distribution respectively. As argued by Bera et al. (2016) that, $\hat{\tau}$ is the most "informative" quantile that determines the quantile of interests that should be performed in the QR so that more efficient and accurate estimates of $\hat{\boldsymbol{\beta}}$ can be obtained. In the following we present a detailed and stable algorithm for obtaining $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}', \hat{\tau})$ for any fixed t.

In step 1, for fixed τ, the optimization problem (3.10) is equivalent to the dual form of (3.5) with an adaptive weights $w_j's$ that assign different weights to the coefficients. We take $w_j = 1/|\tilde{\beta}_j|$, where $\tilde{\beta}$ are the unconstrained estimates of problem (3.10), which has been shown in Bera et al. (2016) to be a $\sqrt{n}$ consistent estimator of the true "slope" coefficients. Thus $\hat{\boldsymbol{\beta}}_\lambda(\tau)$ is obtained as

$$\hat{\boldsymbol{\beta}}_\lambda(\tau) = \underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \beta_0 - \boldsymbol{x}_i'\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|}. \tag{3.11}$$

Note that $|x| = \rho_\tau(x) + \rho_\tau(-x)$, so that (3.11) can be combined into a single quantile check loss function that is given as

$$\hat{\boldsymbol{\beta}}^*_\lambda(\tau) = \underset{\boldsymbol{\beta}^* \in \mathbb{R}^{p+1}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n+2p} \rho_\tau(y_i^* - \boldsymbol{x}_i^* \boldsymbol{\beta}^*),$$

where $y_i^* = y_i$ for $i = 1, \cdots, n$ and $y_i^* = 0$ for $n+1 \leq i \leq n+2p$, and $\boldsymbol{x}_i^*$ is the ith row of matrix $\boldsymbol{X}^*$, where $\boldsymbol{X}^*$ is defined as follows,

$$\boldsymbol{X}^* = \begin{bmatrix} 1 & & \boldsymbol{x}_1' & \\ \vdots & & \vdots & \\ 1 & & \boldsymbol{x}_n' & \\ 0 & -n\lambda/|\tilde{\beta}_1| & & \\ \vdots & & \ddots & \\ 0 & & & -n\lambda/|\tilde{\beta}_p| \\ 0 & n\lambda/|\tilde{\beta}_1| & & \\ \vdots & & \ddots & \\ 0 & & & n\lambda/|\tilde{\beta}_p| \end{bmatrix}$$

In step 2, perform a grid search over τ to identify $\hat{\tau}_\lambda$ which is minimizing $Q(\hat{\beta}_{0,\lambda}(\tau), \hat{\boldsymbol{\beta}}'_\lambda(\tau), \tau)$ in (3.9). As can be seen, the proposed objective function in (3.10) is convex with convex penalties

for any fixed τ, and it has a unique global minimizer $\hat{\boldsymbol{\beta}}_\lambda(\tau)$ rather than multiple possible local minimizers.

3.2.3 Tuning Parameter

Tuning is an important issue in practice. Finite sample performance of the estimate relies on a good tuning procedure. There are many commonly used methods, such as k-fold cross validation, generalized cross-validation (see Golub G., Heath M., Wahba G., 1979; Tibshirani, 1996), AIC, AIC_c (Akaike, 1973; Hurvich CM. and Tsai CL., 1989; Zou et al., 2007), and BIC (Zou et al., 2007; Wang et al., 2007; Lee et al., 2014) can be applied to select the optimal tuning parameter. See Shao (1997) for a discussion of asymptotic properties of these techniques. However, AIC and BIC type methods are more computationally efficient when compared to cross validation type techniques. Both AIC and BIC methods have a theoretical justification to be used to select the correct model asymptotically, and for finite sample since BIC implements larger penalty than AIC, BIC tends to promote more sparsity. In this work, we make use of the connection between our proposed approach to the ALD likelihood, and propose to adopt a BIC type criterion to select the best tuning parameter λ which is shown as follows:

$$BIC_\lambda = \log\left\{\frac{\sum_{i=1}^n \rho_\tau(y_i - \hat{\beta}_0 - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})}{\hat{\tau}(1-\hat{\tau})}\right\} + \frac{\log(n)\hat{df}(\lambda)}{2n}, \tag{3.12}$$

where $\hat{df}(\lambda)$ is the number of non-zero elements of $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}')$, which is shown by Zou (2007) to be an unbiased estimate of the dimensionality of the model. $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}')$ and $\hat{\tau}$ are the estimates for each fixed λ that are obtained from section 3.2.2. The optimal $\hat{\lambda}$ is selected which minimizes the proposed BIC_λ.

Numerically searching for optimal λ based on minimum BIC value requires a uniform upper bound for λ so that for each τ in the first step of the proposed algorithm presented in section 3.2.2, the grid search over λ walks through the minimum λ such that the QR lasso estimate is equivalent to the unconstrained QR estimate, to the maximu m λ such that all "slope" coefficients are penalized to 0. This uniform upper bound λ_{max} can be obtained by

$$\lambda_{max} = \max_{\tau\in(0,1)} |\lambda_{max}(\tau)|, \tag{3.13}$$

where $\lambda_{max}(\tau)$ is the upper bound for QR adaptive lasso when the quantile is τ, which can be obtained using the function *hqreg* in hqreg package (see Yi and Huang, 2016).

3.3 Simulation Results

In this section we summarize a simulation study to evaluate the finite sample performance of the proposed methodology denoted by PZQR, and to compare between the proposed approach and the existing methods including adaptive Lasso (ALA) (Zou, 2006) and adaptive QR Lasso denoted by $QR(\tau)$ (Zheng et al., 2013) with $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The adaptive weights w_j used are $1/|\tilde{\beta}_{OLS,j}|$ for adaptive Lasso; $1/|\tilde{\beta}(\tau)|$ for adaptive QR Lasso with quantile of interest being τ, where $\tilde{\beta}(\tau)$ are the QR estimate (Koenker and Barssett, 1978); $1/|\tilde{\beta}_{ZQR}|$ for the proposed approach, where $\tilde{\beta}_{ZQR}$ is the ZQR estimate (Bera et al., 2016). All methods use corresponding BIC criterion to select the tuning parameter. (See Zou, 2007 for BIC in Lasso and Lee et al., 2014 for BIC in QR Lasso.)

We adopt the following simulation setting. By setting $\beta_0 = 0$ and $\boldsymbol{\beta}$ is a 8×1 vector with $\boldsymbol{\beta} = (3.0, 2.0, 1.5, 0, 0, 0, 0, 0)'$. By doing this, we get a covariate vector with the first three components being important and the rest being redundant. The covariates $\boldsymbol{x}_i$ are generated from a multivariate normal distribution with $\mathbf{0}$ mean and covariance matrix $\boldsymbol{\Sigma}[i, j] = 0.5^{|i-j|}$, for $1 \leq i, j \leq 8$. And the response variables are generated by $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$. The ϵ_i's are generated from different distributions to examine the general performance of the proposed method. In particular, we consider the following six distributions: N(0,1); standard t distribution of degree freedom 4 (t_4); $ALD(0, \sigma^*, 0.25)$ denoted by ALD, where σ^* is the value such that the standard deviation of the ALD is 1; Skewed t distribution (see Fernandez and Steel, 1998) of degree freedom 3 and gamma of 2; two mixture models $0.1N(0, 9) + 0.9N(5, 1)$ and $0.7Laplace(0, 1) + 0.3Laplace(3, 1)$, respectively. These choices reflect various characteristics of error structures that are commonly seen in practical applications; to include symmetry, asymmetry, light tails, heavy tails and slight deviation from light/heavy-tailed distributions. For each of the above error distribution, $m = 1000$ independent data sets were generated, each consisting of $n = 200$ observations.

The results of "slope" coefficients estimates are summarized in Table 3.1 under these six error structures. As can be seen in Table 3.1 that the bias of all coefficients estimates are close to zero for all approaches. In terms of efficiency, except when the error distribution is Normal, adaptive

Lasso is most efficient as expected, the proposed estimator has smallest standard deviation for non-zero coefficients estimates across all other underlying distributions. The variable selection results are summarized in Table 3.2 which shows that all methods perform well on identifying significant/non-zero coefficients. One will note that when the underlying distribution is symmetric, i. e., $N(0,1)$ or t_4, the averaged $\hat{\tau}$ is very close to 0.5, and for other skewed distributions, $\hat{\tau}$ assesses the skewness of the underlying distributions. For example, in case of Laplace mixture as the error structure, the averaged $\hat{\tau} = 0.228$. In Table 3.1 it shows that under Laplace mixture error distribution, the proposed method has the smallest standard deviations of the non-zero coefficients estimates, and adaptive QR Lasso performs better for quantiles of interest being closer to 0.228, so that $QR(0.3)$ is the best among all selected quantiles, whereas the performance can drop drastically for mis-specified quantiles, e.g., $Q(0.9)$. Same conclusion can be drawn for all other error structures. This clearly shows the proposed method provides for more efficient estimates by capturing the skewness of the underlying distribution via estimating the skewness parameter τ, a feature not shared by QR Lasso procedures.

Moreover by further examining the performance of estimation accuracy, the MAE score, i.e., the L_1 norm of the difference between the "slope" regression coefficients estimates and the true, is computed for each sample. Then box plots of values of MAE score over 1000 repetition are reported, across all six error structures. The results are presented in Figure 3.1. It clearly shows that the proposed method generally attains smallest MAE over various error structures, except under Normal distribution, as expected, adaptive Lasso performs the best. Similar simulation results for the MSE score of the estimates are obtained which is omitted here.

3.4 Real data analysis

According to National Institute of Diabetes and Digestive and Kidney Diseases (2017), 29.1 million people or 9.3% of the U.S. population have diabetes. It is well known that many factors increase the risk for diabetes, e.g., obesity, lack of physical activity and generic predisposition (The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus, 2003). Our motivating data which was used in Efron et al. (2004) consists of 442 patients' information including ten covariates; i.e., the participants age (Age), sex (with 1 denoting female and 2 for male), body mass index (BMI), average blood pressure (BP), and six blood serum measurements (S1-S6). The quantity measurement

Table 3.1: Simulation results for coefficients estimates obtained by ALA, QR(τ) with $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and PZQR for six distributions. This summary provides for the averaged estimate for each coefficient and the standard deviation of the estimates in parenthesis.

Error	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
$N(0,1)$									
	ALA	2.996	1.998	1.497	0.000	0.001	-0.001	0.001	-0.001
		(0.082)	(0.089)	(0.082)	(0.023)	(0.020)	(0.015)	(0.021)	(0.020)
	QR(0.1)	2.984	1.977	1.417	0.000	0.000	0.000	0.000	0.000
		(0.139)	(0.165)	(0.151)	(0.006)	(0.005)	(0.003)	(0.004)	(0.004)
	QR(0.3)	2.984	1.976	1.431	0.000	0.000	0.000	0.000	0.000
		(0.109)	(0.126)	(0.116)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.5)	2.982	1.976	1.434	0.000	0.000	0.000	0.000	0.000
		(0.106)	(0.120)	(0.112)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.7)	2.986	1.974	1.431	0.000	0.000	0.000	0.000	0.000
		(0.113)	(0.128)	(0.123)	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)
	QR(0.9)	2.984	1.976	1.423	0.000	0.000	0.000	0.000	0.000
		(0.149)	(0.166)	(0.156)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
	PZQR	2.993	1.990	1.484	0.000	0.000	-0.001	0.000	0.000
		(0.103)	(0.114)	(0.107)	(0.023)	(0.019)	(0.017)	(0.016)	(0.017)
t_4									
	ALA	2.993	1.993	1.492	0.001	-0.002	-0.002	0.002	0.000
		(0.118)	(0.137)	(0.123)	(0.030)	(0.036)	(0.037)	(0.037)	(0.035)
	QR(0.1)	2.971	1.963	1.382	0.001	0.000	0.001	0.000	0.000
		(0.219)	(0.271)	(0.249)	(0.015)	(0.017)	(0.023)	(0.013)	(0.010)
	QR(0.3)	2.984	1.977	1.421	0.000	0.000	0.000	0.000	0.000
		(0.127)	(0.157)	(0.138)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.5)	2.981	1.980	1.426	0.000	0.000	0.000	0.000	0.000
		(0.112)	(0.137)	(0.122)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.7)	2.982	1.975	1.419	0.000	0.000	0.000	0.000	0.000
		(0.124)	(0.152)	(0.137)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.9)	2.972	1.947	1.387	0.000	0.000	0.000	0.001	-0.001
		(0.213)	(0.261)	(0.243)	(0.028)	(0.012)	(0.021)	(0.015)	(0.017)
	PZQR	2.992	1.995	1.474	0.000	0.000	-0.001	-0.001	0.000
		(0.110)	(0.129)	(0.116)	(0.021)	(0.015)	(0.018)	(0.017)	(0.023)
ALD									
	ALA	3.001	1.993	1.498	0.000	0.000	0.001	-0.001	0.000
		(0.083)	(0.091)	(0.079)	(0.019)	(0.018)	(0.017)	(0.018)	(0.021)
	QR(0.1)	2.989	1.987	1.463	0.000	0.000	0.000	0.000	0.000
		(0.079)	(0.090)	(0.078)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.3)	2.992	1.987	1.469	0.000	0.000	0.000	0.000	0.000
		(0.055)	(0.061)	(0.056)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.5)	2.988	1.981	1.451	0.000	0.000	0.000	0.000	0.000
		(0.078)	(0.091)	(0.082)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.7)	2.985	1.972	1.423	0.000	0.000	0.000	0.000	0.000
		(0.119)	(0.138)	(0.123)	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)
	QR(0.9)	2.995	1.939	1.364	0.003	0.001	0.000	-0.001	-0.001
		(0.234)	(0.299)	(0.255)	(0.039)	(0.023)	(0.014)	(0.024)	(0.022)
	PZQR	2.995	1.997	1.493	0.000	0.000	0.000	0.000	0.000
		(0.054)	(0.058)	(0.054)	(0.004)	(0.003)	(0.002)	(0.003)	(0.005)

(*continued*)

Table 3.1 continued: Here Norm Mix and Lap Mix denote the Normal mixture and Laplace mixture described in simulation section respectively.

Error	Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
Skewed t									
	ALA	2.992	1.995	1.469	-0.001	0.000	-0.001	0.000	0.000
		(0.169)	(0.207)	(0.193)	(0.065)	(0.051)	(0.048)	(0.042)	(0.042)
	QR(0.1)	2.981	1.975	1.427	0.000	0.000	0.000	0.000	0.000
		(0.124)	(0.148)	(0.135)	(0.000)	(0.003)	(0.000)	(0.000)	(0.005)
	QR(0.3)	2.975	1.978	1.411	0.000	0.000	0.000	0.000	0.000
		(0.132)	(0.161)	(0.149)	(0.004)	(0.002)	(0.000)	(0.000)	(0.003)
	QR(0.5)	2.968	1.972	1.378	0.000	0.000	0.000	0.000	0.000
		(0.165)	(0.203)	(0.190)	(0.000)	(0.000)	(0.007)	(0.004)	(0.000)
	QR(0.7)	2.968	1.956	1.330	0.000	0.001	0.000	0.000	0.000
		(0.235)	(0.295)	(0.277)	(0.017)	(0.010)	(0.016)	(0.004)	(0.008)
	QR(0.9)	2.974	1.921	1.225	0.014	0.011	0.002	0.003	-0.006
		(0.464)	(0.614)	(0.555)	(0.137)	(0.113)	(0.092)	(0.096)	(0.095)
	PZQR	2.984	1.993	1.464	0.000	0.001	-0.001	0.001	0.000
		(0.118)	(0.136)	(0.123)	(0.024)	(0.029)	(0.026)	(0.025)	(0.026)
Norm Mix									
	ALA	2.996	1.982	1.474	-0.002	0.000	0.000	0.001	-0.002
		(0.172)	(0.196)	(0.187)	(0.058)	(0.053)	(0.044)	(0.051)	(0.052)
	QR(0.1)	2.988	1.848	1.241	0.020	0.012	-0.001	-0.006	0.000
		(0.569)	(0.738)	(0.670)	(0.195)	(0.169)	(0.165)	(0.155)	(0.155)
	QR(0.3)	2.982	1.974	1.410	0.000	0.000	0.000	0.000	0.000
		(0.144)	(0.168)	(0.161)	(0.001)	(0.000)	(0.003)	(0.000)	(0.000)
	QR(0.5)	2.984	1.972	1.420	0.000	0.000	0.000	0.000	0.000
		(0.119)	(0.141)	(0.130)	(0.000)	(0.002)	(0.004)	(0.000)	(0.000)
	QR(0.7)	2.983	1.978	1.422	0.000	0.000	0.000	0.000	0.000
		(0.121)	(0.142)	(0.130)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.9)	2.973	1.981	1.417	0.000	0.000	0.000	0.000	0.000
		(0.157)	(0.182)	(0.164)	(0.003)	(0.007)	(0.005)	(0.003)	(0.005)
	PZQR	2.992	1.987	1.465	0.001	-0.001	0.000	0.000	0.000
		(0.118)	(0.136)	(0.127)	(0.028)	(0.022)	(0.018)	(0.020)	(0.019)
Lap Mix									
	ALA	2.993	1.989	1.476	0.001	0.000	0.002	0.001	0.000
		(0.177)	(0.199)	(0.185)	(0.042)	(0.043)	(0.048)	(0.050)	(0.050)
	QR(0.1)	2.990	1.979	1.442	0.000	0.000	0.000	0.000	0.000
		(0.112)	(0.129)	(0.116)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
	QR(0.3)	2.988	1.986	1.449	0.000	0.000	0.000	0.000	0.000
		(0.088)	(0.097)	(0.090)	(0.000)	(0.000)	(0.003)	(0.000)	(0.000)
	QR(0.5)	2.980	1.972	1.405	0.000	0.000	0.000	0.000	0.000
		(0.151)	(0.181)	(0.167)	(0.000)	(0.000)	(0.005)	(0.000)	(0.006)
	QR(0.7)	2.955	1.938	1.317	0.002	0.000	0.002	-0.001	0.000
		(0.282)	(0.356)	(0.323)	(0.027)	(0.022)	(0.023)	(0.022)	(0.033)
	QR(0.9)	2.960	1.884	1.237	0.022	0.005	0.000	0.004	-0.003
		(0.513)	(0.647)	(0.621)	(0.158)	(0.118)	(0.098)	(0.106)	(0.095)
	PZQR	2.995	1.997	1.482	0.000	0.000	0.000	0.000	0.000
		(0.076)	(0.083)	(0.077)	(0.000)	(0.000)	(0.008)	(0.003)	(0.008)

Table 3.2: Variable selection results for six error distributions. This summary includes the average number of correctly specified zero coefficients (Correct), the average number of incorrectly specified zero coefficients (Wrong), and the averaged $\hat{\tau}$. Here Norm Mix and Lap Mix denote the Normal mixture and Laplace mixture described in simulation section respectively.

		Mean of No. of 0		
Error	Method	Correct	Wrong	$\hat{\tau}$
$N(0,1)$	LS-Lasso	4.927	0.000	-
	QR(0.1)-Lasso	4.927	0.000	-
	QR(0.3)-Lasso	4.997	0.000	-
	QR(0.5)-Lasso	4.999	0.000	-
	QR(0.7)-Lasso	4.997	0.000	-
	QR(0.9)-Lasso	4.942	0.000	-
	ZQR-Lasso	4.914	0.000	0.496
t_4	LS-Lasso	4.912	0.000	-
	QR(0.1)-Lasso	4.827	0.000	-
	QR(0.3)-Lasso	5.000	0.000	-
	QR(0.5)-Lasso	5.000	0.000	-
	QR(0.7)-Lasso	5.000	0.000	-
	QR(0.9)-Lasso	4.853	0.000	-
	ZQR-Lasso	4.933	0.000	0.500
ALD	LS-Lasso	4.936	0.000	-
	QR(0.1)-Lasso	4.990	0.000	-
	QR(0.3)-Lasso	5.000	0.000	-
	QR(0.5)-Lasso	5.000	0.000	-
	QR(0.7)-Lasso	4.993	0.000	-
	QR(0.9)-Lasso	4.751	0.000	-
	ZQR-Lasso	4.987	0.000	0.245
Skewed t	LS-Lasso	4.899	0.000	-
	QR(0.1)-Lasso	4.978	0.000	-
	QR(0.3)-Lasso	4.997	0.000	-
	QR(0.5)-Lasso	4.998	0.000	-
	QR(0.7)-Lasso	4.934	0.001	-
	QR(0.9)-Lasso	4.393	0.047	-
	ZQR-Lasso	4.902	0.000	0.198
Norm Mix	LS-Lasso	4.889	0.000	-
	QR(0.1)-Lasso	4.484	0.119	-
	QR(0.3)-Lasso	4.999	0.000	-
	QR(0.5)-Lasso	4.987	0.000	-
	QR(0.7)-Lasso	4.987	0.000	-
	QR(0.9)-Lasso	4.935	0.000	-
	ZQR-Lasso	4.933	0.000	0.720
Lap Mix	LS-Lasso	4.900	0.000	-
	QR(0.1)-Lasso	4.997	0.000	-
	QR(0.3)-Lasso	4.999	0.000	-
	QR(0.5)-Lasso	4.997	0.000	-
	QR(0.7)-Lasso	4.900	0.000	-
	QR(0.9)-Lasso	4.359	0.065	-
	ZQR-Lasso	4.993	0.000	0.228

Figure 3.1: Box plot of the MAE scores across six error structures for different methods. Here Norm Mix and Lap Mix denote Normal mixture and Laplace mixture described in simulation section respectively.

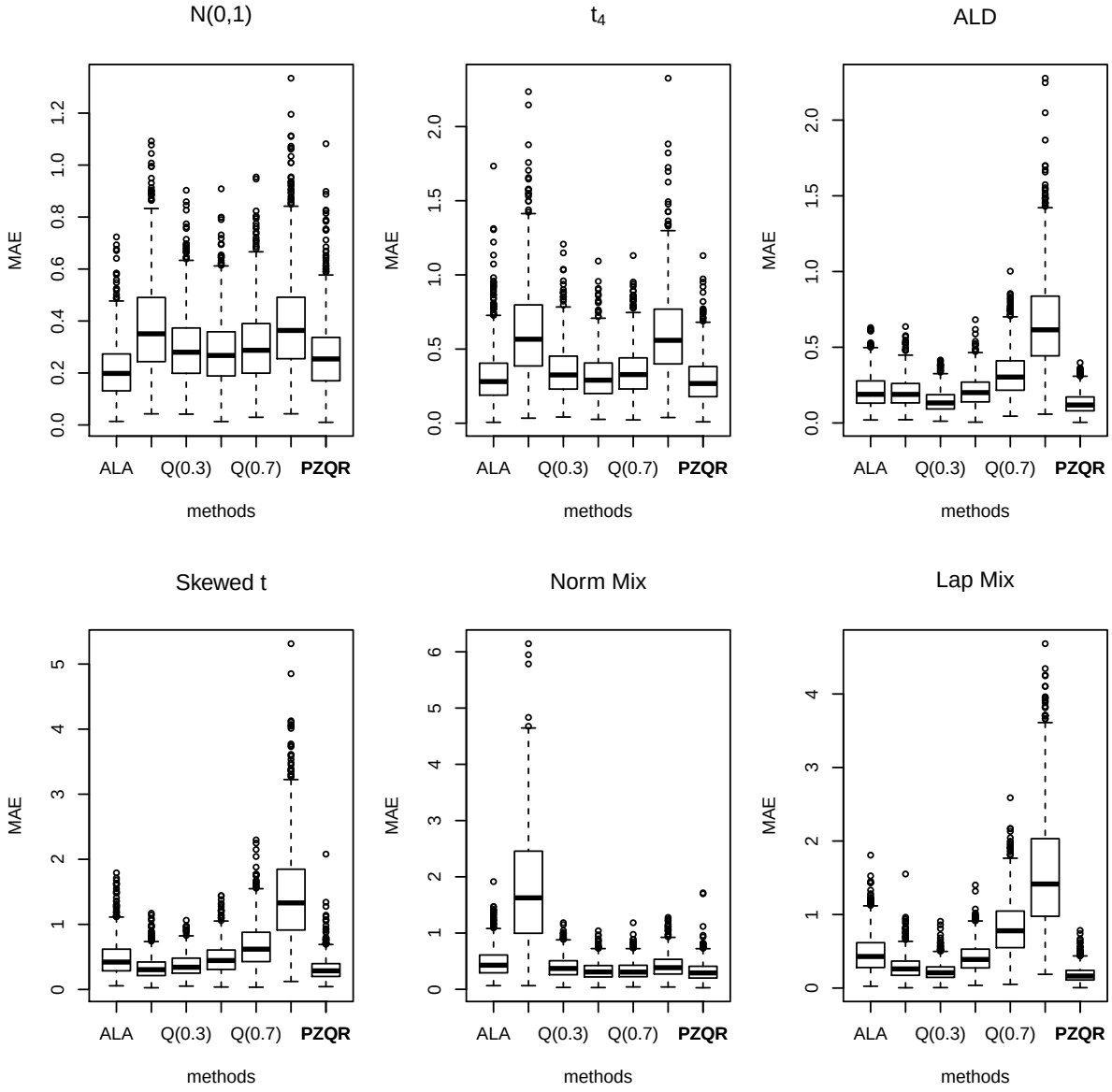


Table 3.3: Coefficients Estimates Result

Methods	Age	Sex	BMI	BP	S1	S2	S3	S4	S5	S6
ALA	0.000	-0.125	0.334	0.194	-0.318	0.166	0.000	0.074	0.422	0.000
QR(0.1)	0.000	0.000	0.132	0.000	0.000	0.000	0.000	0.094	0.198	0.000
QR(0.3)	0.000	-0.029	0.368	0.099	0.000	-0.046	0.000	0.032	0.319	0.000
QR(0.5)	0.000	-0.101	0.326	0.159	-0.167	0.000	0.000	0.018	0.451	0.000
QR(0.7)	0.000	-0.071	0.388	0.132	-0.122	0.000	0.000	0.000	0.494	0.000
QR(0.9)	0.000	-0.048	0.378	0.158	-0.372	0.241	0.000	0.000	0.479	0.000
PZQR	0.000	-0.146	0.303	0.214	-0.204	0.010	0.000	0.109	0.398	0.000

of diabetes progression, which is taken one year after baseline, is the response variable. All covariates and the response variable have been standardized so that the mean is 0 and the standard deviation is 1. This analysis assumes a linear model and uses the proposed methodology as well as ALA and $QR(\tau)$ with the quantile $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, to perform regression coefficients estimation and variable selection simultaneously. All techniques were implemented in the exact same way as was described in Section 3.3.

Table 3.3 reports the regression coefficients estimates and variable selection result. One will note that almost all coefficients estimates across all methods are close to each other, both in magnitude and sign. QQ-plots of the residuals obtained under ALA method indicates this data is close to be normally distributed, and only the proposed approach has the same variable selection result with ALA, that both methods select Sex, BMI, BP, S1,S2, S4 and S5 as important variables. All other QR Lasso approaches will miss either one or several important predictors. The proposed approach delivers $\hat{\tau} = 0.49$, which is consistent with the finding of QQ-plots that the data is close to be symmetric.

3.5 Conclusions

This work has developed an adaptive procedure for variable selection by making use of the log-likelihood of the ALD. In particular, the proposed methodology is robust with respect to the underlying distributions meaning that it can select important predictors under sparse setting when the underlying error structure is symmetric, but can also render reliable estimation and inference for asymmetric error structures with light/heavy tails. This is achieved by the fact that the proposed procedure essentially chooses the loss function in a data driven fashion, such that the additional

parameter estimate $\hat{\tau}$ can assess the skewness of the underlying distribution, which determines the quantile of interest used in the adaptive QR Lasso procedure. Through an extensive Monte Carlo simulation study, the proposed approach is shown to perform as well as the adaptive Lasso and the adaptive QR Lasso procedures in terms of identifying important predictor variables, and from the estimation perspective, the proposed method provides for more accurate estimates across a broad class of distributions which reflect heavy/light-tailed, symmetric/left skewed/right skewed distributions. The strengths of the proposed method are further exhibited through the analysis of one motivating data set. The code (written in R) implements the proposed methodology has been prepared and is efficient and stable to obtain the estimates, which is available upon request.

Chapter 4

Penalized Autocorrelation Estimation in Time Series

4.1 Introduction

Consider a simple case that an independent and identically distributed (*iid*) sample, $\{x_t\}$, $t = 1, \cdots, n$, is given and the sample mean $\bar{x}$ can be used to construct the $1-\alpha$ confidence interval for $\mu = E(X_t)$ which is given as

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{\alpha/2},$$

where s is the sample standard deviation and $t_{\alpha/2}$ is the critical value of Student t distribution with degree freedom of $n-1$. However this often leads to an empirical coverage that is smaller than the nominal level due to the underestimation of the standard error of $\bar{x}$. The usual assumption of independence is not valid for correlated data set in time series, and often positive autocorrelation exists in time series data with the autocorrelation being stronger for closer observations, in which case the actual standard error of $\bar{x}$ is larger than $s/\sqrt{n}$. Discussions about effect of ignoring autocorrelations on producing reliable estimates and type one error in hypothesis testing can found in Cochrane and Orcutt (1949).

The correlation between the two random variables X_t and X_{t+h} in a stationary time series

is called *autocorrelation* and is defined as

$$\rho(h) = \frac{Cov(X_t, X_{t+h})}{Var(X_t)}.$$

Of course, the issue brought by the correlated error term would also extend to the linear regression model for estimating regression coefficients. Test statistics such as Durbin-Watson test in Durbin and Watson (1950) is to detect the presence of autocorrelation and if the conclusion of this hypothesis test indicates the existence of correlation then transformed regression techniques will be applied to test the significance of the regression coefficient. Nakamura and Nakamura (1978) show in presence of correlation, the hypothesis test could yield a type I error much larger than the significance level, e.g., the significance level of $\alpha = 5\%$ hypothesis test for significance of "slope" coefficient brings out a type I error of 40%, with the autocorrelation at lag one $\rho = 0.9$ when the sample size is $n = 30$. The situation improves as sample size increases, but deteriorates as ρ increases.

Sample autocorrelation is used to estimate the population autocorrelation and is defined as

$$\hat{\rho}(h) = \frac{\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2} \tag{4.1}$$

From Cauchy Schwartz Inequality, it is easy to see that $|\hat{\rho}(h)| \leq 1$ at any lag $h \geq 0$. $\hat{\rho}(h)$ is a consistent estimate of $\rho(h)$ and asymptotic normality holds under a mild restriction on stationary linear stochastic process (see Anderson and Walker, 1964). However, for finite samples, $\hat{\rho}(h)$ tends to under estimate $\rho(h)$ in many scenarios and much work has been done to derive explicit approximations for the bias (see Lomnicki and Zaremba, 1957). The analytical expressions of the serial correlation coefficient studied in Marriott and Pope (1954) are derived for two simple stationary processes: the unweighted moving average model of order 1 (UMA(1)), i.e., $X_t = W_{t-1} + W_t$ and the autoregressive model of order 1 (AR(1)), i.e., $X_t = \rho X_{t-1} + W_t$, where W_t is the standard normally distributed white noise, for $t = 1, \cdots, n$. Since $\hat{\rho}(h)$ and the serial correlation coefficient differs by terms of order $1/n^2$, $E\{\hat{\rho}(h)\}$ can be derived and $E\{\hat{\rho}(1)\}$ is given as follows,

$$E\{\hat{\rho}(1)\} = \frac{1}{2} - \frac{2}{n} + O(\frac{1}{n^2}), \qquad \text{for UMA(1)},$$

$$E\{\hat{\rho}(1)\} = \rho - \frac{1+4\rho}{n} + O(\frac{1}{n^2}), \qquad \text{for AR(1)},$$

These results clearly show under these two stationary time series to the order of $1/n^2$, $\hat{\rho}(1)$ underestimates $\rho(1)$ (when $\rho < -0.25$ under AR(1), $\hat{\rho}(1)$ underestimate ρ in absolute sense), the negative bias becomes more severe for smaller sample size and larger ρ. The bias-corrected estimation for ρ becomes possible based on the expectation and Bence (1995) applied the bias-corrected ρ estimate for AR(1) model to adjust the standard error estimate of $\bar{x}$ by a correction factor k such that $E(ks)$ is equals to the standard deviation of $\bar{x}$. An approximate correction factor $k = \{(1+\rho)/(1-\rho)\}^{1/2}$ from which we can see finite sample inference performances depend on the accuracy of ρ estimation.

The underestimation is common to many techniques , e.g., two-step Cochrane-Orcutt, the Durbin estimator, or maximum-likelihood estimation (Park and Mitchell, 1980; Beesley and Griffiths, 1982; Griffiths and Beesley, 1984; King and Giles, 1984). Methods of removing the bias include the test statistics $R = 2r - (r_1 + r_2)/2$ in Quenouille (1949), where r is the sample correlation estimate for the whole series and r_1 and r_2 are for the first and second halves respectively. This procedure reduces the bias to the order of $1/n^2$ but loses efficiency.

In this work we propose a methodology that aims to provide for estimation of autocorrelation with reduced bias and small standard deviation. In particular, the population autocorrelation estimation is obtained as a LS problem from the best linear predictor perspective, and regularization is applied to the parameter of the objective function in order to penalize the autocorrelation estimate being away from ± 1. The proposed procedure is expected to "drag" the autocorrelation estimate toward ± 1 so that it alleviates the underestimation issue. The tuning parameter is chosen in a data driven manner in order to obtain a good performance of the proposed approach in terms of mean squared error (MSE). This chapter is organized as follows. Section 4.2 presents model assumptions, the proposed loss function and the sequence for tuning parameter. Asymptotic properties of the proposed method is established in Section 4.3. Extensive Monte Carlo simulations under AR(1), AR(2), ARMA(1,1) models are studied in section 4.4. Suggestions for better choice of tuning parameter is discussed in section 4.5.

4.2 Methodology

Consider a linear stochastic process defined as

$$X_t = \mu + \sum_{i=0}^{\infty} \psi_i W_{t-i}, \qquad \text{for } t = 1, \cdots, n, \tag{4.2}$$

where $\{W_t\} \sim WN(0, \sigma_w^2)$ is *iid* white noise with zero mean and finite variance, and μ, ψ_i are parameters satisfying $\sum_{i=-\infty}^{\infty} |\psi_i| < \infty$. Note that many commonly seen time series models can be represented as linear process. For example, If $\psi_0 = 1$ and $\psi_1 = \theta$, then the linear process is MA(1) with parameter θ. If $\psi_i = \phi^i$ for $i \geq 0$, then it corresponds to AR(1) with parameter ϕ. In general ARMA models have a linear process representation. Bickel (1996) showed the closure of the linear process under a suitable metric is unexpectedly large, and this work deals with estimation of autocorrelation under linear stochastic models and all of our simulated time series belong to this class.

As we know autocorrelation at lag h can be obtained by the following:

$$\rho(h) = \underset{\rho \in (-1,1)}{\arg\min}\, E\{(X_{t+h} - \mu) - \rho(X_t - \mu)\}^2, \tag{4.3}$$

where $\mu = E(X_t)$. This motivates us to propose a penalized LS procedure, that is, for a realization $\{x_t\}$, $t = 1, \cdots, n$ from the linear process, the loss function is given

$$\sum_{t=1}^{n} \left\{(x^*_{t+h} - \bar{x}) - \rho(x^*_t - \bar{x})\right\}^2 + n\lambda_n(1 - \rho^2), \tag{4.4}$$

where $\lambda_n \geq 0$ is the tuning parameter and x^* is the augmented data set so that $x^*_i = x_i$, for $i = 1, 2, \cdots, n$, and $x^*_i = \bar{x}$ for $i = n+1, \cdots, 2n-1$. The minimizer of the proposed objective function is given

$$\tilde{\rho}(h) = \frac{\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2 - n\lambda_n}. \tag{4.5}$$

When the tuning parameter $\lambda_n = 0$, i.e., no penalization is taken, (4.4) as an empirical analogue of the objective function in (4.3) leads to the sample autocorrelation (4.1). The larger λ_n is, the more penalty is put on ρ being away from ± 1. When $\lambda_n = \infty$, $\tilde{\rho}(h) = \pm 1$. Therefore, selection of λ_n is critical for the performance of the proposed estimate $\tilde{\rho}(h)$. The underestimation of ρ becomes an issue for small sample size, and deteriorates as the absolute value of ρ gets larger. The selection of λ_n should take these into account. We propose $\lambda_n = f(x_1, \cdots, x_n)a_n$, where $f(x_1, \cdots, x_n)$ reflects the penalization depends on ρ through the data, and $a_n \to \infty$. The proposed λ_n in this work to

achieve these characteristics is given as follows,

$$\lambda_n = \{\hat{\gamma}(0) - |\hat{\gamma}(h)|\}n^{-0.5-10^{-6}}, \tag{4.6}$$

where $\hat{\gamma}(h) = \sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(x_t - \bar{x})/n$ for any $h \geq 0$ is a consistent estimate of $\gamma(h) = Cov(X_{t+h}, X_t)$. The proposed λ_n is derived from the restriction that $|\tilde{\rho}(h)| \leq 1$ so that it is a valid autocorrelation estimate, and is chosen so that $\tilde{\rho}(h)$ and $\hat{\rho}(h)$ have the same asymptotic behavior. After simple algebra, the proposed estimator is given as

$$\tilde{\rho}(h) = \frac{\hat{\rho}(h)}{1 - (1 - |\hat{\rho}(h)|)n^{-0.5-10^{-6}}} \tag{4.7}$$

4.3 Asymptotics

Theorem 4.3.1. *For any stationary linear stochastic process defined in (4.2), if* $\sum_{i=-\infty}^{\infty} |i|\psi_i^2 < \infty$, *and* $\sqrt{n}a_n \to 0$ *as* $n \to \infty$, *then for* $\lambda_n = f(x_1, \cdots, x_n)a_n$ *with stochastically bounded f,* $\tilde{\rho}(h)$ *as defined in (4.5) is asymptotically normal; i.e.,*

$$\sqrt{n}\{\tilde{\rho}(h) - \rho(h)\} \xrightarrow{d} N(0, V_h), \tag{4.8}$$

for $-n < h < n$, where $V_h = \sum_{i=-\infty}^{\infty}\{\rho(i)^2 + \rho(i)\rho(i+2h) + 2\rho(h)^2\rho(i)^2 - 4\rho(h)\rho(i)\rho(i+h)\}$. The proof of the theorem is straightforward by using the consistency and asymptotic normality of sample autocorrelation and applying slutsky's theorem. The proposed λ_n in (4.6) satisfies the condition so that $\tilde{\rho}(h)$ defined in (4.7) is asymptotically normal.

4.4 Simulation Study

In order to examine the finite performance of the proposed estimator, a simulation study is conducted to generate stationary time series from AR(1) and ARMA(1,1) model:

- $X_t = \phi X_{t-1} + W_t, \qquad \phi = 0.1, 0.3, 0.5, 0.7, 0.9$
- $X_t = \phi X_{t-1} + \theta W_{t-1} + W_t, \qquad \phi = -0.8, -0.4, 0, 0.4, 0.8$ and $\theta = 1$.

$\rho = 0.1, 0.3, 0.5, 0.7, 0.9$ are considered for time series from both models. For each parameter setting, sample size is taken as $n = 25, 50, 100$. The comparison of the performance is made between the sample autocorrelation at lag one $\hat{\rho}$ and the proposed autocorrelation estimate at lag one $\tilde{\rho}$. 1000 independent dataset for each setting are generated to evaluate the performance and the result is shown in Table 4.1, which provides a summary of the estimators from near non-correlated ($\rho = 0.1$) to highly correlated ($\rho = 0.9$) data sets with small, moderate and large sample sizes. Simulation results for negative ρ are similar and for simplicity we omit the results. In particular, this summary includes the empirical bias, standard deviation of the estimates, and the relative mean squared error (RMSE) i.e., the averaged MSE of the sample autocorrelation divided by the averaged MSE of the proposed estimator. From these results, one will note that the estimates of autocorrelation are subject to quite large biases, especially for small sample size and high correlations. The proposed estimator reduces the bias for all cases and generally has smaller MSE when $\rho \geq 0.5$.

Furthermore, we are interested in using the proposed approach to estimate the partial autocorrelation. In time series X_t, the partial autocorrelation $\alpha(h)$ is defined as the autocorrelation between X_t and X_{t+h} that is not accounted by lags from 1 to $h-1$. Sample partial autocorrelation is defined as

$$\hat{\alpha}(h) = \begin{pmatrix} 1 & \cdots & \hat{\rho}(h-1) \\ \vdots & \ddots & \vdots \\ \hat{\rho}(h-1) & \cdots & 1 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\rho}(1) \\ \vdots \\ \hat{\rho}(h) \end{pmatrix},$$

which is a function of sample autocorrelation. Partial autocorrelation plays an important role in identifying the extent of the lag in $AR(p)$ models. The proposed approach in estimating autocorrelation $\tilde{\rho}$ leads to the partial autocorrelation estimate at lag one $\tilde{\alpha}$. A simulation study under AR(2) model is conducted to examine the performance of estimating partial autocorrelation by the proposed method, where AR(2) is defined as $X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + W_t$.

Note that in AR(2), $\rho(1) = \phi_1/(1-\phi_2)$ and $\alpha(2) = \phi_2$. We take ϕ_1 and ϕ_2 with different combinations so that $\rho(1)$ takes value in $\{0.5, 0.7, 0.9\}$ and $\alpha(2)$ takes value in $\{0.6, 0.7, 0.8, 0.9\}$. Simulation results over 1000 independent time series are shown in Table 4.2, where the bias and standard deviation of both autocorrelation estimate at lag one, $\tilde{\rho}$, and partial autocorrelation estimate at lag of two, $\tilde{\alpha}$, are presented compared with the usual sample autocorrelation $\hat{\rho}$ and usual sample partial autocorrelation $\hat{\alpha}$. The RMSE for both autocorrelation and partial autocorrelation

Table 4.1: Simulation results summarizing the estimates of ρ obtained by the sample autocorrelation at lag one $\hat{\rho}$ and the proposed autocorrelation estimate at lag one $\tilde{\rho}$ under AR(1) and ARMA(1,1) model. This summary includes the average estimate minus the true value (bias), standard deviation (sd) and relative mean squared error (RMSE).

			bias		sd		
Model	n	ρ	$\hat{\rho}$	$\tilde{\rho}$	$\hat{\rho}$	$\tilde{\rho}$	RMSE
AR(1)	25	0.1	-0.0570	-0.0495	0.1880	0.2191	0.7650
		0.3	-0.0773	-0.0442	0.1834	0.2061	0.8916
		0.5	-0.1342	-0.0896	0.1902	0.2037	1.0942
		0.7	-0.1623	-0.1140	0.1717	0.1727	1.3042
		0.9	-0.2186	-0.1764	0.1439	0.1356	1.3839
	50	0.1	-0.0293	-0.0206	0.1405	0.1572	0.8193
		0.3	-0.0423	-0.0150	0.1321	0.1420	0.9428
		0.5	-0.0608	-0.0253	0.1222	0.1252	1.1418
		0.7	-0.0855	-0.0520	0.1140	0.1108	1.3560
		0.9	-0.1054	-0.0828	0.0934	0.0864	1.3859
	100	0.1	-0.0143	-0.0066	0.0984	0.1069	0.8626
		0.3	-0.0217	-0.0010	0.0957	0.1002	0.9587
		0.5	-0.0330	-0.0076	0.0885	0.0891	1.1155
		0.7	-0.0414	-0.0187	0.0789	0.0766	1.2769
		0.9	-0.0494	-0.0368	0.0582	0.0544	1.3512
ARMA(1,1)	25	0.1	-0.0573	-0.0501	0.1752	0.2048	0.7645
		0.3	-0.0681	-0.0321	0.1573	0.1759	0.9193
		0.5	-0.0863	-0.0359	0.1409	0.1461	1.2061
		0.7	-0.1090	-0.0598	0.1263	0.1224	1.5003
		0.9	-0.1327	-0.0970	0.0962	0.0863	1.5944
	50	0.1	-0.0305	-0.0219	0.1354	0.1517	0.8198
		0.3	-0.0326	-0.0040	0.1154	0.1237	0.9389
		0.5	-0.0414	-0.0049	0.0993	0.1006	1.1414
		0.7	-0.0465	-0.0138	0.0847	0.0809	1.3844
		0.9	-0.0630	-0.0437	0.0595	0.0538	1.5611
	100	0.1	-0.0126	-0.0047	0.0945	0.1026	0.8613
		0.3	-0.0109	-0.0029	0.0949	0.1030	0.8591
		0.5	-0.0191	0.0066	0.0743	0.0744	1.0535
		0.7	-0.0237	-0.0014	0.0583	0.0561	1.2562
		0.9	-0.0300	-0.0187	0.0391	0.0362	1.4629

are reported for sample size $n = 25, 50, 100$. As can be seen in Table 4.2, the proposed $\tilde{\alpha}$ has smaller MSE compared to $\hat{\alpha}$ for almost every parameter setting and the efficiency improves when n increases.

4.5 Future Work

4.5.1 Tuning parameter selection

As can be seen from the simulation results in both Table 4.1 and Table 4.2, the proposed estimator $\tilde{\rho}$ and $\tilde{\alpha}$ generally alleviate the underestimation and attain smaller MSE compared to sample autocorrelation and sample partial autocorrelation when the correlation is strong, still there are large biases when time series is short and correlation is strong. For example in Table 4.1, we can see under AR(1) model when $n = 50$, $\rho = 0.9$, the bias is around -0.10 for $\hat{\rho}$ and -0.08 for $\tilde{\rho}$. The proposed tuning parameter sequence λ_n does improve the accuracy of the estimation in terms of mse but there definitely exists a better choice of λ_n to improve further. K fold cross validation method is considered to select the best tuning parameter for a zero mean stationary time series Y_t. The cross validation score can be defined as

$$cv_k = \frac{1}{n-h} \sum_{t=1}^{n-h} (y_{t+h} - \tilde{\rho}_\lambda y_t)^2,$$

where $\tilde{\rho}_\lambda$ is the penalized sample autocorrelation estimate (4.5) which are obtained from the training data set that does not contain kth fold data, and $y_1, \cdots, y_n$ is the testing data set in kth fold. Then the cross validation score is the average of K cv's. Monte Carlo simulation results (not shown here) indicate the estimates obtained in this way has smaller bias compared to the proposed penalized $\tilde{\rho}$ but has larger standard deviations. Other methods to select tuning parameter such as generalized cross validation (GCV), the modified version of GCV in Huang et al. (2011), L-curve approach in Hansen (2000) and BIC Zou et al. (2007) all have similar simulation results. None of those methods work well due to the small sample size and high correlation in time series data. Developing novel cross validation schemes to deal with this issue could be a future research topic.

4.5.2 Estimating autocorrelation at all lags simultaneously

Instead of estimating the autocorrelation at lag h for each fixed h, an estimation of $\boldsymbol{\rho} = \{\rho(0), \cdots, \rho(n-1)\}'$ can be formulated as follows. For a time series with zero mean,

Table 4.2: Simulation results summarizing the estimates of ρ obtained by the sample autocorrelation $\hat{\rho}$ and the proposed methods $\tilde{\rho}$, and the estimates of partial autocorrelation at lag two α by the sample partial autocorrelation $\hat{\alpha}$ and the proposed methods $\tilde{\alpha}$ under AR(2) model. This summary includes the average estimate minus the true value (bias), standard deviation (sd) and RMSE of $\hat{\rho}$ with respect to $\tilde{\rho}$ ($R(\rho)$), and $R(\alpha)$ for the corresponding RMSE for partial autocorrelation.

			bias				sd					
n	ρ	α	$\hat{\rho}$	$\tilde{\rho}$	$\hat{\alpha}$	$\tilde{\alpha}$	$\hat{\rho}$	$\tilde{\rho}$	$\hat{\alpha}$	$\tilde{\alpha}$	R(ρ)	R(α)
25	0.5	0.6	-0.333	-0.314	-0.243	-0.188	0.302	0.337	0.173	0.215	0.954	1.093
	0.7	0.6	-0.419	-0.389	-0.262	-0.218	0.293	0.321	0.175	0.216	1.031	1.053
	0.9	0.6	-0.520	-0.484	-0.286	-0.257	0.296	0.317	0.170	0.213	1.068	0.993
	0.5	0.7	-0.421	-0.411	-0.265	-0.192	0.330	0.368	0.169	0.211	0.938	1.211
	0.7	0.7	-0.515	-0.496	-0.288	-0.226	0.335	0.370	0.176	0.220	0.988	1.147
	0.9	0.7	-0.629	-0.602	-0.298	-0.243	0.324	0.355	0.170	0.213	1.027	1.131
	0.5	0.8	-0.510	-0.511	-0.306	-0.225	0.370	0.411	0.166	0.209	0.923	1.288
	0.7	0.8	-0.649	-0.644	-0.309	-0.229	0.374	0.414	0.169	0.213	0.958	1.267
	0.9	0.8	-0.759	-0.746	-0.335	-0.264	0.375	0.414	0.163	0.206	0.986	1.239
	0.5	0.9	-0.726	-0.744	-0.339	-0.253	0.401	0.435	0.169	0.214	0.924	1.304
	0.7	0.9	-0.853	-0.866	-0.348	-0.261	0.411	0.448	0.169	0.214	0.943	1.308
	0.9	0.9	-0.998	-1.006	-0.349	-0.261	0.409	0.448	0.177	0.224	0.960	1.296
50	0.5	0.6	-0.187	-0.161	-0.125	-0.076	0.241	0.256	0.125	0.144	1.014	1.178
	0.7	0.6	-0.237	-0.206	-0.145	-0.110	0.226	0.233	0.125	0.145	1.107	1.114
	0.9	0.6	-0.305	-0.275	-0.177	-0.156	0.204	0.202	0.131	0.153	1.154	1.014
	0.5	0.7	-0.252	-0.232	-0.136	-0.072	0.272	0.292	0.113	0.132	0.989	1.394
	0.7	0.7	-0.324	-0.299	-0.155	-0.102	0.273	0.286	0.117	0.135	1.049	1.317
	0.9	0.7	-0.398	-0.371	-0.176	-0.134	0.261	0.268	0.122	0.143	1.086	1.186
	0.5	0.8	-0.352	-0.341	-0.161	-0.085	0.331	0.355	0.113	0.134	0.963	1.546
	0.7	0.8	-0.433	-0.414	-0.171	-0.101	0.320	0.340	0.112	0.131	1.007	1.517
	0.9	0.8	-0.541	-0.518	-0.185	-0.121	0.310	0.326	0.116	0.137	1.038	1.416
	0.5	0.9	-0.537	-0.540	-0.180	-0.094	0.380	0.407	0.107	0.127	0.948	1.757
	0.7	0.9	-0.683	-0.682	-0.184	-0.098	0.380	0.407	0.107	0.128	0.969	1.735
	0.9	0.9	-0.797	-0.791	-0.184	-0.100	0.379	0.405	0.103	0.124	0.987	1.762
100	0.5	0.6	-0.099	-0.077	-0.063	-0.027	0.176	0.181	0.085	0.093	1.056	1.186
	0.7	0.6	-0.129	-0.106	-0.068	-0.042	0.161	0.160	0.083	0.092	1.154	1.138
	0.9	0.6	-0.165	-0.147	-0.096	-0.082	0.144	0.139	0.092	0.102	1.169	1.033
	0.5	0.7	-0.130	-0.110	-0.069	-0.023	0.214	0.221	0.077	0.085	1.027	1.382
	0.7	0.7	-0.177	-0.155	-0.080	-0.043	0.197	0.198	0.077	0.085	1.108	1.359
	0.9	0.7	-0.223	-0.204	-0.104	-0.078	0.172	0.168	0.082	0.091	1.136	1.221
	0.5	0.8	-0.206	-0.189	-0.074	-0.016	0.254	0.265	0.070	0.078	1.004	1.640
	0.7	0.8	-0.253	-0.233	-0.089	-0.040	0.240	0.245	0.075	0.084	1.062	1.588
	0.9	0.8	-0.325	-0.304	-0.107	-0.066	0.223	0.223	0.076	0.086	1.089	1.478
	0.5	0.9	-0.364	-0.357	-0.088	-0.019	0.324	0.341	0.063	0.072	0.975	2.125
	0.7	0.9	-0.443	-0.430	-0.094	-0.030	0.324	0.338	0.066	0.075	1.005	2.045
	0.9	0.9	-0.538	-0.523	-0.108	-0.048	0.320	0.331	0.067	0.077	1.023	1.947

$$Cov(Y_i, Y_j) = E(Y_i Y_j) = \gamma(|i-j|),$$

for all $1 \leq i, j \leq n$. Then the $\boldsymbol{\rho} = \{\rho(0), \cdots, \rho(n-1)\}$ can be estimated as

$$\sum_{1 \leq i \leq j \leq n} \{y_i y_j - \rho(j-i)\hat{\gamma}(0)\}^2 + n^2 \lambda \sum_{k=1}^{n-1} \{1 - \rho^2(k)\}, \tag{4.9}$$

where $\hat{\gamma}(0) = \sum_{t=1}^{n} y_t^2/n$. Up to a constant the above equation can be rewritten in the following matrix form:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\rho})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\rho}) - n^2 \lambda (\boldsymbol{A}\boldsymbol{\rho})'(\boldsymbol{A}\boldsymbol{\rho}),$$

where, $\boldsymbol{y} = (y_1^2, \cdots, y_n^2, y_1 y_2^*, \cdots, y_n y_{n+1}^*, \cdots, y_1 y_n^*, \cdots, y_n y_{2n-1}^*)'$ is an $n^2 \times 1$ vector, and $\boldsymbol{y}^* = (y_1, \cdots, y_n, 0, \cdots, 0)'$ is an $(2n-1) \times 1$ vector; $\boldsymbol{X} = diag(\mathbf{1}_n, \cdots, \mathbf{1}_n)$ is an $n^2 \times n$ matrix, $\mathbf{1}_n$ is an $n \times 1$ vector with all entries 1; and $\boldsymbol{A} = diag(0, 1, \cdots, 1)$ is an $n \times n$ matrix. The explicit solution can be obtained as

$$\hat{\rho}(h) = \frac{\hat{\gamma}(0) \sum_{t=1}^{n-h} y_{t+h} y_t}{\hat{\gamma}^2(0) n - n^2 \lambda}, \tag{4.10}$$

for $h = 0, \cdots, n-1$. When $\lambda = 0$, $\hat{\rho}(h)$ becomes the sample autocorrelation, the larger λ is the larger the estimate will be dragged to ± 1. The performance of the proposed estimator depends on the selection of tuning parameter λ, and one handy technique to use is generalized cross validation (GCV) that is defined as

$$GCV(\lambda) = \frac{\frac{1}{n}||(\boldsymbol{I} - \boldsymbol{H}_\lambda)\boldsymbol{y}||^2}{|\frac{1}{n} tr(\boldsymbol{I} - \boldsymbol{H}_\lambda)|^2},$$

where $\boldsymbol{H}_\lambda$ is the hat matrix, defined by

$$\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{X}',$$

and I is an $n^2 \times n^2$ diagonal matrix. This method is currently under investigation.

Bibliography

[1] Agrò G. Maximum likelihood and L_p-norm estimators. *Statistica Applicata* 1992; **4**:171-182.

[2] Akaike H. A New Look at the Statistical Model Identification. *IEEE transaction on automatic control* 1974; AC-**19**.

[3] Anderson TW., Walker AM. On the Asymptotic Distribution of the Autocorrelations of a Sample from a Linear Stochastic Process. *The Annals of Mathematical Statistics* 1964; **35**:1296-1303.

[4] Arcones M, Giné E. On the bootstrap of M-estimators and other statistical functionals. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.). Wiley: New York, 1992; 13-47.

[5] Ayebo A, Kozubowski TJ. An asymmetric generalization of Gaussian and Laplace laws. *Journal of Probability and Statistical Science* 2003; **1**:187-210.

[6] Azzalini A. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 1985; **12**:171-178.

[7] Beesley PAA., Griffiths WE. simulation study of the effects of autocorrelation misspecification in a linear statistical model. *Proceedings of the Fifth Biennial Conference of the Simulation Society of Australia*. Armidale, 1982; 120-127.

[8] Bence JR. Analysis of Short Time Series: Correcting for Autocorrelation. *Ecology* 1995; **76**:628-639.

[9] Bera AK, Galvao AF, Montes-Rojas GV, Park SY. Asymmetric Laplace Regression: Maximum Likelihood, Maximum Entropy and Quantile Regression. *Journal of Econometric Methods* 2016; **5**:79-101.

[10] Bickel P., Bühlmann. What is a linear process? *Proc. Natl. Acad. Sci.* 1996; **93**:12128-12131.

[11] Bradic J., Fan J., Wang W. Penalized Composite Quasi-Likelihood for Ultrahigh-Dimensional Variable Selection. *Journal of Royal Statistical Society: Series B* 2011; **73**:325-349.

[12] Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo L Jr, Jones DW, Materson BJ, Oparil S, Wright JT Jr, Roccella EJ, National Heart, Lung, Blood Institute, National High Blood Pressure Education Program Coordinating Committee. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. *Hypertension* 2003; **42**:1206-1252.

[13] Cochrane D., Orcutt GH. Application of Least Squares Regression to Relationships Containing Auto-Correlated Error Terms. *Journal of the American Statistical Association* 1949; **44**:32-61.

[14] Durbin J., Watson GS. Testing for Serial Correlation in Least Squares Regression: I. *Biometrika* 1950; **37**:409-428.

[15] Efron B. *The Jackknife, the Bootstrap and Other Resampling Plans*. Capital City Press: Montpelier, 1982.

[16] Efron B., Hastie T., Johnstone I., Tibshirani R. Least Angle Regression. *The Annals of Statistics* 2004; **32**:407-499.

[17] Fan J., Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* 2001; **96**.

[18] an J., Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space. *Stat Sin.* 2010; **20**:101-148.

[19] Fernandez C, Steel MFJ. On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association* 1998; **93**:359-371.

[20] Golub G., Heath M., Wahba G. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* 1979; **21**.

[21] Griffiths WE., Beesley PAA. he small sample properties of some preliminary-test estimators in a linear model with autocorrelated errors. *Journal of Econometrics* 1984; **25**:49-62.

[22] Hansen PC. The L-curve and its use in the numerical treatment of inverse problems. In *Computational Inverse Problems in Electrocardiology, Advances in Computational Bioengineering* (P. Johnston, eds.); 119-142.

[23] Hardy JB. The collaborative perinatal project: Lessons and legacy. *Annals of Epidemiology* 2003; **13**:303-311.

[24] Huang C., Hsing T., Cressie N. Spectral Density Estimation Through A Regularized Inverse Problem. *Statistica Sinica* 2011; **21**:1115-1144.

[25] Huber PJ. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 1964; **35**:73-101.

[26] Huber PJ. The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions. In *Fifth Symposium on Mathematical Statistics and Probability*. University of California, Berkeley, California, 1967; 179-195.

[27] Huber PJ. Robust Regression Asymptotics Conjectures and Monte Carlo. *The Annals of Statistics* 1973; **1**:799-821.

[28] Huber PJ, Ronchetti EM. *Robust Statistics*, 2nd ed. Wiley: Hoboken, 2009.

[29] Hurlbert SH. Pseudoreplication and the Design of Ecological Field Experiments. *Ecological Monographs* 1984; **54**:187-211.

[30] Hurvich CM., Tsai CL. Regression and Time Series Model Selection in Small Samples. *Biometrika* 1989; **76**:297-307.

[31] Kai B, Li R, Zou H. Local Composite Quantile Regression Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression. *Journal of the Royal Statistical Society, Series B* 2010; **71**:49-69.

[32] King ML., Giles, DEA. Autocorrelation pre-testing in the linear model: Estimation, testing and prediction. *Journal of Econometrics* 1984; **25**:35-48.

[33] Koenker R, Bassett G. Regression Quantiles. *Econometrica* 1978; **46**:33-50.

[34] Bassett G., Koenker R. Asymptotic Theory of Least Absolute Error Regression. *Journal of the American Statistical Association* 1978; **73**:618-622.

[35] Koenker R. *Quantile Regression*. Cambridge University Press: Cambridge, 2005.

[36] Lee ER., Noh H., Park BU. Model Selection via Bayesian Information Criterion for Quantile Regression Models. *Journal of the American Statistical Association* 2014; **109**.

[37] Li Y., Zhu J. l_1-norm quantile regressions. *J. Comput. Graph. Statist* 2008; **17**:1-23.

[38] Lomnicki ZA., Zaremba SK. On the Estimation of Autocorrelation in time Series. *The Annals of Mathematical Statistics* 1957; **28**:140-158.

[39] Mallow CL. Some Comments on C_p. *Technometric* 1973; **15**:661-675.

[40] Marriott FHC., Pope JA. Bias in the Estimation of Autocorrelations. *Biometrika* 1954; **41**:390-402.

[41] Mineo A. The norm-p estimation of location, scale and simple linear regression parameters. Lecture notes in statistics. *Statistical Modelling Proceedings*. Trento, 1989; 222-233.

[42] Mudholkar GS., Hutson AD. The epsilon-skew-normal distribution for analyzing near-normal data. *Journal of Statistical Planning and Inference* 2000; **83**:291-309.

[43] Nakamura A., Nakamura M. On the Impact of the Tests for Serial Correlation Upon the Test of Significance for the Regression Coefficient. *Journal of Econometrics* 1978; **7**.

[44] National Institute of Diabetes and Digestive and Kidney Diseases. National Diabetes Statistics Report, 2014. Bethesda, MD: Department of Health and Human Services; 2017.

[45] Osborne M., Presnell B., Turlach B. On the LASSO and Its Dual. *Journal of Computational and Graphical Statistics* 2000; **9**:319-337.

[46] Park, RE., Michell, BM. Estimating the Autocorrelated Error Model with Trended Data. *Journal of Econometric* 1980; **13**:185-201.

[47] Quenouille MH. Approximate Tests of Correlation in Time Series. *Journal of the Royal Statistical Society: Series B* 1949; **11**:68-84.

[48] Rao CR. Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* 1945; **37**:81-89.

[49] Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**:461-464.

[50] Shao J. An asymptotic theory for linear model selection (with discussion). *Statist. Sin.* 1997; **7**:221-264.

[51] Subbotin MT. On the law of frequency of error. *Matematicheskii Sbornik* 1923; **31**:296-301.

[52] Sun J., Gai Y., Lin L. Weighted local linear composite quantile estimation for the case of general error distributions. *Journal of Statistical Planning and Inference* 2013; **143**:1049-1063.

[53] The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2003; **26**(Suppl 1):S5–S20.

[54] Tibshirani R. Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996; **58**:267-288.

[55] Wang H., Li G., Tsai CL. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 2007; **69**:63-78.

[56] Whitcomb B, Schisterman E, Klebanoff M, Baumgarten M, Rhoton-Vlasak A, Luo X, Chegini N. Circulating chemokine levels and miscarriage. *American Journal of Epidemiology* 2007; **331**:166-323.

[57] Wood GW, Hausmann E, Choudhuri R. Relative role of CSF-1, MCP-1/JE, and RANTES in macrophage recruitment during successful pregnancy. *Molecular Reproduction and Development* 1997; **46**:62-70.

[58] World Health Organization; 2016. Available from: http://www.who.int/en/.

[59] Yi C., Huang J. Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression. *Journal of Computational and Graphical Statistics* 2016.

[60] Zeckhauser R, Thompson M. Linear regression with non-normal error terms. *The Review of Economics and Statistics* 1970; **52**:280-286.

[61] Zheng Q., Gallagher C., Kulasekera KB. Adaptively weighted kernel regression. *Journal of Nonparametric Statistics* 2013; **25**:855-872.

[62] Zheng Q., Gallagher C., Kulasekera KB. Adaptive penalized quantile regression for high dimensional data. *Journal of Statistical Planning and Inference* 2013; **143**:1029-1038.

[63] Zhu D., Zinde-Walsh V. Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics* 2009; **148**:86-99.

[64] Zou H. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 2006; **101**.

[65] Zou H., Hastie T., Tibshirani R. ON THE "DEGREES OF FREEDOM" OF THE LASSO. *The Annals of Statistics* 2007; **35**:2173-2192.

[66] Zou H, Yuan M. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics* 2008; **36**:1108-1126.