Weighted quantile regression and oracle model selection

Jiang, Xuejun

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Statistics

July, 2009

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree of qualification of this or any other university or other institution of learning.

Abstract of thesis entitled:
Weighted quantile regression and oracle model selection
Submitted by Jiang, Xuejun
for the degree of Doctor of Philosophy
at The Chinese University of Hong Kong in July, 2009

In this dissertation I suggest a new (regularized) weighted quantile regression estimation approach for nonlinear regression models and double threshold ARCH (DTARCH) models. I allow the number of parameters in the nonlinear regression models to be fixed or diverge. The proposed estimation method is robust and efficient and is applicable to other models. I use the adaptive-LASSO and SCAD regularization to select parameters in the nonlinear regression models. I simultaneously estimate the AR and ARCH parameters in the DTARCH model using the proposed weighted quantile regression. The values of the proposed methodology are revealed.

Under regularity conditions, I establish asymptotic distributions of the proposed estimators, which show that the model selection methods perform as well as if the correct submodels are known in advance. I also suggest an algorithm for fast implementation of the proposed methodology. Simulations are conducted to compare different estimators, and a real example is used to illustrate their performance.

Keywords: Weighted quantile regression, Adaptive-LASSO, High dimensionality, Model selection, Oracle property, SCAD, DTARCH models.

Acknowledgement

I would like to thank my supervisors Prof. Sik-Yum Lee and Prof. Xinyuan Song for their generosity of supervision and encouragement during the course of my research program. It is also my pleasure to thank all staff of the Department of Statistics, The Chinese University of Hong Kong, for their kind assistance. Last, but not the least, I would also like to thank my parents, sisters and my elder brother for their everlasting support and encouragement during my PhD research program and my whole life of course.

Contents

List of Tables

Chapter 1

Introduction

1.1 Overview

In this dissertation I focus on two important models, the nonlinear regression model with a fixed or diverging number of parameters and the DTARCH models.

The real world is nonlinear. From the economic society to human intelligence, nonlinearity widely exists. Our economy could not expand in a linear way, and artificial neural networks are complicate nonlinear models. Existing statistical theory is beautiful for linear models, but there is too much work to be done about nonlinear models. There is a genuine demand for us to advance statistical techniques for various nonlinear models. I work on the above two models to introduce a new modeling methodology, the weighted quantile regression (WQR) with associated model selection strategies, which can be extended to other models, such as transformation models and semiparametric/nonparametric models.

Quantiles regression (QR) is a statistical technique designed to estimate and conduct inference about conditional quantile functions. Advantages of QR over mean regression are advocated by Koenker and Bassett (1978) and Chaudhuri, Doksum and Samarov (1997). In addition to more accurate portrayal of the stochastic relationship between random variables, QR provides more robust and

consequently more efficient estimates than the mean regression when the error is non-normal (Koenker and Bassett, 1978; Koenker and Zhao, 1996).

The proposed WQR includes the QR as a special case and is considerably more efficient than QR (see Chapters 2-4). Since the WQR involves a vector of weights which may take negative values, I develop a data-driven strategy for deciding the weights. The WQR with data-driven weights is adaptive, in the sense that it performs as well as if the optimal weights were known, and hence it achieves maximum asymptotic efficiency among all WQR.

The nonlinear regression model is very useful in statistics, econometrics and human intelligence. It includes linear models and generalized linear models with continuous responses as specific examples. It can also be used when the effects of some covariates are linear and the remaining are nonlinear. A number of artificial neural networks are special cases of the nonlinear regression model with a lot of parameters. For these nonlinear models, it is fundamental to efficiently and robustly estimate the parameters and to select the best model from them. This motivates me to propose the WQR estimation. Common stepwise deletion and subset selection procedures have difficulty in implementation and derivation of sampling properties. To this end, I study regularized WQR which simultaneously estimates the parameters and selects the models.

The DTARCH model is also nonlinear. It is useful for capturing changing volatility. Modeling volatility lies at the core of activity in financial markets. Since volatility is fundamental for asset pricing, monetary policymaking, proprietary trading, portfolio management and risk analysis, it is especially important to accurately forecast volatility. The celebrated ARCH model (Engle, 1982) is an important tool in modeling the changing volatility. It has received tremendous

attention. A number of variants of the ARCH model have been proposed as important tools in modeling the changing volatility. See for example, Bollerslev et al. (1992), Bera and Higgins (1993), Fan and Yao (2003), and Peng and Yao (2003). One example is the DTARCH model in Li and Li (1996). The DTARCH model is very useful in detecting nonlinear structures in the mean and volatility of an asset return, and heteroscedasticity with clustering in the volatility. As financial returns can be very heavy-tailed, the maximum likelihood (ML) method may create serious problems in parameter estimation and conditional prediction intervals. Given that, the efficiency of the ML estimators may be very low under this case. Therefore, more efficient estimators than the ML estimators are required, and an attractive alternative is to use quantile regression (QR) estimation. This again motivates me to consider the WQR estimation approach for estimating the model parameters, especially for simultaneously estimating the parameters in the AR and ARCH parts.

The (regularized) WQR estimators admit no close form and involve minimizing complicate nonlinear functions, so it is challenging to derive asymptotic properties and to implement the methodology. Theoretically, I establish asymptotic normality of the resulting estimators and show their optimality, no matter whether the error variance is finite or not. Practically, I develop an algorithm for fast implementation of the proposed methodology, based on the "interior point algorithm" [Vanderbei, Meketon and Freedman (1986) and Koenker and Park (1996)]. The resulting algorithm is easy to implement. The advantages of the proposed methodology are illustrated by simulations and real data examples.

The findings of this dissertation will contribute to the theory and practice of statistics and econometrics in many important aspects. Our methodology will

be particularly attractive in QR settings and should greatly expand the scope of inference currently described in the statistics and econometrics literature.

1.2 Outline of the Thesis

The thesis is organized as follows. In Chapter 2, I proposed a weighted QR method for nonlinear models with finite parameters and use adaptive-Lasso and SCAD penalties to select parameters in the models. In Chapter 3, I study the same problems as in Chapter 2 but allow for a diverging number of parameters. Implementation of the proposed methodology, simulations and real data analysis are presented in this chapter. In Chapter 4, I study the weighted QR method for DTARCH models and propose a simultaneous estimation method for the parameters in the AR and ARCH parts. Conclusions and further developments are presented in Chapter 5. Finally, proofs of theorems are given in the Appendices.

□ **End of chapter.**

Chapter 2

Weighted QR with a fixed number of parameters and oracle model selection

2.1 Introduction

Various techniques have been developed for simultaneous variable selection and coefficient estimation, based on the penalized likelihood or least squares principles. Examples include the nonnegative garrote [Breiman (1995) and Yuan and Lin (2007)], the LASSO [Tibshirani (1996, 1997)], the bridge regression [Fu (1998) and Knight and Fu (2000)], the SCAD [Fan and Li (2001)], etc. These methods have advantages over traditional stepwise deletion and subset selection procedures in implementation and derivation of sampling properties, and have been extended by several authors to achieve robustness. For instance, for linear models, Wang, Li and Jiang (2007) considered the LASSO for least absolute regression (LAD-LASSO), and Zou and Yuan (2008a) studied the LASSO for composite quantile regression (CQR-LASSO), among others. These endeavors have enriched the variable selection theory for different models by using different regularized estimation methods, with aim at oracle model selection procedures [see Fan and Li (2006) for a comprehensive overview] and robustness and efficiency of the estimation [Zou

and Yuan (2008a)].

The CQR-LASSO in Zou and Yuan (2008a) is robust and efficient and performs nearly like an oracle model selector. The CQR they used is a sum of different quantile regression (QR) [Koenker and Bassett (1978)] at predetermined quantiles, which can be regarded as a weighted quantile regression (WQR) with equal weights (see also Section 2 for details). Intuitively, equal weights are not optimal in general, and hence a more efficient WQR should exist. Therefore, in this article we suggest a WQR estimation method and let the data decide the weights to improve efficiency while keeping robustness from the QR. The WQR method is applicable to various models, but in this chapter we focus only on the nonlinear model

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{2.1.1}$$

where ε_i's are independent random errors with an unknown distribution function $G(\cdot)$, and the function $f(\cdot, \boldsymbol{\beta})$ is known up to a p-dimensional vector of parameters $\boldsymbol{\beta}$. Model (2.1.1) contains many submodels. Linear models and generalized linear models with continuous responses are specific examples. The nonlinear model can also be used when the effects of some covariates are linear and the remaining are nonlinear.

QR estimation of model (2.1.1) with a fixed p has received much attention. See Oberhofer (1982), Dupačová and Wets (1988), Powell (1986, 1991), Jurečková and Procházka (1994), and Wang (1995), among others, but for model (2.1.1) with a diverging number of parameters, to the best of our knowledge, there is no formal work using QR in the literature. Hence, the WQR for model (2.1.1) is new and fundamental. We will address the issue of variable/parameter selection using the penalized WQR with the adaptive LASSO and SCAD penalties. The

resulting method is robust against outliers and heavy-tailed error distributions, like the Cauchy distribution, and efficient as nearly as the MLE when the error is normal. In addition, the weights in the WQR are allowed to be negative, so the proposed WQR is essentially different from the common QR and the CQR (see also Section 2). When the weights are all equal and the model is linear with a fixed number of parameters, our method reduces to that of Zou and Yuan (2008a) if the LASSO penalty is employed. Since the proposed WQR involves a vector of weights, we develop a data-driven weighting strategy which maximizes the efficiency of the WQR estimators. The resulting estimation is adaptive in the sense that it performs asymptotically the same as if the theoretically optimal weights were used.

The penalized WQR estimators admit no close form and involve minimizing complicate nonlinear functions, so it is challenging to derive asymptotic properties and to implement the methodology. Theoretically, we will establish asymptotic normality of the resulting estimators and show their optimality, no matter whether the error variance is finite or not. Practically, we will develop an algorithm for fast implementation of the proposed methodology. This algorithm solves a succession of (penalized) linearized WQR problems, each of whose dual problems is derived. We will extend the "interior point algorithm" [Vanderbei, Meketon and Freedman (1986) and Koenker and Park (1996)] to solve these dual problems. The resulting algorithm is easy to implement. Simulations endorse our discovery.

This chapter is organized as follows. In Section 2.2 we introduce the WQR for model (2.1.1) with a fixed number of parameters and use a data-driven weighting scheme to maximize the efficiency for the resulting WQR estimators. In section 2.3, we study the variable/parameter selection problem using the adaptive LASSO and SCAD penalties. In section 2.4, we consider the computation method for the

proposed methodology. Conclusion is given in section 2.5. Finally, we give proofs of theorems in the appendix A.

2.2 Weighted Quantile Regression

Our idea can be well motivated from the linear model,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{2.2.2}$$

where $\{\varepsilon_i\}$ are i.i.d. noise with unknown distribution $G(\cdot)$ and density $g(\cdot)$.

By Koenker and Basset (1978), the τ-th QR estimate of $\boldsymbol{\beta}$ can be obtained via minimizing

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

over $\boldsymbol{\beta}$ and b_τ, where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the check function with derivative $\psi_\tau(u) = \tau - I(u < 0)$ for $u \neq 0$. Noticing that the regression coefficients are the same across different QR models, Zou and Yuan (2008a) proposed to estimate $\boldsymbol{\beta}$ by minimizing

$$\sum_{k=1}^{K}\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k}), \tag{2.2.3}$$

over $\boldsymbol{\beta}$ and b_{τ_k} and used the adaptive LASSO penalty [Zou (2006)] for (2.2.3) to select variables, where $\{\tau_k\}_{k=1}^K$ are predetermined over $(0, 1)$. This is the aforementioned CQR-LASSO.

Note that the CQR method uses the same weight for different QR models. Intuitively, it is more effective if different weights are used, which leads to minimizing

$$\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k}),$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)'$ is a vector of weights such that $\|\boldsymbol{\omega}\| = 1$ with $\|\cdot\|$ denoting the Euclidean norm. The weight ω_k controls the amount of contribution of the τ_k-th QR. The components in the weight vector $\boldsymbol{\omega}$ are allowed to be negative, since

$\{\sum_{i=1}^{n} \rho_{\tau_k}(y_i - \mathbf{x}_i'\boldsymbol{\beta} - b_{\tau_k})\}_{k=1}^{K}$ may not be positively correlated. Thus, the WQR is essentially different from the CQR. Applying the weighting scheme to model (2.1.1), one can estimate $\boldsymbol{\beta}$ by minimizing

$$L_n(\boldsymbol{\beta}, \mathbf{b}) \equiv \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - f(\mathbf{x}_i, \boldsymbol{\beta}) - b_{\tau_k}), \tag{2.2.4}$$

over $\boldsymbol{\beta}$ and $\mathbf{b} = (b_{\tau_1}, \ldots, b_{\tau_K})'$. For convenience, we denote by $\tilde{\boldsymbol{\beta}}_1$ the minimizer of $\boldsymbol{\beta}$ for (2.2.4) and refer to it as "the WQR estimator". The CQR method can be regarded as an example of the WQR estimation with $\omega_i = 1/\sqrt{K}$. In general, given K, one can use the equally spaced quantiles at $\tau_k = k/(K+1)$ for $k = 1, 2, \ldots, K$. Typically, one can use at least the three quantiles at $\tau_k = 0.25, 0.5$, and 0.75.

In order to derive the asymptotic property of the proposed estimator, in the following we introduce some notations and conditions. Let $\boldsymbol{\beta}^*$ be the true value of $\boldsymbol{\beta}$, $b^*_{\tau_k}$ be the τ_k-th quantile of ε, and $\mathbf{b}^* = (b^*_{\tau_1}, \ldots, b^*_{\tau_K})'$. Denote by $f_i^* = f(\mathbf{x}_i, \boldsymbol{\beta}^*)$, $\boldsymbol{\nabla} f_i^* = [\partial f(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}] \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$, and $\boldsymbol{\nabla}^2 f_i^* = [\partial^2 f(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'] |_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$. Assume that

(a) $\mathbf{G} = \mathrm{var}(\boldsymbol{\nabla} f_1^*) > 0$.

(b) The error ε_i has the distribution function $G(\cdot)$ and density function $g(\cdot)$. The density function g is positive and continuous at the τ_k-th quantiles $b^*_{\tau_k}$.

(c) There is a large enough open subset $\Omega \in \mathbf{R}^p$ which contains the true parameter point $\boldsymbol{\beta}^*$, such that for all $\mathbf{x}_i$ the second derivative matrix $\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta})$ of $f(\mathbf{x}_i, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ satisfies that

$$\|\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_1) - \nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_2)\| \leq M(\mathbf{x}_i)\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|$$

$$|\partial^2 f(\mathbf{x}_i, \boldsymbol{\beta})/(\partial \beta_j \partial \beta_k)| \leq N_{jk}(\mathbf{x}_i)$$

for all $\boldsymbol{\beta}_i \in \Omega$, where $E[M^2(\mathbf{x}_i)] < \infty$ and $E[N_{jk}^2(\mathbf{x}_i)] < C_1 < \infty$ for all j, k.

Under these mild conditions, we have the following asymptotic normality result.

Theorem 2.2.1. *Under the conditions (a)-(c),*

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(\boldsymbol{0}, \sigma^2(\boldsymbol{\omega})\mathbf{G}^{-1}\Big),$$

where

$$\sigma^2(\boldsymbol{\omega}) = \{\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})\}^{-2} \sum_{k,k'=1}^{K} \omega_k \omega_{k'} \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'})).$$

For linear models (2.2.2), $\mathbf{G} = \mathrm{var}(\mathbf{x}_1)$. If all ω_k are equal, then Theorem 2.2.1 reduces to the asymptotic normality of the CQR estimators in Zou and Yuan (2008a). When $K = 1$ and $\tau_1 = \tau$, it follows from the above theorem that the τ-th QR estimate of $\boldsymbol{\beta}$ is $\sqrt{n}$-consistent and asymptotically normal with mean zero and variance $g^{-2}(b^*_\tau)\tau(1-\tau)\mathbf{G}^{-1}$. Note that the asymptotic variance of the least squares estimator is $\sigma^2\mathbf{G}^{-1}$ [see for example Jennrich (1969) and Wu (1981)], where σ^2 is the variance of the error. It follows that the asymptotic relative efficiency (ARE) of the WQR estimation with respect to the least squares estimation is

$$\mathrm{ARE}(\boldsymbol{\omega}, g) = \sigma^2\sigma^{-2}(\boldsymbol{\omega}).$$

Since $\mathbf{G}$ does not involve $\boldsymbol{\omega}$, the weights should be selected to minimize $\sigma^2(\boldsymbol{\omega})$. Let $\mathbf{g} = (g(b^*_{\tau_1}), \ldots, g(b^*_{\tau_K}))'$, and let $\boldsymbol{\Omega}$ be a $K \times K$ matrix with the (k, k') element being $\Omega_{kk'} = \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'}))$. Then the optimal weight $\boldsymbol{\omega}_{opt}$, which minimizes $\sigma^2(\boldsymbol{\omega})$, can be shown as

$$\boldsymbol{\omega}_{opt} = (\mathbf{g}'\boldsymbol{\Omega}^{-2}\mathbf{g})^{-1/2}\boldsymbol{\Omega}^{-1}\mathbf{g}.$$

The optimal weight components can be very different, and some of them may even be negative. In fact, in our simulations we also experience such a scenario.

This reflects the necessity to use a data-driven weighting scheme. The density function $g(\cdot)$ of ε_i can be estimated by running the kernel smoother over residuals. Let the resulting estimate of $\mathbf{g}$ be $\hat{\mathbf{g}}$. Then $\hat{\boldsymbol{\omega}} = (\hat{\mathbf{g}}'\boldsymbol{\Omega}^{-2}\hat{\mathbf{g}})^{-1/2}\boldsymbol{\Omega}^{-1}\hat{\mathbf{g}}$ provides a nonparametric estimator of $\boldsymbol{\omega}$. This leads to an adaptive estimator of $\boldsymbol{\beta}^*$ by minimizing

$$\sum_{k=1}^{K}\hat{\omega}_k\sum_{i=1}^{n}\rho_{\tau_k}(y_i - f(\mathbf{x}_i;\boldsymbol{\beta}) - b_{\tau_k}) \tag{2.2.5}$$

over b_{τ_k} and $\boldsymbol{\beta}$, where $\hat{\omega}_k$ is the k-th component of $\hat{\boldsymbol{\omega}}$. Let the resulting estimator of $\boldsymbol{\beta}$ be $\tilde{\boldsymbol{\beta}}_2$. Then $\tilde{\boldsymbol{\beta}}_2$ is asymptotically normal from the following theorem.

Theorem 2.2.2. *Under the same conditions as in Theorem 2.2.1,*

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(\mathbf{0}, (\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g})^{-1}\mathbf{G}^{-1}\Big).$$

Since $\sigma^2(\boldsymbol{\omega}_{opt}) = (\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g})^{-1}$, $\tilde{\boldsymbol{\beta}}_2$ has the same asymptotic variance matrix as $\tilde{\boldsymbol{\beta}}_1$, if $\boldsymbol{\omega}_{opt}$ were known. That is, the estimator $\tilde{\boldsymbol{\beta}}_2$ is adaptive. By Theorem 2.2.2, the asymptotic relative efficiency (ARE) of the optimal WQR estimation with respect to the least squares estimation is $\sigma^2\mathbf{g}'\boldsymbol{\Omega}^{-1}\mathbf{g}$.

2.3 Penalized WQR and oracle model selection

In the section, we consider the model selection method using the penalized WQR and establish the oracle properties of the resulting estimates.

2.3.1 Parameter selection with SCAD penalty

Partition the parameter vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1 \in \mathbf{R}^s$ and $\boldsymbol{\beta}_2 \in \mathbf{R}^{p-s}$. Assume that the true regression coefficients are $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*T}, \boldsymbol{\beta}_2^{*T})'$ with each component in $\boldsymbol{\beta}_1^*$ being nonzero and in $\boldsymbol{\beta}_2^*$ being zero. Denote $f_{i1}^* = f(\mathbf{x}_i, \boldsymbol{\beta}_1^*)$, $\boldsymbol{\nabla} f_{i1}^* = [\partial f(\mathbf{x}_i, \boldsymbol{\beta})/\partial\boldsymbol{\beta}]\,|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1^*}$, and $\mathcal{A} = \{j : j = 1, 2, \ldots, s\}$. Following Fan and Li (2001), for

an estimation procedure ξ, the resulting estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}(\xi) = (\hat{\beta}_1(\xi), \ldots, \hat{\beta}_p(\xi))'$, is called a WQR-oracle estimator if it satisfies

(i) consistency in selection: $P(\{j : \hat{\beta}_j(\xi) \neq 0\} = \mathcal{A}) \to 1$;

(ii) efficient estimation: $\sqrt{n}(\hat{\boldsymbol{\beta}}_1(\xi) - \boldsymbol{\beta}_1^*) \to \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{G}_{11}^{-1})$, where $\mathbf{G}_{11}$ is the sub-matrix of $\mathbf{G}$ with both row and column indices in $\mathcal{A}$ and $\sigma^2(\boldsymbol{\omega})$ is defined in Theorem 2.2.1.

Zou and Yuan (2008a) obtained a CQR-oracle estimator of $\boldsymbol{\beta}$ for model (2.2.2) using the adaptive LASSO penalty. This stimulates us to study the penalized WQR estimation for the nonlinear model (2.1.1) using the adaptive LASSO and SCAD penalties. The SCAD penalty (Fan and Li, 2001) is mathematically defined in terms of its first order derivative and is symmetric about the origin. For $\theta > 0$, its first order derivative is given by

$$p'_\lambda(\theta) = \lambda\Big\{I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\Big\},$$

where $a > 2$ and $\lambda > 0$ are tuning parameters. We define the SCAD penalized WQR by solving

$$(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mathbf{b},\boldsymbol{\beta}} Q^{SC}(\boldsymbol{\beta}, \mathbf{b}), \tag{2.3.6}$$

where $Q^{SC}(\boldsymbol{\beta}, \mathbf{b}) = L_n(\boldsymbol{\beta}, \mathbf{b}) + n\sum_{j=1}^p p_{\lambda_n}(|\beta_j|)$. For convenience, the estimation is coined as WQR-SCAD method.

Theorem 2.3.1. (**Consistency**) *Assume that Conditions (a)-(c) hold. If* $\lambda_n \to 0$ *as* $n \to \infty$, *there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-\frac{1}{2}})$.

The following condition from Fan and Li (2001) is needed for deriving asymptotic normality of the WQR-SCAD estimator:

(d) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$.

Theorem 2.3.2. (**Oracle**) *Assume that Conditions (a)-(d) hold. If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then with probability tending to one the* $\sqrt{n}$*-consistent estimator* $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2'\right)'$ *in Theorem 2.3.1 must satisfy that*

(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_2 = 0$; *and*

(ii) *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{G}_{11}^{-1}\right)$.

Theorem 2.3.2 demonstrates that $\hat{\boldsymbol{\beta}}$ is a WQR-oracle estimator. Note that $\sigma^2(\boldsymbol{\omega})$ involves the density but not the variance of the error term. The theorem still holds when the error term has an infinite variance, like the Cauchy distribution. This property is possessed by the CQR-oracle estimator but not shared by the least-squares oracle estimators, which demonstrates that the WQR-SCAD is quite robust.

2.3.2 Parameter selection with adaptive-LASSO

As a variable selection method, LASSO was proposed by Tibshirani (1996) using the L_1 penalty. Zou (2006) introduced the adaptive LASSO by penalizing different parameters with adaptive weights, which makes the LASSO be an oracle method. In what follows we develop the adaptive LASSO theory for the WQR estimation of model (2.1.1). By Theorem 2.1, $\tilde{\boldsymbol{\beta}}_1$ is root-n consistent. Using $\tilde{\boldsymbol{\beta}}_1$, we construct the adaptive LASSO penalized estimator:

$$(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}^{AL}) = \arg\min_{\mathbf{b},\boldsymbol{\beta}} Q^{AL}(\boldsymbol{\beta}, \mathbf{b}), \tag{2.3.7}$$

where $Q^{AL}(\boldsymbol{\beta}, \mathbf{b}) = L_n(\boldsymbol{\beta}, \mathbf{b}) + nh_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|$, and the weights are set to be $\tilde{w}_j = |\tilde{\beta}_{1j}|^{-\gamma}$ for some $\gamma > 0$. The estimation approach is referred to as the adaptive

WQR-LASSO for convenience. When f is linear about $\boldsymbol{\beta}$ and the weights are equal, it reduces to the adaptive CQR-LASSO in Zou and Yuan (2008a), which has been shown to enjoy the CQR-oracle property. The following theorem extends the result to the WQR-LASSO for nonlinear model (2.1.1).

Theorem 2.3.3. *Assume that Conditions (a)-(c) hold. If* $n^{1/2}h_n \to 0$ *and* $h_n n^{(\gamma+1)/2} \to \infty$, *then*

(i) *Sparsity:* $\hat{\boldsymbol{\beta}}_2^{AL} = 0$.

(ii) *Asymptotic normality:* $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{AL} - \boldsymbol{\beta}_1^*) \to \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{G}_{11}^{-1})$.

Theorems 2.3.2 and 2.3.3 indicate that the adaptive WQR-LASSO is asymptotically equivalent to the WQR-SCAD.

2.4 Numerical Implementation

Since the target functions are high-dimensional with singularities, and minimization in (2.2.4), (2.3.6) and (2.3.7) involves complicate nonlinear optimization problems, it is challenging to implement the proposed methodology. In the following we introduce a fast algorithm for computation. This algorithm solves a succession of penalized linearized WQR problems, each of which is solved by extending the interior point algorithm [see Osborne and Watson (1971) and Koenker and Park (1996)]. Matlab codes are available upon request for the proposed methods.

Minimization problem (2.2.4) is a specific example of (2.3.7) with $h_n = 0$, and (2.3.6) can be solved using a similar method to (2.3.7). First, we consider the minimization of (2.3.7). This problem is equivalent to

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - l_{ik}(\boldsymbol{\theta})) + nh_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|, \tag{2.4.8}$$

where $l_{ik}(\boldsymbol{\theta}) = f(\mathbf{x}_i, \boldsymbol{\beta}) + b_{\tau_k}$ and $\boldsymbol{\theta} = (\mathbf{b}', \boldsymbol{\beta}')'$. Following Osborne and Watson (1971), we solve (2.4.8) using the following algorithm:

(1) Given the current value, $\boldsymbol{\theta}^{(r)}$, of $\boldsymbol{\theta}$, calculate $\mathbf{t}$ to minimize

$$\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - l_{ik}(\boldsymbol{\theta}^{(r)}) - \boldsymbol{\nabla} l_{ik}(\boldsymbol{\theta}^{(r)})\mathbf{t}\} + nh_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|, \tag{2.4.9}$$

where $\boldsymbol{\nabla} l_{ik}(\boldsymbol{\theta}^{(r)}) = \frac{\partial l_{ik}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(r)}}$ and β_j is the $(K+j)$th component of $\boldsymbol{\theta}^{(r)}+\mathbf{t}$. Let the minimizer be $\mathbf{t} = \mathbf{t}^{(r)} = (t_1^{(r)}, \ldots, t_{K+p}^{(r)})'$.

(2) Calculate λ to minimize

$$\begin{aligned} &\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - l_{ik}(\boldsymbol{\theta}^{(r)} + \lambda \mathbf{t}^{(r)})\} \\ &+nh_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j^{(r)} + \lambda t_{K+j}^{(r)}|. \end{aligned} \tag{2.4.10}$$

Let the minimizer be $\lambda = \lambda^{(r)}$.

(3) Set $\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} + \lambda^{(r)}\mathbf{t}^{(r)}$. Update the current value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}^{(r+1)}$, and repeat the above procedure until the new iteration fails to improve the objective function by a specified tolerance.

For the above method, the problem (2.4.10) can easily be solved by line search in the resulting direction $\mathbf{t} = \mathbf{t}^{(r)}$, but one has to solve a succession of penalized linearized WQR problems in (2.4.9). Let $y_{ik}^* = y_i - l_{ik}(\boldsymbol{\theta}^{(r)})$ and $\mathbf{a}_{ik}' = \nabla l_{ik}(\boldsymbol{\theta}^{(r)})$. Then the problem (2.4.9) becomes

$$\min_{\mathbf{t}}\{\sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}) + nh_n \sum_{j=1}^{p} \tilde{w}_j |\beta_j|\}. \tag{2.4.11}$$

For $j = 1, \ldots, p$ and $k = 1, \ldots, K$, let $y_{(n+j)k}^* = 0$ and $\mathbf{a}_{(n+j)k} = nh_n \tilde{w}_j \mathbf{e}_{K+j}$, where $\mathbf{e}_{K+j}$ is a $(K+p) \times 1$ vector with the $(K+j)$th entry being one and others being

zeros. Then (2.4.11) becomes the linear programming problem:

$$\min_{\mathbf{t}} \sum_{k=1}^{K} \omega_k \Big\{ \sum_{i=1}^{n} \rho_{\tau_k}(y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}) + \sum_{i=n+1}^{n+p} \omega_k |y_{ik}^* - \mathbf{a}_{ik}'\mathbf{t}| \Big\}. \tag{2.4.12}$$

For $k = 1, \ldots, K$, let $\mathbf{y}_k^* = (y_{1k}^*, \ldots, y_{n+p,k}^*)'$, $\mathbf{u}_k = \text{vec}(\mathbf{y}_k^*, \mathbf{0}_{p\times 1})$, $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_k')'$, and $\mathbf{A}_k = (\mathbf{a}_{1k}, \ldots, \mathbf{a}_{n+p,k})'$, and $\mathbf{A} = (\mathbf{A}_1', \ldots, \mathbf{A}_K')'$. Then, the dual problem of (2.4.12) can be shown as

$$\max_{\mathbf{d}} \{ \mathbf{u}'\mathbf{d} \,|\mathbf{A}'\mathbf{d} = 0 \}, \tag{2.4.13}$$

where $\mathbf{d} = \text{vec}(\mathbf{d}_1, \ldots, \mathbf{d}_K)$, $\mathbf{d}_k = \text{vec}(\mathbf{d}_k^{(1)}, \mathbf{d}_k^{(2)})$, $\mathbf{d}_k^{(1)} = (d_{1k}, \ldots, d_{nk})' \in [\omega_k(\tau_k - 1), \omega_k\tau_k]^n$, $\mathbf{d}_k^{(2)} = (d_{n+1,k}, \ldots, d_{n+p,k})' \in [-\omega_k^2, \omega_k^2]^p$.

There are two methods, the simplex and the interior point [see Vanderbei, Meketon and Freedman (1986)], for solving (2.4.13). Here we opt for the latter due to its two advantages [Bassett and Koenker (1992) and Koenker and Park (1996)]: a) computational simplicity and natural extensions to nonlinear problems; and b) unlike the simplex-based method, the interior point algorithm can be shown to converge to the correct answer. Algorithmic details for the dual problem (2.4.13) proceed as follows:

1. For any initial feasible $\mathbf{d}$, e.g., $\mathbf{d} = \mathbf{0}$, following Meketon (1986), set an $n \times n$ diagonal matrix $\mathbf{D}_k^{(1)}$ with the ith diagonal element being $\min\{\omega_k\tau_k - d_{ik}, d_{ik} - \omega_k(\tau_k - 1)\}$, and a $p \times p$ diagonal matrix $\mathbf{D}_k^{(2)}$ with the ith diagonal element being $\min\{\omega_k^2 - d_{ik}, d_{ik} + \omega_k^2\}$. Let $\mathbf{D}_k = \text{diag}(\mathbf{D}_k^{(1)}, \mathbf{D}_k^{(2)})$, for $k = 1, \ldots, K$, $\mathbf{D} = \text{diag}(\mathbf{D}_1, \ldots, \mathbf{D}_K)$, $\mathbf{s} = \mathbf{D}^2(\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{D}^2\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^2)\mathbf{u}$, and $\mathbf{t} = (\mathbf{A}'\mathbf{D}^2\mathbf{A})^{-1}\mathbf{A}'\mathbf{D}^2\mathbf{u}$.

2. Set $\mathbf{d}^* = \mathbf{d} + (a_0/\gamma)\mathbf{s}$, where $\mathbf{s} = \text{vec}(\mathbf{s}_1, \ldots, \mathbf{s}_K)$, $\mathbf{s}_k = (s_{1k}, \ldots, s_{n+p,k})'$,

$\gamma = \max(\gamma_1, \ldots, \gamma_K)$, $\gamma_k = \max(\gamma_k^{(1)}, \gamma_k^{(2)})$,

$$\gamma_k^{(1)} = \max_{1 \le i \le n} (\max\{s_{ik}/(\omega_k \tau_k - d_{ik}), -s_{ik}/(d_{ik} - \omega_k(\tau_k - 1))\}),$$

$$\gamma_k^{(2)} = \max_{n+1 \le i \le n+p} (\max\{s_{ik}/(\omega_k^2 - d_{ik}), -s_{ik}/(d_{ik} + \omega_k^2)\}),$$

for $k = 1, \ldots, K$, and $a_0 \in (0, 1)$ is a constant chosen to insure feasibility. As suggested by Koenker and Park (1996), we take $a_0 = 0.97$.

3. Set $\mathbf{d} = \mathbf{d}^*$. Updating $\mathbf{D}$, $\mathbf{s}$ and $\mathbf{d}$ continues the iteration.

After solving (2.4.13) using the above interior point algorithm, we arrive at the next loop which uses the current value $\boldsymbol{\theta} = \boldsymbol{\theta}^{(r+1)}$ for the primal problem in (2.4.12). This leads to the updated dual problem (2.4.13) with $y_{ik}^* = y_i - l_{ik}(\boldsymbol{\theta}^{(r+1)})$ and $\mathbf{a}'_{ik} = \nabla l_{ik}(\boldsymbol{\theta}^{(r+1)})$ for $i = 1, \ldots, n$. The current $\mathbf{d}$ should be adjusted to ensure that it is feasible for the new value of $\mathbf{A}$. Similar to Koenker and Park (1996), we project the current $\mathbf{d}$ onto the null space of the new $\mathbf{A}$, i.e. $\widehat{\mathbf{d}} = (\mathbf{I} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\mathbf{d}$, and then shrink it to insure that $\mathbf{d}_k^{(1)} \in [\omega_k(\tau_k - 1), \omega_k \tau_k]^n$ and $\mathbf{d}_k^{(2)} \in [-\omega_k^2, \omega_k^2]^p$. So the adjusted $\mathbf{d}$ becomes $\mathbf{d} = \widehat{\mathbf{d}}/(m + \delta)$ for some $\delta > 0$, where $m = \max(m_1, m_2, \ldots, m_K)$, with $m_k = \max(m_k^{(1)}, m_k^{(2)})$, $m_k^{(1)} = \max_{1 \le i \le n} \{\max(\frac{\hat{d}_{ik}}{\omega_k(\tau_k - 1)}, \frac{\hat{d}_{ik}}{\omega_k \tau_k})\}$ and $m_k^{(2)} = \max_{n+1 \le i \le n+p} \{|\hat{d}_{ik}/\omega_k^2|\}$.

As noted by Koenker and Park (1996), the difficulty with the above method is twofold: first, it is required to fully solve a linearized problem (2.4.9) or equivalently (2.4.12) at each iteration; second, the resulting search directions may actually be inferior to directions determined by incomplete solutions to the sequence of linearized problems. As they suggested, when f_i is nonlinear, there is no longer a compelling argument for fully solving (2.4.9), and only a few iterations to refine the dual vector is preferable. This reduces the computational burden.

Next, we consider the minimization problem (2.3.6). By Taylor's expansion for the SCAD penalty at an initial consistent estimate $\boldsymbol{\beta}^0$ (for example the common L_1-norm estimate), we have

$$p_{\lambda_n}(|\beta_j|) \approx p'_{\lambda_n}(|\beta_j^0|)|\beta_j| + \{p_{\lambda_n}(|\beta_j^0|) - p'_{\lambda_n}(|\beta_j^0|)|\beta_j^0|\},$$

where $p_{\lambda_n}(|\beta_j^0|) - p'_{\lambda_n}(|\beta_j^0|)|\beta_j^0|$ is a constant. Therefore, minimization in (2.3.6) is reduced to

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \rho_{\tau_k}(y_i - l_{ik}(\boldsymbol{\theta})) + n \sum_{j=1}^{p} p'_{\lambda_n}(|\beta_j^0|)|\beta_j|,$$

which can be solved using the same algorithm as for (2.4.8). Update the initial value for $\boldsymbol{\beta}$ and do iterations until convergence, where a few steps can lead to convergence since $\boldsymbol{\beta}^0$ is close to the true parameter.

2.5 Conclusion

In this chapter, we have proposed the WQR for model (2.1.1) with a fixed number of parameters and used the the adaptive WQR-LASSO and the WQR-SCAD to study variable/parameter selection problem. We have demonstrated that the adaptive WQR-LASSO and the WQR-SCAD estimators all enjoy oracle properties and are asymptotically equivalent. We have also developed an efficient algorithm to implement the WQR, but simulations and real data analysis are delegated to next chapter.

Chapter 3

Weighted QR with infinite parameters and model selection

3.1 Introduction

Model selection with a fixed number of parameters has been widely pursued in the last decades. However, to reduce possible modeling biases, many variables are introduced in practice. As noted in Huber (1973, 1988), Portnoy (1984, 1988) and Donoho (2000), the number of parameters p is often large and should be modeled as p_n, which tends to ∞. Fan and Peng (2004) and Lam and Fan (2007) advocated that in most model selection problems the number of parameters should be large and grow with the sample size. In a recent seminal paper, Fan and Lv (2008) also studied model selection for linear models with the number of parameters higher than the sample size. Therefore, in this chapter the dimensionality p is allowed to be independent or dependent of the sample size n.

QR estimation of model (2.1.1) with a fixed p has received much attention. See Oberhofer (1982), Dupačová and Wets (1988), Powell (1986, 1991), Jurečková and Procházka (1994), and Wang (1995), among others, but for model (2.1.1) with a diverging number of parameters, to the best of our knowledge, there is no formal work using QR in the literature. Hence, the WQR for model (2.1.1) is

new and fundamental. We will address the issue of variable/parameter selection using the penalized WQR with the adaptive LASSO and SCAD penalties. The resulting method is robust against outliers and heavy-tailed error distributions, like the Cauchy distribution, and efficient as nearly as the MLE when the error is normal. In addition, the weights in the WQR are allowed to be negative, so the proposed WQR is essentially different from the common QR and the CQR (see also Section 2). When the weights are all equal and the model is linear with a fixed number of parameters, our method reduces to that of Zou and Yuan (2008a) if the LASSO penalty is employed. Since the proposed WQR involves a vector of weights, we develop a data-driven weighting strategy which maximizes the efficiency of the WQR estimators. The resulting estimation is adaptive in the sense that it performs asymptotically the same as if the theoretically optimal weights were used.

The penalized WQR estimators admit no close form and involve minimizing complicate nonlinear functions, so it is challenging to derive asymptotic properties and to implement the methodology. Theoretically, we will establish asymptotic normality of the resulting estimators and show their optimality, no matter whether the error variance is finite or not. Practically, we will develop an algorithm for fast implementation of the proposed methodology. This algorithm solves a succession of (penalized) linearized WQR problems, each of whose dual problems is derived. We will extend the "interior point algorithm" [Vanderbei, Meketon and Freedman (1986) and Koenker and Park (1996)] to solve these dual problems. The resulting algorithm is easy to implement. Simulations endorse our discovery.

This chapter is organized as follows. In Section 3.2 we introduce the penalized WQR for model (3.2.1) with a diverging number of parameters and study the variable/parameter selection problem using the adaptive LASSO and SCAD

penalties. In Section 3.3 we present numerical studies which include the choice of tuning parameters, simulations and a real example. Conclusion is given in section 3.4. Finally, we give proofs of the theorems in the Appendix B.

3.2 Weighted QR with a diverging number of parameters

As discussed in the introduction, in most model selection problems the number of parameters should be large and grow with the sample size. In practice, many variables are introduced to reduce possible modeling biases. The variable/parameter selection methods in Section 2 are limited to the finite-parameter setting. In this section, we allow the number of regression parameters tends to infinity as the sample size increases and study the sampling properties of the penalized WQR estimators with adaptive LASSO and SCAD penalties. To stress dependence on the sample size, we denote the p_n-vector of parameters by $\boldsymbol{\beta}_n = (\beta_{n1}, \ldots, \beta_{np_n})'$ and rewrite the nonlinear model (2.1.1) as:

$$y_i = f(\mathbf{x}_i, \boldsymbol{\beta}_n) + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{3.2.1}$$

Without loss of generality, we assume that the first s_n components of the true regression coefficients $\boldsymbol{\beta}_n^*$, denoted by $\boldsymbol{\beta}_{n1}^*$, do not vanish and the remaining $p_n - s_n$ components, denoted by $\boldsymbol{\beta}_{n2}^*$, are zeros. Let $f_{ni}^* = f(\mathbf{x}_i, \boldsymbol{\beta}_{ni}^*)$ and $\boldsymbol{\nabla} f_{ni}^* = [\partial f(\mathbf{x}_i, \boldsymbol{\beta}_n)/\partial \boldsymbol{\beta}_n] \mid_{\boldsymbol{\beta}_n = \boldsymbol{\beta}_n^*}$.

3.2.1 Regularity conditions

The following conditions are needed for our theoretical results.

(i) Regularity condition on penalty. Let $n_p = n/p_n$,

$$a_n = \max_{1 \le j \le p_n} \{p'_{\lambda_n}(|\beta_{nj}^*|), \beta_{nj}^* \neq 0\},$$

and $b_n = \max_{1\le j\le p_n}\{p'_{\lambda_n}(|\beta^*_{nj}|), \beta^*_{nj} \neq 0\}$. The conditions on penalty functions we require are:

(A_1) $\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n > 0$;

(A_2) $a_n = O(n^{-\frac{1}{2}})$;

(A_3) $b_n \to 0$ as $n \to +\infty$.

(A_4) there are constants C and D such that $|p''_{\lambda_n}(\theta_1) - p''_{\lambda_n}(\theta_2)| \le D|\theta_1 - \theta_2|$, where $\theta_1, \theta_2 > C\lambda_n$.

Conditions (A_1)-(A_4) are also the regularity conditions on the penalty given in Fan and Peng (2004).

(ii) Regularity condition on regression function.

(B_1) There is a large enough open subset $\Omega_n \in \mathbf{R}^{p_n}$ which contains the true parameter point $\boldsymbol{\beta}^*_n$, such that for all $\mathbf{x}_i$ the second derivative matrix $\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ of $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ with respect to $\boldsymbol{\beta}_n$, satisfies that

$$\|\nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_{n1}) - \nabla^2 f(\mathbf{x}_i, \boldsymbol{\beta}_{n2})\| \le M(\mathbf{x}_i)\|\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n2}\|$$

$$|\partial^2 f(\mathbf{x}_i, \boldsymbol{\beta}_n)/(\partial\beta_{nj}\partial\beta_{nk})| \le N_{jk}(\mathbf{x}_i)$$

for all $\boldsymbol{\beta}_n \in \Omega_n$, and $E[M^2(\mathbf{x}_i)] < \infty$, $E[N^2_{jk}(\mathbf{x}_i)] < C_1 < \infty$ for all j, k.

(B_2) $\mathrm{var}((\nabla f^*_{ni})^{\otimes 2}) = \mathbf{G}_n$, and $0 < d_1 < \lambda_{\min}(\mathbf{G}_n) \le \lambda_{\max}(\mathbf{G}_n) < d_2 < \infty$, for all n, where $\lambda_{\min}(\mathbf{G}_n)$ and $\lambda_{\max}(\mathbf{G}_n)$ represent the smallest and largest eigenvalues of $\mathbf{G}_n$, respectively.

(B_3) $\beta^*_{n1}, \beta^*_{n2}, \ldots, \beta^*_{ns_n}$ satisfy

$$\min_{1\le j\le s_n} |\beta^*_{nj}|/\lambda_n \to \infty \quad \text{as} \quad n \to \infty.$$

(B_4) $\beta^*_{n1}, \beta^*_{n2}, \ldots, \beta^*_{ns_n}$ satisfy

$$\min_{1\le j\le s_n} |\beta^*_{nj}|/(\sqrt{n}h_n) \to \infty \ \text{ as } \ n \to \infty.$$

Condition (B_1) is a natural extension to Condition (c) in §2. Condition (B_2) is similar to the condition (F) placed on the information matrix of Fan and Peng (2004). Condition (B_3) is the same condition of Fan and Peng (2004) used to obtain the oracle property. Condition (B_4) acts the same role as (B_3) does, which is used to obtain the oracle property when using the adaptive LASSO penalty.

(iii) Regularity condition on error distribution.

(C) The error ε_i has the distribution function $G(\cdot)$ and density function $g(\cdot)$. The density function g is positive and continuous at the τ_k-th quantiles $b^*_{\tau_k}$.

3.2.2 Model selection with SCAD penalty

Similar to (2.3.6), the WQR-SCAD estimators for model (3.2.1) are defined as

$$(\hat{b}_{\tau_1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}_n) = \arg\min_{\mathbf{b},\boldsymbol{\beta}_n} Q^{SC}_n(\boldsymbol{\beta}_n, \mathbf{b}),$$

where $Q^{SC}_n(\boldsymbol{\beta}_n, \mathbf{b}) = L_n(\boldsymbol{\beta}_n, \mathbf{b}) + n\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$. The following theorems establish consistency and asymptotic normality of the SCAD penalized estimator.

Theorem 3.2.1. (Consistency) *Suppose that the density $g(\cdot)$ satisfies Condition (C), penalty function $p_{\lambda_n}(\cdot)$ satisfies Conditions (A_2)-(A_4), and regression function $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ satisfies Conditions (B_1)-(B_2). If $p_n^3/n \to 0$ as $n \to \infty$, then there is a local minimizer $\hat{\boldsymbol{\beta}}_n$ of $Q^{SC}_n(\boldsymbol{\beta}_n, \mathbf{b})$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*_n\| = O_p(\sqrt{p_n}(n^{-\frac{1}{2}} + a_n))$.*

Denote

$$\mathbf{b}_n = \left\{p'_{\lambda_n}(|\beta^*_{n1}|)\mathrm{sgn}(\beta^*_{n1}), \ldots, p'_{\lambda_n}(|\beta^*_{ns_n}|)\mathrm{sgn}(\beta^*_{ns_n})\right\}'$$

and

$$\boldsymbol{\Sigma}_{\lambda_n} = \text{diag}\big\{p''_{\lambda_n}(\boldsymbol{\beta}^*_{n1}), \ldots, p''_{\lambda_n}(\boldsymbol{\beta}^*_{ns_n})\big\}.$$

Let $\mathbf{G}_{n11}$ be the $s_n \times s_n$ sub-matrix of $\mathbf{G}_n$ corresponding to $\boldsymbol{\beta}_{n1}$, and let $\mathbf{A}_n$ be a $q \times s_n$ matrix such that $\mathbf{A}_n\mathbf{A}'_n \to \mathbf{B}$, where $\mathbf{B}$ is a $q \times q$ nonnegative symmetric matrix and q is a fixed positive integer.

Theorem 3.2.2. (Oracle property) *Suppose the conditions in Theorem 3.2.1, Condition (A_1), and Condition (B_3) hold. If $\lambda_n \to 0$, $\sqrt{n_p}\lambda_n \to \infty$, and $p_n^3/n \to 0$ as $n \to \infty$, then, with probability tending to 1, the root-n_p consistent local minimizer $\hat{\boldsymbol{\beta}}_n = (\hat{\boldsymbol{\beta}}'_{n1}, \hat{\boldsymbol{\beta}}'_{n2})'$ in Theorem 3.2.1 must satisfy*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$; *and*

(ii) Asymptotic normality:

$$\begin{aligned} & \sqrt{n}\mathbf{A}_n\mathbf{G}_{n11}^{-\frac{1}{2}}\big(\mathbf{G}_{n11} + \boldsymbol{\Sigma}_{\lambda_n}/\boldsymbol{\omega}'\mathbf{g}\big) \\ \times \quad & [(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}^*_{n1}) + \big(\mathbf{G}_{n11} + \frac{\boldsymbol{\Sigma}_{\lambda_n}}{\boldsymbol{\omega}'\mathbf{g}}\big)^{-1}\mathbf{b}_n/\boldsymbol{\omega}'\mathbf{g}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B}). \end{aligned}$$

Remark 3.2.1. Fan and Peng (2004) set the oracle property condition for the penalized likelihood estimator with the assumption of convergence rate $p_n^5/n \to 0$. As they noted, this condition may be weakened to $p_n^3/n \to 0$, which is verified by the current WQR-SCAD estimation.

3.2.3 Variable selection with adaptive-LASSO

Denote by $\tilde{\boldsymbol{\beta}}_n$ the resulting solution to $\min_{\boldsymbol{\beta}_n, \mathbf{b}} L_n(\boldsymbol{\beta}_n, \mathbf{b})$. Then using the same argument as for Theorem 3.2.1, $\tilde{\boldsymbol{\beta}}_n$ is $\sqrt{n_p}$-consistent. Thus, we can use $\tilde{\boldsymbol{\beta}}_n$ to construct the adaptive LASSO penalty. Let $\tilde{w}_{nj} = |\tilde{\beta}_{nj}|^{-\gamma}$ for some $\gamma > 0$, and

define the adaptive WQR-LASSO estimator as

$$(\hat{b}_{\tau 1}, \ldots, \hat{b}_{\tau_K}, \hat{\boldsymbol{\beta}}_n^{AL}) = \arg\min_{\mathbf{b}, \boldsymbol{\beta}_n} Q_n^{AL}(\boldsymbol{\beta}_n, \mathbf{b}),$$

where

$$Q_n^{AL}(\boldsymbol{\beta}_n, \mathbf{b}) = L_n(\boldsymbol{\beta}_n, \mathbf{b}) + nh_n \sum_{j=1}^{p_n} \tilde{w}_{nj} |\beta_{nj}|.$$

Theorem 3.2.3. (Consistency) *Suppose that the density $g(\cdot)$ satisfies Condition (C), and regression function $f(\mathbf{x}_i, \boldsymbol{\beta}_n)$ satisfies Conditions (B_1)-(B_2). If $p_n^3/n \to 0$ and $\sqrt{n}h_n \to 0$ as $n \to \infty$, then there is a local minimizer $\hat{\boldsymbol{\beta}}_n^{AL}$ of $Q_n^{AL}(\boldsymbol{\beta}_n, b_\tau)$ such that $\|\hat{\boldsymbol{\beta}}_n^{AL} - \boldsymbol{\beta}_n^*\| = O_p(n_p^{-1/2})$.*

Denote $\mathbf{d}_n = (\mathrm{sgn}(\beta_{n1}^*)/|\tilde{\beta}_{n1}|^\gamma, \ldots, \mathrm{sgn}(\beta_{ns_n}^*)/|\tilde{\beta}_{ns_n}|^\gamma)'$.

Theorem 3.2.4. (Oracle property) *Suppose the conditions of Theorem 3.2.3 and the condition (B_4) hold. If $h_n n_p^{(\gamma+1)/2} \to \infty$, then, with probability tending to 1, the $\sqrt{n_p}$-consistent local minimizer $\hat{\beta}_n^{AL} = (\{\hat{\boldsymbol{\beta}}_{n1}^{AL}\}', \{\hat{\boldsymbol{\beta}}_{n2}^{AL}\}')'$ in Theorem 3.2.3 must satisfy*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_{n2}^{AL} = \mathbf{0}$; *and*

(ii) Asymptotic normality:

$$\sqrt{n}\mathbf{A}_n \mathbf{G}_{n11}^{\frac{1}{2}}[(\hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}_{n1}^*) + \mathbf{G}_{n11}^{-1} h_n \mathbf{d}_n / \boldsymbol{\omega}'\mathbf{g}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B}).$$

Note that when n is finite and large enough, $\boldsymbol{\Sigma}_{\lambda_n} = \mathbf{0}$ and $\mathbf{b}_n = \mathbf{0}$ for the WQR-SCAD, but $\mathbf{d}_n$ is not zero and hence the bias term for the WQR-LASSO in Theorem 3.2.4 cannot be ignored. By Condition (B_4), $\sqrt{n}h_n\mathbf{d}_n \to \mathbf{0}$, as $n \to \infty$. Hence, Theorem 3.2.2 (ii) and 3.2.4 (ii) become

$$\sqrt{n}\mathbf{A}_n \mathbf{G}_{n11}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}_{n1}^*) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B})$$

and

$$\sqrt{n}\mathbf{A}_n\mathbf{G}_{n11}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}_{n1}^{*}) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B}),$$

respectively. This demonstrates that the adaptive WQR-LASSO and WQR-SCAD estimators have the same asymptotic efficiency as the WQR estimator of $\boldsymbol{\beta}_{n1}$ based on the submodel with $\boldsymbol{\beta}_{n2} = 0$ known in advance. That is, the penalized WQR estimators enjoy the oracle properties.

3.3 Numerical studies

3.3.1 Choice of the tuning parameters

For the penalized WQR estimators, one has to select the tuning parameters λ_n and h_n, respectively for the SCAD and LASSO penalties. The two parameters can be chosen using the same method. We here focus on the choice of λ_n.

There are several methods for selecting λ_n, which include the generalized cross-validation (GCV) criterion [Wang, Li and Tsai (2007)] and the Schwartz Information Criterion (SIC) [see Koenker, Ng and Portnoy (1994) and Zou and Yuan (2008b)]. Since the resulting estimators depend on λ_n, we denote the estimators by $(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n})$ to stress such dependence. Applying the SIC method, we propose to select λ_n by minimizing

$$SIC(\lambda_n) = \log\Big\{\frac{1}{nK}L_n(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n})\Big\} + \frac{\log(nK)}{2nK}df(\lambda_n)$$

over λ_n, where $df(\lambda_n)$ is the effective degrees of freedom of the fitted model. For a given tuning parameter λ_n, we define a set $\mathcal{E}_{\lambda_n}$ as

$$\mathcal{E}_{\lambda_n} = \{(k,i) : y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_{\lambda_n}) - \hat{b}_{\lambda_n,\tau_k} = 0\}.$$

Let $|\mathcal{E}_{\lambda_n}|$ denote the size of the set $\mathcal{E}_{\lambda_n}$. Koenker, Ng and Portnoy (1994) conjectured that $|\mathcal{E}_{\lambda_n}|$ is the effective degrees of freedom in the quantile regression. Li, Liu and

Zhu (2007) and Li and Zhu (2008) verified the conjecture. Therefore, we use

$$SIC(\lambda_n) = \log\Big\{\frac{1}{nK} L_n(\hat{\boldsymbol{\beta}}_{\lambda_n}, \hat{\mathbf{b}}_{\lambda_n})\Big\} + \frac{\log(nK)}{2nK}|\mathcal{E}_{\lambda_n}|.$$

The final turning parameter is estimated by $\hat{\lambda}_n = \arg\min_{\lambda_n} SIC(\lambda_n)$.

3.3.2 Simulations

In this section we conduct simulations to investigate finite sample performance of the WQR estimation and the associated model selection. The following exponential regression model is used:

$$y = 1 + b\exp(\mathbf{c}'\mathbf{x}) + \varepsilon,$$

where b and $\mathbf{c} = (c_1, c_2, c_3)'$ are parameters, ε is the error. The true values of parameters are set as $b = 1.5$, and $\mathbf{c} = (-0.6, -0.8, -0.7)'$.

When the penalized WQR methods are considered, we shall allow the lengths of $\mathbf{c}$ and the relevant $\mathbf{x}$ increasing with the sample size, by setting

$$\mathbf{c} = (-0.6, -0.8, -0.7, 0, \ldots, 0)'.$$

The following two penalties are employed:

(i) the adaptive LASSO penalty, defined by $nh_n \sum_{j=1}^{p_n} |\beta_j|/|\tilde{\beta}_j|$,

(ii) the SCAD penalty, defined by $\sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_j|)/|\tilde{\beta}_j|$,

where h_n and λ_n are tuning parameters and $\tilde{\beta}_j$'s are consistent estimators of β_j's. In simulations, the tuning parameters are determined by the *SIC* method.

Let $\boldsymbol{\beta} = (b, \mathbf{c}')'$ be the $p_n \times 1$ vector of parameters in the working model. We draw from the working model 400 samples of sizes 100 and 400 with $p_n = [n^{1/3}] + 3$. In each simulation, the first component of $\mathbf{x}$ is generated from $U[-1, 1]$, and the remaining components of $\mathbf{x}$ are generated from the joint normal distribution

with the pairwise correlation coefficient being 0.4 and the standard normal as the marginal. We consider three sets of errors: $N(0,1)$, $t(3)$, and $\chi^2(3)$. All of them are centralized and scaled so that the medians of the absolute errors are ones.

We compare four estimation methods: the penalized QR (L_1, CQR, and WQR) estimation and the WQR-oracle estimation. In each simulation the "root of mean squared error (RMS)" for different penalized QR estimators and their summation are calculated, and their average over simulations are reported in Tables 1-3, where Σ denotes the sum of RMS for all components in $\boldsymbol{\beta}$. Therefore, better methods should have smaller values in the tables. As expected, the WQR-oracle estimator performs the best, the penalized WQR performs comparably to the oracle estimator, and the penalized L_1 is the worst. This exemplifies the theory about the penalized WQR estimation: asymptotically the penalized WQR estimation performs as well as if the correct submodel were known. The penalized WQR performs much better than the penalized CQR and L_1 when the error is chi-squared, but the two methods are comparable when the errors are normal and $t(3)$. In Table 4 we report the frequency that zero coefficients are set to zero correctly if their estimates are less than 10^{-8}. It seems that the WQR-SCAD has higher probability of correctly identifying zero coefficients than the WQR-LASSO.

3.3.3 A real example

The patients in hospital faces an infection risk. To study the Efficacy of Nosocomial Infection Control (SENIC), the Hospital Infections Program was conducted by Robert W. Haley and his collaborators, Center for Infectious Diseases, Centers for Disease Control, Atlanta, Georgia 30333. This resulted in the SENIC dataset for the 1975-76 study period, consisting of a random sample of 113 hospitals selected

Table 3.1: *RMS (multiplied by 10^3) of penalized estimators under the normal error;* $\boldsymbol{\omega}_{opt} = (0.6362, 0.4365, 0.6362)'$.

	$n = 100$					$n = 400$				
Method	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SCAD-L_1	427	117	97	97	791	144	45	34	32	259
SCAD-CQR	339	98	80	81	652	126	40	29	28	231
SCAD-WQR	329	98	79	82	637	124	40	28	27	231
LASSO-L_1	412	112	94	94	782	145	45	34	32	266
LASSO-CQR	331	97	79	82	656	126	40	29	28	241
LASSO-WQR	329	98	78	81	647	124	40	28	27	240
WQR-oracle	321	99	78	80	578	124	40	28	28	219

Table 3.2: *RMS (multiplied by 10^3) of penalized estimators under the normalized t(3) error;* $\boldsymbol{\omega}_{opt} = (0.4856, 0.7269, 0.4856)'$.

	$n = 100$					$n = 400$				
Method	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SCAD-L_1	412	122	104	89	806	154	45	39	33	289
SCAD-CQR	393	117	94	82	755	145	42	35	31	272
SCAD-WQR	386	117	95	83	748	145	42	35	31	268
LASSO-L_1	411	123	104	90	819	156	46	39	34	298
LASSO-CQR	392	115	93	82	763	145	43	35	31	283
LASSO-WQR	385	117	94	84	760	144	42	36	31	273
WQR-oracle	373	116	93	82	664	144	42	35	31	253

Table 3.3: *RMS (multiplied by 10^3) of penalized estimators under the normalized $\chi^2(3)$ error;* $\boldsymbol{\omega}_{opt} = (0.9916, 0.1115, 0.0658)'$.

	$n = 100$					$n = 400$				
Method	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ	$\hat{b}$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	Σ
SCAD-L_1	374	118	96	89	688	133	40	31	28	237
SCAD-CQR	329	101	82	75	617	109	33	26	23	199
SCAD-WQR	249	76	70	65	466	93	28	23	20	166
LASSO-L_1	379	115	95	88	697	134	41	31	29	242
LASSO-CQR	330	101	82	75	626	109	33	26	23	200
LASSO-WQR	250	77	70	65	473	93	28	23	20	170
WQR-oracle	247	77	70	65	460	93	28	23	20	164

Table 3.4: *The frequency of zero coefficients set to zero correctly under the adaptive LASSO and SCAD penalties.*

		$n = 100$		$n = 400$	
Method	error	LASSO	SCAD	LASSO	SCAD
	normal	0.519	0.556	0.705	0.762
L_1	$t(3)$	0.320	0.393	0.559	0.644
	$\chi^2(3)$	0.663	0.705	0.700	0.761
	normal	0.531	0.567	0.764	0.851
CQR	$t(3)$	0.453	0.528	0.598	0.691
	$\chi^2(3)$	0.618	0.671	0.808	0.867
	normal	0.543	0.598	0.718	0.804
WQR	$t(3)$	0.452	0.546	0.674	0.759
	$\chi^2(3)$	0.839	0.890	0.862	0.921

from the original 338 hospitals surveyed (see Kutner *et al.* 2005). For each single hospital there are 11 variables:

- Infection risk (y): Average estimated probability of acquiring infection in hospital.
- Length of stay (x_1): Average length of stay of all patients in hospital (in days).
- Age (x_2): Average age of patients (in years).
- Routine culturing ratio (x_3): Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100.
- Routine chest X-ray ratio (x_4): Ratio of number of X-rays performed to numbers of patients without signs or symptoms of pneumonia, times 100.
- Number of beds (x_5): Average number of beds in hospital during study pe-

riod.

- Medical school affiliation (x_6): 1=Yes, 2=No.
- Region (x_7-x_9): Geographic region, where: 1=NE, 2=NC, 3=S, 4=W.
- Average daily census (x_{10}): Average number of patients in hospital per day during study period.
- Number of nurses (x_{11}): Average number of full-time equivalent registered and licensed practical nurses during study period (number full time plus one half the number part time).
- Available facilities and services (x_{12}): Percent of 35 potential facilities and services that are provided by the hospital.

Now we study whether the infection risk depends on the possible influential factors and target at providing a good estimate for the infection risk, after adjusting contributions from confounding factors. Since the medical school affiliation and region are categorical, we introduce a dummy variable x_6 for the medical school affiliation and three dummy variables (x_7, x_8, x_9) for the region as covariates. Note that the response y (infection risk) is the average estimated probability of acquiring infection in hospital. It is sensible to use the following logistic model with all of covariates,

$$y_i = \frac{\exp(\beta_0 + \sum_{i=1}^{12} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{12} \beta_i x_i)} + \varepsilon_i, \quad i = 1, \ldots, 113,$$

to model the relationship between the infection risk and all possible infection factors, where all of covariates are used to reduce possible modeling biases and the number of non-zero parameters is assumed to depend on the sample size.

We apply the L_2-penalized least squares estimation (LSE) and the penalized CQR and WQR methods with adaptive LASSO and SCAD penalties to select the non-zero parameters or significant variables. The SIC criterion (Section 5.1) is applied to choose the tuning parameters. The results of variable selection are presented in Table 3.5. From Table 3.5, we can see that penalized SCAD and penalized LASSO methods both select four variables: age (x_2), routine chest X-ray ratio (x_4), number of beds (x_5), and average daily census (x_{10}), but the penalized LSE selects all variables (note that x_7-x_9 together represents the region). Similar to the ridge regression for linear models, the LSE with L_2-penalty fails in shrinking any coefficients directly to zero for the nonlinear model.

Table 3.5: *Estimates and standard errors (in parentheses, multiplied by* 10^4*)*

Penalty	L_2	LASSO		SCAD	
Method	LSE	CQR	WQR	CQR	WQR
x_1	574 (335)	0 (-)	0 (-)	0 (-)	0 (-)
x_2	-667 (105)	-743 (113)	-705 (102)	-745 (114)	-713 (103)
x_3	55 (40)	0 (-)	0 (-)	0 (-)	0 (-)
x_4	-31 (23)	-25 (36)	-32 (33)	-25 (37)	-23 (32)
x_5	-18 (12)	-12 (17)	-5 (16)	-10 (17)	-5 (16)
x_6	229 (1302)	0 (-)	0 (-)	0 (-)	0 (-)
x_7	66 (1512)	0 (-)	0 (-)	0 (-)	0 (-)
x_8	-100 (1359)	0 (-)	0 (-)	0 (-)	0 (-)
x_9	250 (1343)	0 (-)	0 (-)	0 (-)	0 (-)
x_{10}	15 (14)	23 (21)	12 (21)	21 (21)	12 (20)
x_{11}	9 (7)	0 (-)	0 (-)	0(-)	0 (-)
x_{12}	-14 (46)	0 (-)	0 (-)	0 (-)	0 (-)

To check the significance of the selected model, we consider the hypothesis testing problem:

$$H_0 : \beta_2 = \beta_4 = \beta_5 = \beta_{10} = 0 \text{ versus}$$

$$H_1 : \text{at least one of them are non-zeros.}$$

Use the LSE to estimate the parameters in the null and alternative models and let $SSE(H_0)$ and $SSE(H_1)$ be the residual sum of squares under H_0 and H_1, respectively. Define the F statistic:

$$F = \frac{SSE(H_0) - SSE(H_1)}{df_0 - df_1} \Big/ \frac{SSE(H_1)}{df_1},$$

where $df_0 = n - 1$ and $df_1 = n - 5$ are degrees of freedom for the null and alternative models, respectively. The approximate null distribution of F-statistic is $F(df_0 - df_1, df_1)$. The realized value of F is calculated as 124.541 with approximate p-value equal to zero. Therefore, the selected model is significant.

3.4 Conclusion

In this chapter, we have suggested the penalized-WQR for model (3.2.1) with parameters depending on the sample size and study the variable/parameter selection problem by the adaptive LASSO and SCAD penalties. We have established the asymptotic properties of penalized-WQR estimators and proved that these estimators all enjoy oracle properties. We have also proposed an algorithm to implement the penalized WQR and analyzed a real data. Simulation results and SENIC dataset analysis all endorse the use the proposed methodology.

Chapter 4

Quantile Regression and its application in DTARCH models

4.1 Introduction

Modelling volatility is important in financial data analysis. One of the most widely used tools in modelling the changing volatility is the autoregressive conditional heteroscedasticity (ARCH) model pioneered by Engle (1982). ARCH models and its extensions have been widely applied in finance and econometrics (Bollerslev et al., 1992, Bera and Higgins, 1993, Bollerslev et al., 1994, and Fan and Yao, 2003). Li and Li (1996) proposed a double-threshold autoregressive conditional heteroscedastic (DTARCH) model to study the piecewise linear patterns of the conditional mean and the conditional variance. They studied model identification, estimation and diagnostic check based on the maximum likelihood principle. This approach is useful for detecting nonlinear structures such as asymmetric behavior in the mean and the volatility of an asset return, and heteroscedasticity with clustering in the volatility. In practice, it is observed that financial returns tend to have thicker tails than normal distributions. Note that misspecification of the conditional distribution in the likelihood approach may create serious problems in parameter estimation. Moreover, likelihood based testing methods may fail in

detecting false structures in the conditional variance of asset return. It is worth investigating robust modelling techniques without specific distribution assumptions. This motivates us to consider DTARCH models for conditional scale based on quantile regression. The advantage of such an approach was discussed in Koenker and Zhao (1996) for ARCH models.

Quantiles regression (QR) is a statistical technique designed to estimate, and conduct inference about conditional quartile functions. The basic motivation for using quantiles rather than simple mean regression is that the stochastic relationship between random variables can be portrayed much better and with much more accuracy. See for example Chaudhuri, Doksum and Samarov (1997). The QR provides more robust and consequently more efficient estimates than the mean regression when the error is non-normal (Koenker and Bassett, 1978; Koenker and Zhao, 1996). This approach has been widely used in time series analysis (see for example, Koenker and Zhao, 1996; Davis and Dunsmuir, 1997; and Peng and Yao, 2003), but not for the DTARCH models.

The existing work considered only the MLE and L_1 estimation for the DTARCH model. Jiang, Zhao and Hui (2001) and Hui and Jiang (2005) studied the ARCH models and the DTARCH models, respectively, for the conditional scale (standard deviation) based on L_1 regression.

For the DTARCH model, both the MLE of Li and Li (1996) and the L_1 estimation of Hui and Jiang (2005) are useful in practice. The former is efficient when the error is normal, but it is sensitive to outliers and not robust against the error distribution, while the latter is resistant to outliers in the Y-space but not efficient when the error is normal. Hence, there is a genuine need for us to study robust and efficient estimation of the model.

An important problem now is how to efficiently and robustly estimate, and to make inference about, the DTARCH models. Related results can be used to analyse various financial data including those highly related to economy. This motivates us to use the proposed "weighted quantile regression (WQR)" in Chapter 2 for analyzing the DTARCH models. The proposed WQR is more efficient than the traditional QR while inheriting robustness.

This chapter is organized as follows. In §4.2 we give a review on DTARCH models. In §4.3 we apply the WQR to analyze DTARCH models and establish the asymptotic properties of the WQR estimators derived, where Data-driven weights are introduced to maximize the asymptotic efficiency of the estimators. Computational aspects and simulation results are presented in §4.4. Conclusion is given in §4.5. Proofs of theorems are presented in the Appendix C.

4.2 Review on DTARCH models

Given a time series y_t, $t = 1, \cdots, n$, let $\mathcal{F}_t$ be the σ-field generated from the realized value $\{y_t, y_{t-1}, \cdots\}$ at time t. Assume that y_t is generated by

$$y_t = \mathbf{X}'_{t,j}\boldsymbol{\alpha}^{(j)} + \varepsilon_t \quad \text{if} \quad r_{j-1} < y_{t-d} \leq r_j, \tag{4.2.1}$$

where $j = 1, \cdots, m$; the delay parameter d is a positive integer; the threshold parameters r_j satisfy $-\infty = r_0 < r_1 < r_2 < \cdots < r_m = \infty$; $\mathbf{X}_{t,j} = (1, y_{t-1}, \cdots, y_{t-p_j})'$ is a $(p_j+1)\times 1$ vector of lagged variables; $\boldsymbol{\alpha}^{(j)} = (\alpha_0^{(j)}, \alpha_1^{(j)}, \cdots, \alpha_{p_j}^{(j)})'$ is a $(p_j+1)\times 1$ parameter vector. The stochastic error satisfies $\varepsilon_t = h_t(\boldsymbol{\beta})u_t$ with

$$h_t(\boldsymbol{\beta}) = \sum_{j=1}^{m} I_{t,j}[\beta_0^{(j)} + \beta_1^{(j)}|\varepsilon_{t-1}| + \cdots + \beta_{q_j}^{(j)}|\varepsilon_{t-q_j}|] \equiv \sum_{j=1}^{m} I_{t,j}\mathbf{Z}'_{t,j}\boldsymbol{\beta}^{(j)}, \tag{4.2.2}$$

where $I_{t,j} = I(r_{j-1} < y_{t-d} \leq r_j)$; $\mathbf{Z}_{t,j} = (1, |\varepsilon_{t-1}|, \ldots, |\varepsilon_{t-q_j}|)'$; the parameters in the conditional scales satisfy $\beta_0^{(j)} > 0$, $\beta_i^{(j)} \geq 0$ $(i = 1, \cdots, q_j)$; and the inno-

vations $\{u_t\}$ are independently identically distributed random variables with an unknown distribution $F(x)$ and a density function $f(u)$. For convenience, as in Tsay (1989) and Li and Li (1996), we refer to the model in (4.2.1) and (4.2.2) as a DTARCH$(p_1, \cdots, p_m; q_1, \cdots, q_m)$ model, where the first m integers p's represent the AR orders in the m regimes and the last m integers q's denote the ARCH orders. The interval $r_{j-1} < y_{t-d} \leq r_j$ is the j-th regime of y_t. The proposed model is similar to that in Li and Li (1996) where the conditional scale instead of the conditional variance is specified as the ARCH structure.

The distinguished features of the model in (4.2.1) and (4.2.2) are: (i) the conditional scale h_t is a piecewise linear function of the absolute values of the lagged errors and each piece has an ARCH structure, which depicts the clustering of deviations at different regions of the lagged variable y_{t-d}; (ii) the double-threshold structure extends Tong's threshold model in a natural way and is capable to capture nonlinear phenomena such as asymmetric cycles, jump resonance and amplitude-frequence dependence (see Tong and Lim, 1980); and (iii) no assumption on the form of error distribution enables a robust inference for the model.

Modeling the conditional scale is important. As noted by Bickel and Lehmann (1976), scale provides a more natural dispersion concept than variance, and also offers substantial advantages from the robustness viewpoint (see Bickel, 1978; Carroll and Ruppert, 1988). Therefore, model (4.2.2) is especially appropriate for QR modeling.

Let $\boldsymbol{\alpha} = \text{vec}(\boldsymbol{\alpha}^{(1)}, ..., \boldsymbol{\alpha}^{(m)})$ and $\boldsymbol{\beta} = \text{vec}(\boldsymbol{\beta}^{(1)}, \cdots, \boldsymbol{\beta}^{(m)})$. Denote by $\mathbf{X}_t = \text{vec}(I_{t,1}\mathbf{X}_{t,1}, \cdots, I_{t,m}\mathbf{X}_{t,m})$ and $\mathbf{Z}_t = \text{vec}(I_{t,1}\mathbf{Z}_{t,1}, \cdots, I_{t,m}\mathbf{Z}_{t,m})$. Then $h_t(\boldsymbol{\beta}) = \mathbf{Z}_t'\boldsymbol{\beta}$.

4.3 Quantile regression estimation of the DTARCH model

4.3.1 The purely conditional heteroscedastic linear model

For explicit exposure of our methodology, we first consider the DTARCH model without the AR part:

$$y_t = \varepsilon_t = h_t(\boldsymbol{\beta})u_t, \tag{4.3.3}$$

where $h_t(\boldsymbol{\beta})$ is the same as in (4.2.2).

Quantile Regression. Since the τ-th conditional quantile of ε_t given $\mathcal{F}_{t-1}$ is $Q_\tau(\varepsilon_t|\mathcal{F}_{t-1}) = h_t(\boldsymbol{\beta})F^{-1}(\tau)$. It is apparent that $\boldsymbol{\beta}$ is only identifiable up to a scale. For identifiability of $\boldsymbol{\beta}$, we assume that the first nonzero component of $\boldsymbol{\beta}$ is 1, that is, $\beta_0^{(1)} = 1$.

Note that $Q_\tau(h_t^{-1}(\boldsymbol{\beta})\varepsilon_t|\mathcal{F}_{t-1}) = F^{-1}(\tau) \equiv b_\tau$. Following the idea in Koenker and Bassett (1978), one can define the τ-th regression quantile estimator for $\tau \in (0,1)$ by minimizing

$$\sum_{t=s+1}^{n} \rho_\tau\{h_t^{-1}(\boldsymbol{\beta})\varepsilon_t - b_\tau\}, \tag{4.3.4}$$

over b_τ and $\boldsymbol{\beta}$, where $s = \max(q_1, \cdots, q_m)$ and $\rho_\tau(u) = u(\tau - I(u<0))$ is the check function and with derivative $\psi_\tau(u) = \tau - I(u<0)$ for $u \neq 0$. Let the resulting estimator be $\hat{\boldsymbol{\beta}}_0$.

In order to derive the asymptotic property of the proposed estimator, we introduce some notations and conditions. Let $\boldsymbol{\beta}^*$ be the true value of $\boldsymbol{\beta}$, b_τ^* be the τ-th quantile of u_t. Denote by $h_t = h_t(\boldsymbol{\beta}^*)$, $E(\frac{\mathbf{Z}_t}{h_t}) = \boldsymbol{\mu}$, $a_1 = E(u_t\psi_\tau(u_t - b_\tau^*))$, $b_1 = -E(u_t\delta(u_t - b_\tau^*))$, $b_2 = E(u_t^2\delta(u_t - b_\tau^*))$, where $\delta(.)$ is Dirac delta function such that $\int_{-\infty}^{\infty}\delta(x)F(x)dx = F(0)$ for any continuous function $F(x)$. Assume that

(a0) $E|y_t|^{2+\delta} < +\infty$ for $\delta > 0$ and y_t is strictly stationary and ergodic.

(a1) $E(\mathbf{Z}_t\mathbf{Z}_t' h_t^{-2}) = \mathbf{G}_2$ is positive definite.

(a2) The density of u_t, $f(u)$, is positive and continuous at b_τ.

Conditions (a0) is used to insure the ergodicity of $\mathbf{Z}_t$ such that the mean ergodic theorem for $\frac{\mathbf{Z}_t}{h_t}$ holds, that is $\lim_{n\to\infty} \frac{1}{n} \sum_{t=s+1}^{n} \frac{\mathbf{Z}_t}{h_t} \to_p E(\frac{\mathbf{Z}_t}{h_t})$. Conditions (a1) is the natural extension to the conditions for establishing the asymptotic normality of single quantile regression [Koenker (2005) and Zou and Yuan (2008a)].

Theorem 4.3.1. Suppose that the threshold and the delay parameters are known. Under assumptions $(a0)$-$(a2)$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*) = -\{(b_2 + 2a_1)\mathbf{G}_2 - \frac{b_1^2 \boldsymbol{\mu}\boldsymbol{\mu}'}{f(b_\tau^*)}\}^{-1}(\boldsymbol{\gamma}_n - \frac{b_1 \boldsymbol{\mu}}{f(b_\tau^*)} q_{n,0}) + o_p(1),$$

where $q_{n,0} = -n^{-1/2} \sum_{t=s+1}^{n} \psi_\tau(u_t - b_\tau^*)$ and $\boldsymbol{\gamma}_n = -n^{-1/2} \sum_{t=s+1}^{n} u_t \psi_\tau(u_t - b_\tau^*) \frac{\mathbf{Z}_t}{h_t}$.

The resulting estimator $\hat{\boldsymbol{\beta}}_0$ is, unfortunately, biased because $\boldsymbol{\gamma}_n$ is not asymptotically unbiased. To overcome this shortcoming, we define a modified form of QR estimator. Note that $\log(|\varepsilon_t|) = \log\{h_t(\beta)\} + e_t$, where $e_t = \log(|u_t|)$, the τ-th QR estimate of $\boldsymbol{\beta}$ can be obtained by minimizing

$$\sum_{t=s+1}^{n} \rho_\tau(\log|\varepsilon_t| - \log\{h_t(\boldsymbol{\beta})\} - c_\tau) \tag{4.3.5}$$

over $\boldsymbol{\beta}$ and c_τ.

The distribution of $|\varepsilon_t|$ is confined to the nonnegative half-axis and is typically skewed. Intuitively the log-transform will make the distribution less skewed. Peng and Yao (2003) advocated the log-transform and studied the L_1 regression for the ARCH/GARCH models.

Weighted quantile regression. As we discussed in Chapter 2, the WQR is more efficient than traditional QR such as the single QR [Koenker (2005)] and CQR [Zou

and Yuan (2008a)] while inheriting robustness. Applying this weighting scheme to DTARCH models and combining with (4.3.5) motivate us to estimate the model parameters in (4.3.3) by minimizing

$$\sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \rho_{\tau_k}(\log|\varepsilon_t| - \log\{h_t(\boldsymbol{\beta})\} - c_{\tau_k}) \tag{4.3.6}$$

over c_{τ_k} and $\boldsymbol{\beta}$, where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_K)'$ is a vector of weights such that $\|\boldsymbol{\omega}\| = 1$, and c_{τ_k} be the τ_k-th quantile of u_t. Denote by $\hat{\boldsymbol{\beta}}_1$ the resulting solution of $\boldsymbol{\beta}$. For convenience, we refer to it as the WQR estimator.

For $\omega_i = 1/\sqrt{K}$, the above method can be regarded as an extension of the CQR estimation to the DTARCH model. Typically, one can use the equally spaced qauntiles: $\tau_k = k/(K+1)$ for $k = 1, 2, \ldots, K$. The weight ω_k controls the amount of contribution of the τ_k-th quantile regression. Since some weights are allowed to be negative, the WQR method is essentially different from the CQR method.

In order to derive the asymptotic properties of $\hat{\boldsymbol{\beta}}_1$, we introduce the following conditions and assumptions:

(b1) $\boldsymbol{\Gamma} = E(\frac{\mathbf{Z}_t\mathbf{Z}_t'}{h_t^2}) > 0$.

(b2) The innovation u_t has cumulative distribution function $G(\cdot)$ with density $g(\cdot)$ being positive and continuous at $c^*_{\tau_k}$.

Theorem 4.3.2. Suppose that the threshold and the delay parameters are known. Under condition (a0) and conditions $(b1) - (b2)$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}) \to_d \mathcal{N}\Big(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{\Gamma}^{-1}\Big).$$

where

$$\sigma^2(\boldsymbol{\omega}) = \frac{\sum_{k,k'=1}^{K} \omega_k\omega_{k'} \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'}))}{(\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k}))^2}.$$

Theorem 4.3.2 indicates $\hat{\boldsymbol{\beta}}_1$ is asymptotically unbiased, which confirms the log-transformation of $|e_t|$ we made previously is reasonable.

4.3.2 Choice of weights

Since $\boldsymbol{\Gamma}$ does not involve $\boldsymbol{\omega}$, the weights should be selected to minimize $\sigma(\boldsymbol{\omega})$. Let $\mathbf{g} = (g(c_{\tau_1}), \ldots, g(c_{\tau_K}))'$, and $\mathbf{A}$ be a $K \times K$ matrix with the (k, k') element being $A_{kk'} = \min(\tau_k, \tau_{k'})(1 - \max(\tau_k, \tau_{k'}))$. Then the optimal weight $\boldsymbol{\omega}_{opt}$, which minimizes $\sigma(\boldsymbol{\omega})$, can be shown as

$$\boldsymbol{\omega}_{opt} = (\mathbf{g}'\mathbf{A}^{-2}\mathbf{g})^{-1/2}\mathbf{A}^{-1}\mathbf{g}$$

by the maximization lemma (see for example Richard A and Dean W (2007)) under the condition of $\|\boldsymbol{\omega}\| = 1$. The optimal weight components can be very different and some of them may even be negative, which reflects the necessity to use a data-driven weighting scheme. In fact, in our simulations we also experience such a scenario.

4.3.3 Adaptive Estimation

The density function $g(\cdot)$ of e_t can be estimated by running the kernel smoother over residuals. Let the resulting estimate of $\mathbf{g}$ be $\hat{\mathbf{g}}$. Then $\hat{\boldsymbol{\omega}} = (\hat{\mathbf{g}}'\mathbf{A}^{-2}\hat{\mathbf{g}})^{-1/2}\mathbf{A}^{-1}\hat{\mathbf{g}}$ provides a nonparametric estimator of $\boldsymbol{\omega}_{opt}$. This leads to an adaptive estimator of $\boldsymbol{\beta}$ by minimizing:

$$\sum_{k=1}^{K} \hat{\omega}_k \sum_{t=s+1}^{n} \rho_{\tau_k}(\log|\varepsilon_t| - \log(h_t(\boldsymbol{\beta})) - c_{\tau_k}). \tag{4.3.7}$$

over c_{τ_k} and $\boldsymbol{\beta}$. Let the resulting estimator of $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}_2$.

Theorem 4.3.3. *Under the same conditions as in Theorem 4.3.2,*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}^*) \to_d \mathcal{N}\Big(\mathbf{0}, (\mathbf{g}'\mathbf{A}^{-1}\mathbf{g})^{-1}\boldsymbol{\Gamma}^{-1}\Big).$$

Since $\sigma^2(\boldsymbol{\omega}_{opt}) = (\mathbf{g}'\mathbf{A}^{-1}\mathbf{g})^{-1}$, $\hat{\boldsymbol{\beta}}_2$ has the same asymptotic variance matrix as $\hat{\boldsymbol{\beta}}_1$, if $\boldsymbol{\omega}_{opt}$ were known. That is, the estimator $\hat{\boldsymbol{\beta}}_2$ is adaptive.

4.3.4 Estimation of the DTARCH model with AR part

In this section, we introduce two estimation methods. In both scenarios, asymptotic properties of quantile regression estimators will be derived.

Estimates based on residuals.

Residual-based modeling of heteroskedasticity was studied by Engle (1982) and Koenker and Zhao (1996). This approach is carried out in two steps: in the first step they estimated the autoregressive parameters and computed the residuals, and in the second step they estimated the ARCH parameters by regressing the (squared) residuals on the lagged (squared) residuals. In the following we use an analogous procedure to study the asymptotic behavior of weighted QR estimators.

Rewrite the DTARCH model in (4.2.1) and (4.2.2) as

$$y_t = \mathbf{X}'_{t,j}\boldsymbol{\alpha}^{(j)} + h_t(\boldsymbol{\beta})u_t, \quad \text{if} \quad r_{j-1} < y_{t-d} \leq r_j,$$

which is equivalent to

$$y_t = \mathbf{X}_t\boldsymbol{\alpha} + h_t(\boldsymbol{\beta})u_t. \tag{4.3.8}$$

Suppose there exists an estimator, $\hat{\boldsymbol{\alpha}}$, of the AR parameter, $\boldsymbol{\alpha}$, such that $\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = O_p(1)$. Let $\varepsilon_t(\hat{\boldsymbol{\alpha}}) = y_t - \mathbf{X}_t\hat{\boldsymbol{\alpha}}$ be the residuals. Then similar to (4.3.6) we can estimate the ARCH parameters by the solution $\hat{\boldsymbol{\beta}}_3$ of the following minimization problem:

$$\min_{c_k,\boldsymbol{\beta}} \sum_{k=1}^{K} \omega_k \sum_{t=s'+1}^{n} \rho_{\tau_k}(\log|\hat{\varepsilon}_t| - \log\{\hat{h}_t(\boldsymbol{\beta})\} - c_k). \tag{4.3.9}$$

where $\hat{\varepsilon}_t = \varepsilon_t(\hat{\boldsymbol{\alpha}})$, $s' = \max(p_1, \cdots, p_m, q_1, \cdots, q_m)$ and

$$\hat{h}_t(\boldsymbol{\beta}) = \sum_{j=1}^{m} I_{t,j}[\beta_0^{(j)} + \beta_1^{(j)}|\hat{\varepsilon}_{t-1}| + \cdots + \beta_{q_j}^{(j)}|\hat{\varepsilon}_{t-q_j}|].$$

Theorem 4.3.4. *Suppose that the threshold and the delay parameters are known and there exists an estimator, $\boldsymbol{\alpha}_n$, of the AR parameter, $\boldsymbol{\alpha}$, such that $\sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}) = O_p(1)$. Under assumptions (a0) and (b1)-(b2),*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_3 - \boldsymbol{\beta}) \quad = \quad \mathbf{N} - \boldsymbol{\Gamma}^{-1}\mathbf{C}'\sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*) + o_p(1),$$

where $\mathbf{N}$ is a normal random variable with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\omega)\boldsymbol{\Gamma}^{-1}$ and the definition of $\mathbf{C}'$ is delegated to the notations in next theorem.

The result of the theorem reduces to that of Theorem 4.3.2 if the initial estimate $\boldsymbol{\alpha}_n$ is superefficient (i.e. $\sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*) = o_p(1)$) or $\mathbf{C} = \mathbf{0}$ when the innovation density is symmetric about zero. The result also demonstrates that the residual-based WQR estimate of $\boldsymbol{\beta}$ is consistent but generally depends on the initial estimate of $\boldsymbol{\alpha}$. In general, when the innovation is asymmetric, the initial estimate $\boldsymbol{\alpha}_n$ inflates the asymptotic variance of the estimator of $\boldsymbol{\beta}$ in the second step. This is similar to the QR estimate of the ARCH parameters in Koenker and Zhao (1996), which is not a desired property.

Simultaneous estimation of the AR and ARCH parameters.

Simultaneous estimating the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ receives attention in Koenker and Zhao (1996). However, the mathematical property of the joint estimation is challenging and remains unknown. Although the simultaneous estimation is computationally more demanding than the two-step estimation, it avoids the symmetrical assumption on the innovation and does not require a $\sqrt{n}$-consistent initial

estimator of the $\boldsymbol{\alpha}$ parameters. In the following we jointly estimate the model parameters using the WQR.

Let $\varepsilon_t(\boldsymbol{\alpha}) = y_t - \mathbf{X}_t\boldsymbol{\alpha}$. By model (4.3.8),

$$Q_{\tau_k}([\log|\varepsilon_t(\boldsymbol{\alpha})|] \mid \mathcal{F}_{t-1}) = \log(h_t(\boldsymbol{\alpha},\boldsymbol{\beta})) + c_{\tau_k}.$$

Similar to (4.3.6), we propose to estimate the parameters by minimizing

$$\min \sum_{k=1}^{K} \omega_k \sum_{t=s'+1}^{n} \rho_{\tau_k}(\log|\varepsilon_t(\boldsymbol{\alpha})| - \log\{h_t(\boldsymbol{\alpha},\boldsymbol{\beta})\} - c_{\tau_k}). \tag{4.3.10}$$

over c_{τ_k}, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $h_t(\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{j=1}^{m} I_{t,j}(\beta_0^{(j)} + \beta_1^{(j)}|\varepsilon_{t-1}(\boldsymbol{\alpha})| + \cdots + \beta_{q_j}^{(j)}|\varepsilon_{t-q_j}(\boldsymbol{\alpha})|)$. Let the resulting estimators be $(\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\beta}}_4)$.

Let $\boldsymbol{\alpha}^*$ be the true value of $\boldsymbol{\alpha}$. In order to keep the accordance of signs, we abuse some signs to mark $\varepsilon_t = \varepsilon_t(\boldsymbol{\alpha}^*)$, $h_t = h_t(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$, $\mathbf{Z}_t = \mathbf{Z}_t(\boldsymbol{\alpha}^*)$ such that $\varepsilon_t = h_t u_t$, $h_t = \mathbf{Z}_t'\boldsymbol{\beta}^*$. Denote by $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}^{*\prime}, \boldsymbol{\beta}^{*\prime})'$, $\mathbf{D}_t = \frac{\partial \log(h_t(\boldsymbol{\alpha},\boldsymbol{\beta}))}{\partial \boldsymbol{\alpha}}|_{\boldsymbol{\theta}^*}$, $\mathbf{J}_t = \frac{\mathbf{X}_t}{\varepsilon_t} + \mathbf{D}_t$ $\mathbf{C} = \mathrm{cov}(\mathbf{J}_t, \frac{\mathbf{Z}_t}{h_t})$, $\boldsymbol{\mu}_a = E(\mathbf{J}_t)$, $\boldsymbol{\Omega} = E(\mathbf{J}_t\mathbf{J}_t')$, $\boldsymbol{\Pi} = \mathrm{var}(\mathbf{J}_t)$ and $\mathbf{D} = \boldsymbol{\Gamma} - \mathbf{C}'\boldsymbol{\Pi}^{-1}\mathbf{C}$. Then, we have the following asymptotic results for simultaneous estimation of AR and ARCH parameters.

Theorem 4.3.5. *Suppose that the threshold and the delay parameters are known. Under the assumptions* $(a0)$ *and* $(b1)-(b2)$,

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}^* \\ \hat{\boldsymbol{\beta}}_4 - \boldsymbol{\beta}^* \end{pmatrix} = \{\sum_{k=1}^{K} \omega_k g(c_{\tau_k}^*)\}^{-1} \begin{pmatrix} \boldsymbol{\Pi} & \mathbf{C} \\ \mathbf{C}' & \boldsymbol{\Gamma} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \omega_k q_k \\ \mathbf{z} - \boldsymbol{\mu} \sum_{k=1}^{K} \omega_k q_k \end{pmatrix} + o_p(1),$$

where $\mathbf{q} = (q_1, \ldots, q_K)'$, $(\mathbf{q}', \mathbf{r}', \mathbf{z}')'$ being jointly normal with $\mathrm{cov}(\mathbf{q},\mathbf{r}) = \mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}_a'$, $\mathrm{cov}(\mathbf{q},\mathbf{z}) = \mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}'$, $\mathrm{cov}(\mathbf{r},\mathbf{z}) = \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\mathbf{C}$, $\mathbf{q} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0},\mathbf{A})$, $\mathbf{r} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0},\boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\boldsymbol{\Omega})$, $\mathbf{z} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0},\boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\boldsymbol{\Pi})$ and $\mathrm{cov}(\mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \omega_k q_k, \mathbf{z} - \boldsymbol{\mu} \sum_{k=1}^{K} \omega_k q_k) = \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}(\mathbf{C} - \boldsymbol{\mu}_a'\boldsymbol{\mu})$.

Corollary 4.3.1. *Under conditions of Theorem 4.3.5, we have*

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}^*) &= -\{\sum_{k=1}^{K}\omega_k g(c^*_{\tau_k})\}^{-1}[(\boldsymbol{\Pi}^{-1} + \boldsymbol{\Pi}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}'\boldsymbol{\Pi}^{-1}) \\ &\times \ (\mathbf{r} - \boldsymbol{\mu}_a\sum_{k=1}^{K}\eta_k) - \boldsymbol{\Pi}^{-1}\mathbf{C}\mathbf{D}^{-1}(\mathbf{z} - \boldsymbol{\mu}\sum_{k=1}^{K}\eta_k)] + o_p(1),\end{aligned}$$

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_4 - \boldsymbol{\beta}^*) &= -\{\sum_{k=1}^{K}\omega_k g(c^*_{\tau_k})\}^{-1}[\mathbf{D}^{-1}(\mathbf{z} - \boldsymbol{\mu}\sum_{k=1}^{K}\eta_k) \\ &- \ \mathbf{D}^{-1}\mathbf{C}'\boldsymbol{\Pi}^{-1}(\mathbf{r} - \boldsymbol{\mu}_a\sum_{k=1}^{K}\eta_k)] + o_p(1),\end{aligned}$$

where $(\boldsymbol{\omega}'\mathbf{g})^{-1}(\mathbf{r} - \boldsymbol{\mu}_a\sum_{k=1}^{K}\eta_k) \sim \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{\Pi})$ *and* $(\boldsymbol{\omega}'\mathbf{g})^{-1}(\mathbf{z} - \boldsymbol{\mu}\sum_{k=1}^{K}\eta_k) \sim \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{\Gamma})$.

Remark 4.3.1. *If* u_t *is symmetric and* $\boldsymbol{\alpha}_0 = (\alpha_0^1, \ldots, \alpha_0^m) = \mathbf{0}$, *then* $\mathbf{C} = 0$.

Remark 4.3.2. *If there are no AR part in DTARCH models, then* $\mathbf{C} = \mathbf{0}$. *In this case, Theorem 4.3.5 reduces to Theorem 4.3.2*

4.4 Computational Issue

Minimization problems (4.3.6)-(4.3.7) and (4.3.9) are special cases of (4.3.10). We use the following algorithm to solve (4.3.10):

(a) Given a consistent estimate of $\boldsymbol{\alpha}$, say $\boldsymbol{\alpha}_0$, use the interior point algorithm to find the solution $(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_4)$ to

$$\arg\min_{\alpha,\beta,c_{\tau_k}}\sum_{k=1}^{K}\omega_k\sum_{t=s'+1}^{n}\rho_{\tau_k}(\log|\varepsilon_t(\boldsymbol{\alpha}_0)| - \log\{h_t(\boldsymbol{\alpha},\boldsymbol{\beta})\} - c_{\tau_k}).$$

The optimization problem can be solved by the updated interior algorithm developed in §2.3 of Chapter2.

(b) Use $\boldsymbol{\alpha}_1$ to replace the above $\boldsymbol{\alpha}_0$, and then solve the above optimization problem to get an updated estimate of $\boldsymbol{\alpha}$. Repeat this procedure until convergence.

4.4.1 Simulation results

In this section we conduct 1000 simulations to investigate the advantages of the WQR estimation. The following DTARCH model is studied:

$$y_t = \begin{cases} \alpha_1^{(1)} y_{t-1} + \varepsilon_t & \text{if } y_{t-1} \le 0 \\ \alpha_1^{(2)} y_{t-1} + \varepsilon_t & \text{if } y_{t-1} > 0, \end{cases}$$

where $(\alpha_1^{(1)}, \alpha_1^{(2)}) = (0.20, 0.35)$ and $\varepsilon_t = h_t u_t$, with

$$h_t = \begin{cases} 1 + \beta_1^{(1)} |\varepsilon_{t-1}| & \text{if } y_{t-1} \le 0 \\ \beta_0^{(2)} + \beta_1^{(2)} |\varepsilon_{t-1}| & \text{if } y_{t-1} > 0, \end{cases}$$

where $\beta_1^{(1)} = 0.3$ and $(\beta_0^{(2)}, \beta_1^{(2)}) = (1.00, 0.25)$. We employ three sets of innovation variables: $N(0,1)$, $t(3)$, and $\chi^2(3)$, which are centralized and normalized so that the medians of the absolute innovations are 1's. In each simulation, we draw a sample of size $n = 2400$. For the CQR and WQR methods, we take equal spaced quantiles at $\tau_k = 0.25, 0.5$, and 0.75.

We compare the proposed robust estimation approach with that based on the MLE when the innovation is normal. For non-normal innovations, we compare the robust approach with the quasi-likelihood based method (QMLE). In each simulation the bias and the "root of mean squared error (RMS)" for different estimators are calculated, and their average over simulations are reported in Tables 1-3, which lead to the following two points:

(i) When the innovation is normal, the MLE dominates the others for estimating the parameters of AR part, but the WQR and CQR estimates for the ARCH part compare favorably to the MLE. However, the QMLE is quite unsatisfactory in the two nonnormal innovation cases since the averages of $\boldsymbol{\beta}$ estimates are far away from the true values even though the parameters $\boldsymbol{\alpha}$ are well

Table 4.1: Comparison of different estimators of parameters under the scaled normal innovation. $\omega_{opt} = (0.1031, 0.2622, 0.9595)'$.

Esti	Measures	$\alpha_1^{(1)}$	$\alpha_1^{(2)}$	$\beta_1^{(1)}$	$\beta_0^{(2)}$	$\beta_1^{(2)}$
QMLE	Bias	-0.0019	-0.0017	-0.0018	0.0020	-0.0012
	RMS	0.0385	0.0304	0.0199	0.0389	0.0258
L_1	Bias	0.0022	-0.0053	0.0076	0.0025	0.0033
	RMS	0.0813	0.0728	0.0530	0.0765	0.0367
CQR	Bias	-0.0040	-0.0064	0.0047	0.0041	0.0012
	RMS	0.0444	0.0385	0.0236	0.0359	0.0182
CWQR	Bias	0.0002	-0.0084	0.0047	0.0012	-0.0004
	RMS	0.0425	0.0341	0.0215	0.0323	0.0163

*u_t is normalized to satisfy that $E(u_t) = 0$ and $Median(|u_t| = 1)$.

estimated. This result queries the use of QMLE for DTARCH models, which is different for ARCH models (see Jiang et al. 2001).

(ii) The WQR estimation with data-driven weights uniformly dominates the L_1 estimation and the CQR estimation. This endorses the value of our WQR method.

4.5 Conclusion

In this chapter, we have studied the WQR estimation for DTARCH models. Simultaneous WQR estimation of AR and ARCH parameters has been proposed, and the asymptotic properties of WQR estimators have been established. Theoretical and computational results all support our finding that the data-driven WQR estimation uniformly dominates the L_1 estimation and the CQR estimation.

Table 4.2: Comparison of different estimators of parameters under the scaled $t(3)$ innovation. $\omega_{opt} = (0.2032, 0.4409, 0.8742)'$.

Esti	Measures	$\alpha_1^{(1)}$	$\alpha_1^{(2)}$	$\beta_1^{(1)}$	$\beta_0^{(2)}$	$\beta_1^{(2)}$
QMLE	Bias	-0.0000	-0.0024	0.6080	0.6943	0.1758
	RMS	0.0646	0.0384	0.7633	0.7405	0.2386
L_1	Bias	-0.0041	-0.0054	0.0076	0.0060	0.0024
	RMS	0.0683	0.0589	0.0505	0.0845	0.0367
CQR	Bias	-0.0055	-0.0093	-0.0001	-0.0104	0.0009
	RMS	0.0614	0.0564	0.0382	0.0651	0.0310
CWQR	Bias	0.0033	-0.0060	0.0014	-0.0085	0.0019
	RMS	0.0448	0.0364	0.0247	0.0488	0.0195

*u_t is normalized to satisfy that $E(u_t) = 0$ and $Median(|u_t| = 1)$.

Table 4.3: Comparison of different estimators of parameters under the scaled $\chi^2(3)$ innovation. $\omega_{opt} = (-0.0053, 0.0785, 0.9969)'$.

Esti	Measures	$\alpha_1^{(1)}$	$\alpha_1^{(2)}$	$\beta_1^{(1)}$	$\beta_0^{(2)}$	$\beta_1^{(2)}$
QMLE	Bias	0.0015	-0.0034	0.7977	0.5504	0.0757
	RMS	0.0467	0.0326	0.8018	0.5636	0.0941
L_1	Bias	0.00186	-0.0023	0.00059	0.0059	0.0039
	RMS	0.1157	0.0503	0.0617	0.0803	0.0377
CQR	Bias	0.0197	-0.0043	0.0146	0.0053	-0.0003
	RMS	0.0987	0.0486	0.0422	0.0619	0.0301
CWQR	Bias	0.0183	-0.0081	0.0085	0.0037	-0.0075
	RMS	0.0800	0.0407	0.0296	0.0509	0.0238

*u_t is normalized to satisfy that $E(u_t) = 0$ and $Median(|u_t| = 1)$.

Chapter 5

Conclusions and Further Developments

5.1 Conclusions

The proposed weighted QR has been demonstrated as a powerful tool for modeling nonlinear models. Its advantages have been advocated for the nonlinear regression models and the DTARCH models. Our results provide new insights into quantile regression. The values of the WQR and regularized WQR are revealed in statistical modeling.

5.2 Further Developments

The proposed WQR is applicable to other models and can be extended to the nonparametric smoothing world. Further topics include but not limited to

(i) Extensions of the proposed WQR to other models such as the transformation models, nonparametric/semiparametric regression models and time-varying or functional coefficient models. Due to the nonparametric nature, we need to develop local weighted QR modeling methods.

(ii) Regularized estimation of the DTARCH models. This is an important problem. Since the GARCH(1,1) model with $ARCH(\infty)$ representation has suc-

cessfully applied in modeling real financial data, if one use a DTARCH model to fit such data, the order of model may be very high. It is natural to consider estimation of the DTARCH model with a diverging number of parameters and to develop the regularized WQR modeling strategy for model selection.

(iii) Hypothesis testing based on the WQR fitting. The generalized likelihood ratio tests in Fan et al. (2001) and Fan and Jiang (2005, 2007) can be accommodated to the WQR but with more technical challenges.

(iv) Extensions of the WQR to multivariate cases. This is challenging since there is no unique definition for multivariate quantiles.

□ **End of chapter.**

Appendix A

Proofs of Theorems for Chapter 2

A.1 Proofs of Theorems

In this appendix we give proofs of our theorems. Let

$$S_n(\mathbf{u},\mathbf{v}) = L_n(\boldsymbol{\beta}^* + n^{-1/2}\mathbf{u}, \mathbf{b}^* + n^{-1/2}\mathbf{v}) - L_n(\boldsymbol{\beta}^*, \mathbf{b}^*).$$

$$\eta_{n,k} = n^{-1/2}\omega_k \sum_{i=1}^{n} [I(\varepsilon_i < b^*_{\tau_k}) - \tau_k],$$

$$\mathbf{z}_n = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nabla} f_i^* \sum_{k=1}^{K} \omega_k [I(\varepsilon_i < b^*_{\tau_k}) - \tau_k],$$

Proof of Theorem 2.2.1. Let $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$, $\sqrt{n}(b_{\tau_k} - b^*_{\tau_k}) = v_k$, $\mathbf{v} = (v_1, \ldots, v_K)'$, and $\xi_i(\mathbf{u}, v_k) = (\boldsymbol{\nabla} f_i^*)'\mathbf{u} + v_k$. Then minimizing the $L_n(\boldsymbol{\beta}, \mathbf{b})$ in (2.2.4) is equivalent to minimizing $S_n(\mathbf{u}, \mathbf{v})$. Put

$$\begin{aligned} S_n^* &= \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big\{ \rho_{\tau_k}(y_i - f_i^* - (\boldsymbol{\nabla} f_i^*)' n^{-1/2}\mathbf{u} - (b^*_{\tau_k} + n^{-1/2} v_k)) \\ &\quad - \rho_{\tau_k}(y_i - f_i^* - b^*_{\tau_k}) \Big\}, \end{aligned}$$

$$\begin{aligned} S_n^{**} &= \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big\{ \rho_{\tau_k}\big(y_i - f_i^* - (\boldsymbol{\nabla} f_i^*)' n^{-1/2}\mathbf{u} \\ &\quad - \frac{1}{2n}\mathbf{u}'(\boldsymbol{\nabla}^2 f_i^*)\mathbf{u} - (b^*_{\tau_k} + n^{-1/2} v_k)\big) - \rho_{\tau_k}(y_i - f_i^* - b^*_{\tau_k}) \Big\}. \end{aligned}$$

Then

$$S_n^* = \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big\{ \rho_{\tau_k}\big(\varepsilon_i - b^*_{\tau_k} - \frac{1}{\sqrt{n}}\xi_i(\mathbf{u}, v_k)\big) - \rho_{\tau_k}(\varepsilon_i - b^*_{\tau_k}) \Big\}, \tag{A.1.1}$$

$$\begin{aligned} S_n^{**} \;=\; & \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big\{ \rho_{\tau_k}\big(\varepsilon_i - b^*_{\tau_k} - \frac{1}{\sqrt{n}}\xi_i(\mathbf{u}, v_k) - \frac{1}{2n}\mathbf{u}'(\boldsymbol{\nabla}^2 f_i^*)\mathbf{u}\big) \\ & -\rho_{\tau_k}(\varepsilon_i - b^*_{\tau_k}) \Big\}. \end{aligned}$$

Denote by $\boldsymbol{\mu} = E(\boldsymbol{\nabla} f_1^*)$ and $\boldsymbol{\Gamma} = E[(\boldsymbol{\nabla} f_1^*)^{\otimes 2}]$. Then $\mathbf{G} = \boldsymbol{\Gamma} - \boldsymbol{\mu}^{\otimes 2}$. We will show that

$$S_n(\mathbf{u}, \mathbf{v}) \xrightarrow{d} S(\mathbf{u}, \mathbf{v}), \tag{A.1.2}$$

where

$$S(\mathbf{u}, \mathbf{v}) = \sum_{k=1}^{K} \eta_k v_k + \mathbf{z}'\mathbf{u} + \frac{1}{2}\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'\boldsymbol{\Gamma}\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u})$$

with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)'$ and $(\boldsymbol{\eta}', \mathbf{z}')'$ being jointly normal with mean zero and covariance matrix $Cov(\boldsymbol{\eta}, \mathbf{z}) = \mathrm{diag}(\boldsymbol{\omega})\mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}'$.

Statement (i): Asymptotic expression for S_n^*:

$$S_n^* = \sum_{k=1}^{K} \eta_{n,k} v_k + \mathbf{z}_n'\mathbf{u} + \frac{1}{2}\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'\boldsymbol{\Gamma}\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u}) + o_p(1). \tag{A.1.3}$$

In fact, by the identity [Knight (1998)]

$$|r - s| - |r| = -s(I(r > 0) - I(r < 0) + 2\int_0^s [I(r \le x) - I(r \le 0)]dx,$$

we have

$$\rho_\tau(r - s) - \rho_\tau(r) = s[I(r < 0) - \tau] + \int_0^s [I(r \le x) - I(r \le 0)]dx. \tag{A.1.4}$$

Combining (A.1.4) and (A.1.1), we obtain that

$$\begin{aligned} S_n^* &= \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \xi_i(\mathbf{u}, v_k)[I(\varepsilon_i < b^*_{\tau_k}) - \tau_k] \\ &+ \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} [I(\varepsilon_i \le b^*_{\tau_k} + x) - I(\varepsilon_i \le b^*_{\tau_k})]dx \\ &= \{\sum_{k=1}^{K} \eta_{n,k} v_k\} + \mathbf{z}_n' \mathbf{u} + \sum_{k=1}^{K} \omega_k A_n^{(k)}, \end{aligned} \tag{A.1.5}$$

where

$$A_n^{(k)} = \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} [I(\varepsilon_i \le b^*_{\tau_k} + x) - I(\varepsilon_i \le b^*_{\tau_k})]dx.$$

Decompose $A_n^{(k)}$ into $A_n^{(k)} = A_{n1}^{(k)} + A_{n2}^{(k)}$ with

$$A_{n1}^{(k)} = \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} [G(b^*_{\tau_k} + x) - G(b^*_{\tau_k})]\, dx$$

and

$$A_{n2}^{(k)} = \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} \Big\{ I(\varepsilon_i \le b^*_{\tau_k} + x) - I(\varepsilon_i \le b^*_{\tau_k}) - [G(b^*_{\tau_k} + x) - G(b^*_{\tau_k})] \Big\}\, dx.$$

By the mean value theorem,

$$\begin{aligned} A_{n1}^{(k)} &= \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} x g(b^{**}_{\tau_k})\, dx \\ &= \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} x g(b^*_{\tau_k})\, dx + \sum_{i=1}^{n} \int_0^{\frac{1}{\sqrt{n}}\xi_i(\mathbf{u},v_k)} x[g(b^{**}_{\tau_k}) - g(b^*_{\tau_k})]\, dx \\ &\equiv A_{n11}^{(k)} + A_{n12}^{(k)}, \end{aligned}$$

where $b^{**}_{\tau_k}$ is between $b^*_{\tau_k}$ and $b^*_{\tau_k} + x$. By simple algebra, the central limit theorem and the continuity of $g(\cdot)$, we have

$$A_{n11}^{(k)} = \frac{1}{2} g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'\mathbf{\Gamma}\mathbf{u} + 2v_k \boldsymbol{\mu}'\mathbf{u}) + o_p(1)$$

and $A_{n12}^{(k)} = o_p(1)$. Note that $E[A_{n2}^{(k)}] = 0$ and $\text{var}[A_{n2}^{(k)}] = o(1)$. It follows that $A_{n2}^{(k)} = o_p(1)$. Therefore,

$$A_n^{(k)} = \frac{1}{2} g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}'\boldsymbol{\Gamma}\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u}) + o_p(1).$$

It follows from (A.1.5) that (A.1.3) holds.

Statement (ii): $S_n^{**} - S_n^* \longrightarrow_p 0$. This can be obtained using the same argument as in Davis (1997).

Statement (iii): $S_n(\mathbf{u}, \mathbf{v}) - S_n^{**} = o_p(1)$. Using the inequality $|\rho_\tau(r_1) - \rho_\tau(r_2)|/|r_1 - r_2| \leq \max(\tau, 1-\tau) < 1$, we obtain that

$$\begin{aligned}
|S_n(\mathbf{u}, \mathbf{v}) - S_n^{**}| &= \Big| \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big[\rho_{\tau_k}\big(y_i - f_i^* - n^{-1/2}(\boldsymbol{\nabla} f_i^*)'\mathbf{u} \\
&\quad - \frac{1}{2n}\mathbf{u}'\boldsymbol{\nabla}^2 f(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})\mathbf{u} - (b_{\tau_k}^* + n^{-1/2}v_k)\big)\Big] \\
&\quad - \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \Big[\rho_{\tau_k}\big(y_i - f_i^* - n^{-1/2}(\boldsymbol{\nabla} f_i^*)'\mathbf{u} \\
&\quad - \frac{1}{2n}\mathbf{u}'\boldsymbol{\nabla}^2 f_i^*\mathbf{u} - (b_{\tau_k}^* + n^{-1/2}v_k)\big)\Big] \Big| \\
&\leq \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \frac{1}{2n} \Big| \mathbf{u}'[\boldsymbol{\nabla}^2 f_i^* - \boldsymbol{\nabla}^2 f(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})]\mathbf{u} \Big|,
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + n^{-1/2}\mathbf{u}$. Then by Condition (c), $S_n(\mathbf{u}, \mathbf{v}) - S_n^{**} \rightarrow_p 0$.

Combining Statements (i)-(iii) leads to

$$\begin{aligned}
S_n(\mathbf{u}, \mathbf{v}) &= \sum_{k=1}^{K} \eta_{n,k} v_k + \mathbf{z}_n'\mathbf{u} \\
&\quad + \frac{1}{2} \sum_{k=1}^{K} \omega_k g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}'\boldsymbol{\Gamma}\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u}) + o_p(1). \qquad \text{(A.1.6)}
\end{aligned}$$

Let $\boldsymbol{\eta}_n = (\eta_{n,1}, \ldots, \eta_{n,K})'$. Using the Cramér-Wold device and the multivariate central limit theorem, we establish that $(\boldsymbol{\eta}_n', \mathbf{z}_n')' \xrightarrow{\mathcal{D}} (\boldsymbol{\eta}', \mathbf{z}')'$, and hence

$$\sum_{k=1}^{K} \eta_{n,k} v_k + \mathbf{z}_n'\mathbf{u} \xrightarrow{\mathcal{D}} \sum_{k=1}^{K} \eta_k v_k + \mathbf{z}'\mathbf{u}, \qquad \text{(A.1.7)}$$

where $\eta_{n,k} \xrightarrow{\mathcal{D}} \eta_k$ and $\mathbf{z}_n \xrightarrow{\mathcal{D}} \mathbf{z}$. Therefore, by Slutsky's theorem and (A.1.6), (A.1.2) is true.

Let $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{u}}$ be the minimizers of $S_n(\mathbf{u}, \mathbf{v})$ and $S(\mathbf{u}, \mathbf{v})$ for $\mathbf{u}$, respectively. Since $S(\mathbf{u}, \mathbf{v})$ is a quadratic form of $(\mathbf{u}', \mathbf{v}')'$, $\hat{\mathbf{u}}$ is unique. Simple algebra gives that

$$\hat{\mathbf{u}} = -[\sum_{k=1}^{K} \omega_k f(b^*_{\tau_k})\mathbf{G}]^{-1}\{\mathbf{z} - \boldsymbol{\mu}(\sum_{k=1}^{K} \eta_k)\}.$$

Since the minimization operator is continuous under the infimum topology, by (A.1.2) and the continuous mapping theorem [see for example Theorem 25.7 in Billingsley (1995)],

$$\hat{\mathbf{u}}_n \xrightarrow{\mathcal{D}} \hat{\mathbf{u}}. \tag{A.1.8}$$

Since $[\sum_{k=1}^{K} \omega_k f(b^*_{\tau_k})]^{-1}\{\mathbf{z} - \boldsymbol{\mu}(\sum_{k=1}^{K} \eta_k)\}$ is normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\boldsymbol{\omega})\mathbf{G}$ and $\hat{\mathbf{u}}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*)$,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*) \rightarrow_d N\Big(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{G}^{-1}\Big).$$

Proof of Theorem 2.2.2. Note that $\hat{\boldsymbol{\omega}} = \boldsymbol{\omega}_{opt}(1 + o_p(1))$ and $\sigma(\hat{\boldsymbol{\omega}}) = \sigma(\boldsymbol{\omega}_{opt})(1 + o_p(1)) = (\mathbf{g}'\mathbf{A}^{-1}\mathbf{g})^{-1}(1 + o_p(1))$. The result can be proven by using the same argument as in Theorem 2.2.1.

Proof of Theorem 2.3.1. Let $\sqrt{n}(\beta_j - \beta^*_j) = u_j$, $\sqrt{n}(b_{\tau_k} - b^*_{\tau_k}) = v_k$, $\mathbf{u} = (u_1, \ldots, u_p)'$, and $\mathbf{v} = (v_1, \ldots, v_K)'$. To prove consistency of $\hat{\boldsymbol{\beta}}$, it suffices to show that for any $\delta > 0$, there exists a large constant C such that

$$P\{\inf_{(\mathbf{u},\mathbf{v})\in\mathcal{C}} Q^{SC}(\boldsymbol{\beta}^* + n^{-1/2}\mathbf{u}, \mathbf{b}^* + n^{-1/2}\mathbf{v}) > Q^{SC}(\boldsymbol{\beta}^*, \mathbf{b}^*)\} \geq 1 - \delta, \tag{A.1.9}$$

where $\mathcal{C} = \{(\mathbf{u}, \mathbf{v}) : \|\mathbf{u}\| = \|\mathbf{v}\| = C\}$. This implies that with probability tending

to one there exists a local minimum $\hat{\boldsymbol{\beta}}$ in the ball $\{(\boldsymbol{\beta}^* + n^{-\frac{1}{2}}\mathbf{u}, \mathbf{b}^* + n^{-\frac{1}{2}}\mathbf{v}) : \|\mathbf{u}\| \leq C, \|\mathbf{v}\| \leq C\}$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| = O_p(n^{-\frac{1}{2}})$.

Define $D^{SC}(\mathbf{u}, \mathbf{v}) = Q^{SC}(\boldsymbol{\beta}^* + n^{-\frac{1}{2}}\mathbf{u}, \mathbf{b}^* + n^{-\frac{1}{2}}\mathbf{v}) - Q^{SC}(\boldsymbol{\beta}^*, \mathbf{b}^*)$. Then

$$D^{SC}(\mathbf{u}, \mathbf{v}) = S_n(\mathbf{u}, \mathbf{v}) + n\sum_{j=1}^{p}\{p_{\lambda_n}(|\beta_j^* + n^{-\frac{1}{2}}u_j|) - p_{\lambda_n}(|\beta_j^*|)\}. \tag{A.1.10}$$

Using $p_{\lambda_n}(0) = 0$, we get

$$D^{SC}(\mathbf{u}, \mathbf{v}) \geq S_n(\mathbf{u}, \mathbf{v}) + n\sum_{j=1}^{s}\big\{p_{\lambda_n}(|\beta_j^* + n^{-\frac{1}{2}}u_j|) - p_{\lambda_n}(|\beta_j^*|)\big\}. \tag{A.1.11}$$

Note that, for large n

$$n\sum_{j=1}^{s}\big\{p_{\lambda_n}(|\beta_j^* + n^{-\frac{1}{2}}u_j|) - p_{\lambda_n}(|\beta_j^*|)\big\} = 0, \tag{A.1.12}$$

uniformly in any compact set of $\mathbf{R}^s$ due to the facts that $|\beta_j^*| > 0$ (for $j = 1, 2, \ldots, s$) and the SCAD penalty $p_{\lambda_n}(|\beta_j^*|)$ is flat for coefficients of magnitude larger than $a\lambda_n$, and $\lambda_n \to 0$. It follows from (A.1.6) and (A.1.7) that the right hand side of (A.1.11) is dominated by the positive quadratic term

$$\frac{1}{2}\sum_{k=1}^{K}\omega_k g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}'\boldsymbol{\Gamma}\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u})$$

as long as $\|\mathbf{u}\|$ and $\|\mathbf{v}\|$ are allowed to be large enough. This means (C.2.18) holds. The proof is completed.

Lemma A.1.1. *Suppose Conditions* $(a) - (d)$ *hold. If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then with probability tending to* 1, *for any given* $(\boldsymbol{\beta}_1, \mathbf{b})$ *satisfying* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\| = O_p(n^{-\frac{1}{2}})$ *and* $\|\mathbf{b} - \mathbf{b}^*\| = O_p(n^{-\frac{1}{2}})$ *and for any positive constant* C,

$$Q^{SC}((\boldsymbol{\beta}_1', \mathbf{0}')', \mathbf{b}) = \min_{\|\boldsymbol{\beta}_2\| \leq Cn^{-\frac{1}{2}}} Q^{SC}((\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \mathbf{b}).$$

Proof. Let $\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) = \mathbf{u}_1$, $\sqrt{n}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2^*) = \mathbf{u}_2$, $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$, and $\sqrt{n}(\mathbf{b} - \mathbf{b}^*) = \mathbf{v}$. By the definitions of $Q^{SC}(\boldsymbol{\beta}, \mathbf{b})$ and $S_n(\mathbf{u}, \mathbf{v})$, we have

$$\begin{aligned} & Q^{SC}((\boldsymbol{\beta}_1', \mathbf{0}')', \mathbf{b}) - Q^{SC}((\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \mathbf{b}) \\ = \; & S_n((\mathbf{u}_1', \mathbf{0}'))', \mathbf{v}) - S_n((\mathbf{u}_1', \mathbf{u}_2')', \mathbf{v}) - n \sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|). \end{aligned}$$

From (A.1.2), we obtain that $S_n((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) = O_p(1)$ and $S_n((\mathbf{u}_1', \mathbf{u}_2')', \mathbf{v}) = O_p(1)$. By $p_{\lambda_n}(0) = 0$ and the mean value theorem, there exists $\beta_j^\dagger$ between 0 and $|\beta_j|$ such that for large n,

$$\begin{aligned} n \sum_{j=s+1}^{p} p_{\lambda_n}(|\beta_j|) &= \sqrt{n}\lambda_n \sum_{j=s+1}^{p} \frac{p'_{\lambda_n}(|\beta_j^\dagger|)}{\lambda_n} |\sqrt{n}\beta_j^\dagger| \\ &\geq \sqrt{n}\lambda_n (\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n) \sum_{j=s+1}^{p} |\sqrt{n}\beta_j^\dagger|. \end{aligned}$$

If $\boldsymbol{\beta}_2 \neq 0$, then for $\|\boldsymbol{\beta}_2\| \leq Cn^{-\frac{1}{2}}$, we have $0 < \sum_{j=s+1}^{p} |\sqrt{n}\beta_j^\dagger| \leq \sum_{j=s+1}^{p} |\sqrt{n}\beta_j| < \sqrt{p}\sqrt{n}\|\boldsymbol{\beta}_2\| \leq Cp$. Then, by Condition (d) and $\sqrt{n}\lambda_n \to \infty$, for larger n, $Q^{SC}((\boldsymbol{\beta}_1', \mathbf{0}')', \mathbf{b}) - Q^{SC}((\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')', \mathbf{b})$ is dominated by the term $-n \sum_{s+1}^{p} p_{\lambda_n}(|\beta_j|)$, which is less than zero. Hence, the lemma holds.

Proof of Theorem 2.3.2.

(i) (Sparsity) It follows Lemma C.1.1.

(ii) (Asymptotic normality) Let $\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) = \mathbf{u}_1$, $\sqrt{n}(b_{\tau_k} - b_{\tau_k}^*) = v_k$ and $\mathbf{v} = (v_1, \ldots, v_K)'$. By (A.1.10), we have

$$D^{SC}((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) = S_n((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) + n \sum_{j=1}^{s} \{p_{\lambda_n}(|\beta_j^* + n^{-1/2}u_j|) - p_{\lambda_n}(|\beta_j^*|)\}.$$

Note that $\hat{\mathbf{u}}_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ minimizes

$$\begin{aligned} D^{SC}((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) &= Q^{SC}((\boldsymbol{\beta}_1^{*T}, \mathbf{0})' + n^{-\frac{1}{2}}(\mathbf{u}_1', \mathbf{0}')', \mathbf{b}^* + n^{-\frac{1}{2}}\mathbf{v}) \\ & \quad - Q^{SC}((\boldsymbol{\beta}_1^{*T}, \mathbf{0})', \mathbf{b}^*). \end{aligned}$$

It follows from (A.1.2) and (A.1.12) that

$$D^{SC}((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) \xrightarrow{\mathcal{D}} S((\mathbf{u}_1', \mathbf{0}')', \mathbf{v})$$

with

$$S((\mathbf{u}_1', \mathbf{0}')', \mathbf{v}) = \sum_{k=1}^{K} \eta_k v_k + \mathbf{z}_1' \mathbf{u}_1 + \frac{1}{2} \sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}_1' \mathbf{\Gamma}_{11} \mathbf{u}_1 + 2 v_k \boldsymbol{\mu}_1' \mathbf{u}_1),$$

where $\mathbf{z}_1$ and $\boldsymbol{\mu}_1$ are the sub-vectors consisting of the first s components of $\mathbf{z}$ and $\boldsymbol{\mu}$, respectively, and $\mathbf{\Gamma}_{11}$ is the sub-matrix of $\mathbf{\Gamma}$ with both row and column indices in $\mathcal{A}$. Therefore, by the same argument as for (A.1.8), the minimizer $\hat{\mathbf{u}}_{n1} = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)$ of $D^{SC}((\mathbf{u}_1', \mathbf{0}')', \mathbf{v})$ converges in distribution to the minimizer $\hat{\mathbf{u}}_1$ of $S((\mathbf{u}_1', \mathbf{0}')', \mathbf{v})$. Tote that

$$\hat{\mathbf{u}}_1 = -[\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})]^{-1} \mathbf{G}_{11}^{-1} \{\mathbf{z}_1 - \boldsymbol{\mu}_1 (\sum_{k=1}^{K} \eta_k),$$

where $[\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})]^{-1} \{\mathbf{z}_1 - \boldsymbol{\mu}_1 (\sum_{k=1}^{K} \eta_k)\}$ is normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\boldsymbol{\omega}) \mathbf{G}_{11}$. It follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \xrightarrow{\mathcal{D}} \mathcal{N}\big(\mathbf{0}, \sigma^2(\boldsymbol{\omega}) \mathbf{G}_{11}^{-1}\big).$$

Proof of Theorem 2.3.3.

(i) follows the same argument as in Zou and Yuan (2008a).

(ii) Let $\sqrt{n}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) = \mathbf{u}_1$, $\sqrt{n}(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_2^*) = \mathbf{u}_2$, $\mathbf{u} = (\mathbf{u}_1', \mathbf{u}_2')' = (u_1, \ldots, u_p)'$, $\sqrt{n}(b_{\tau_k} - b^*_{\tau_k}) = v_k$, and $\mathbf{v} = (v_1, v_2, \ldots, v_k)'$. Define $D^{AL}(\mathbf{u}, \mathbf{v}) = Q^{AL}(\boldsymbol{\beta}^* + n^{-\frac{1}{2}}\mathbf{u}, \mathbf{b}^* + n^{-\frac{1}{2}}\mathbf{v}) - Q^{AL}(\boldsymbol{\beta}^*, \mathbf{b}^*)$. Then minimizing $Q^{AL}(\boldsymbol{\beta}, \mathbf{b})$ in (2.3.7) is equivalent to minimizing

$$D^{AL}(\mathbf{u}, \mathbf{v}) = S_n(\mathbf{u}, \mathbf{v}) + \sum_{j=1}^{p} \frac{n h_n}{\sqrt{n}|\tilde{\beta}_j|^{\gamma}} \sqrt{n}(|\beta_j^* + n^{-1/2} u_j| - |\beta_j^*|). \tag{A.1.13}$$

If $u_j \neq 0$, then $\sqrt{n}(|\beta_j^* + n^{-1/2} u_j| - |\beta_j^*|) \to u_j sgn(\beta_j^*)$. If $\beta_j^* \neq 0$, then by the sparsity, $\tilde{\beta}_j \neq 0$ holds with probability tending to one, and hence $\frac{n h_n}{\sqrt{n}|\beta_j|^{\gamma}} \to 0$ by

the assumption $n^{1/2}h_n \to 0$. If $\beta_j^* = 0$, then by $\sqrt{n}\tilde{\beta}_j = O_p(1)$ and the assumption $h_n n^{(\gamma+1)/2} \to \infty$, we have $\frac{nh_n}{\sqrt{n}|\tilde{\beta}_j|^\gamma} = \frac{h_n n^{(\gamma+1)/2}}{|\sqrt{n}\tilde{\beta}_j|^\gamma} \to_p \infty$. Therefore, by Slutsky's theorem,

$$\begin{aligned} & \sum_{j=1}^{p} \frac{nh_n}{\sqrt{n}|\tilde{\beta}_j|^\gamma} \sqrt{n}(|\beta_j^* + n^{-1/2}u_j|) - |\beta_j^*|) \\ \to_p & \sum_{j=1}^{p} W(\beta_j^*, u_j) = \begin{cases} \infty, & \text{if } \beta_j^* = 0 \text{ and } u_j \neq 0; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

It follows from (A.1.2) and (A.1.13) that

$$D^{AL}(\mathbf{u}, \mathbf{v}) \xrightarrow{\mathcal{D}} M(\mathbf{u}, \mathbf{v}) \equiv S(\mathbf{u}, \mathbf{v}) + \sum_{j=1}^{p} W(\beta_j^*, u_j).$$

Then similar to (A.1.8), the minimizer $\hat{\mathbf{u}}_n = (\hat{\mathbf{u}}_{n1}', \hat{\mathbf{u}}_{n2}')'$ of $D^{AL}(\mathbf{u}, \mathbf{v})$ converges in distribution to the minimizer $\hat{\mathbf{u}} = (\hat{\mathbf{u}}_1', \hat{\mathbf{u}}_2')$ of $M(\mathbf{u}, \mathbf{v})$. Note that $\hat{\mathbf{u}}_2 \xrightarrow{\mathcal{D}} \mathbf{0}$, $\hat{\mathbf{u}}_1 \xrightarrow{\mathcal{D}} N\Big(0, \sigma^2(\boldsymbol{\omega})\mathbf{G}_{11}^{-1}\Big)$ and $\hat{\mathbf{u}}_{n1} = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^{AL} - \boldsymbol{\beta}_1^*)$. It follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{AL} - \boldsymbol{\beta}_1^*) \to_d N\Big(0, \sigma^2(\boldsymbol{\omega})\mathbf{G}_{11}^{-1}\Big).$$

This completes the proof of the theorem.

Appendix B

Proofs of Theorems for Chapter 3

B.1 Proofs of Theorems

In this appendix we give proofs of our theorems. Let

$$S_n(\mathbf{u},\mathbf{v}) = L_n(\boldsymbol{\beta}^* + n^{-1/2}\mathbf{u}, \mathbf{b}^* + n^{-1/2}\mathbf{v}) - L_n(\boldsymbol{\beta}^*, \mathbf{b}^*).$$

$$\eta_{n,k} = n^{-1/2}\omega_k \sum_{i=1}^{n} [I(\varepsilon_i < b^*_{\tau_k}) - \tau_k],$$

$$\mathbf{z}_n = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nabla} f_i^* \sum_{k=1}^{K} \omega_k [I(\varepsilon_i < b^*_{\tau_k}) - \tau_k],$$

Proof of Theorem 3.2.1. The idea of proof is similar to that for Theorem 2.3.1, but much more techniques are involved. Let $\alpha_n = \sqrt{p_n}(n^{-\frac{1}{2}} + a_n)$, and set $\mathcal{C}_n = \{(\mathbf{u}_n, \mathbf{v}) : \|\mathbf{u}_n\| = \|\mathbf{v}\| = C\}$. We will show that, for any $\delta > 0$, there is a large constant C such that, for large n,

$$P\{\inf_{(\mathbf{u}_n,\mathbf{v})\in\mathcal{C}_n} Q_n^{SC}(\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n, \mathbf{b}^* + \alpha_n\mathbf{v}) > Q_n^{SC}(\boldsymbol{\beta}_n^*, \mathbf{b}^*)\} \geq 1 - \delta, \qquad \text{(B.1.1)}$$

which implies that, with probability tending to one, there is a local minimum $\hat{\boldsymbol{\beta}}_n$ in the ball $\{(\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n, \mathbf{b}^* + \alpha_n\mathbf{v}) : \|\mathbf{u}_n\| \leq C, \|\mathbf{v}\| \leq C\}$ such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n^*\| = O_p(\alpha_n)$.

Let $D_n^{SC}(\mathbf{u}_n, \mathbf{v}) = Q_n^{SC}(\boldsymbol{\beta}_n^* + \alpha_n\mathbf{u}_n, \mathbf{b}^* + \alpha_n\mathbf{v}) - Q_n^{SC}(\boldsymbol{\beta}_n^*, \mathbf{b}^*)$ and $P_{\lambda_n}(\mathbf{u}_n) =$

$n\sum_{j=1}^{p_n}[p_{\lambda_n}(|\beta_{nj}^*+\alpha_n u_{nj}|)-p_{\lambda_n}(|\beta_{nj}^*|)]$. Then

$$D_n^{SC}(\mathbf{u}_n,\mathbf{v})=S_n(\mathbf{u}_n,\mathbf{v})+P_{\lambda_n}(\mathbf{u}_n), \tag{B.1.2}$$

where

$$S_n(\mathbf{u}_n,\mathbf{v})=L_n(\boldsymbol{\beta}_n^*+\alpha_n\mathbf{u}_n,\mathbf{b}_n^*+\alpha_n\mathbf{v})-L_n(\boldsymbol{\beta}^*,\mathbf{b}^*).$$

By Taylor's expansion for $f(\mathbf{x}_i,\boldsymbol{\beta}_n^*+\alpha_n\mathbf{u}_n)$ at $\boldsymbol{\beta}_n^*$, there exists a $\tilde{\boldsymbol{\beta}}_n$ between $\boldsymbol{\beta}_n^*$ and $\boldsymbol{\beta}_n^*+\alpha_n\mathbf{u}_n$, such that

$$f(\mathbf{x}_i,\boldsymbol{\beta}_n^*+\alpha_n\mathbf{u}_n)=f_{ni}^*+\alpha_n\boldsymbol{\nabla}f(\mathbf{x}_i,\tilde{\boldsymbol{\beta}}_n)'\mathbf{u}_n.$$

Let $s_{ik}=\alpha_n v_k+\alpha_n\boldsymbol{\nabla}f(\mathbf{x}_i,\tilde{\boldsymbol{\beta}}_n)'\mathbf{u}_n$. Then using the identity (A.1.4), we obtain that

$$\begin{aligned}
S_n(\mathbf{u}_n,\mathbf{v}) &= \sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}\Big\{\rho_{\tau_k}\big(\varepsilon_i-b_{\tau_k}^*-s_{ik}\big)-\rho_{\tau_k}(\varepsilon_i-b_{\tau_k}^*)\Big\}\\
&= \sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}s_{ik}[I(\varepsilon_i<b_{\tau_k}^*)-\tau_k]\\
&\quad+\sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}\int_0^{s_{ik}}[I(\varepsilon_i\le b_{\tau_k}^*+x)-I(\varepsilon_i\le b_{\tau_k}^*)]dx. \qquad \text{(B.1.3)}
\end{aligned}$$

Let

$$B_n^{(k)}=\sum_{i=1}^{n}\int_0^{s_{ik}}[I(\varepsilon_i\le b_{\tau_k}^*+x)-I(\varepsilon_i\le b_{\tau_k}^*)]dx,$$

and

$$\delta_n(\mathbf{u}_n)=\sqrt{n}\alpha_n\mathbf{u}_n'(\tilde{\mathbf{z}}_n-\mathbf{z}_n),$$

where $\tilde{\mathbf{z}}_n=n^{-1/2}\sum_{k=1}^{K}\omega_k\sum_{i=1}^{n}\nabla f(\mathbf{x}_i,\tilde{\boldsymbol{\beta}}_n)[I(\varepsilon_i<b_{\tau_k}^*)-\tau_k]$. Then $S_n(\mathbf{u}_n,\mathbf{v})$ can be rewritten as

$$S_n(\mathbf{u}_n,\mathbf{v})=\sqrt{n}\alpha_n\Big(\sum_{k=1}^{K}\eta_{n,k}v_k+\mathbf{z}_n'\mathbf{u}_n\Big)+\sum_{k=1}^{K}\omega_k B_n^{(k)}+\delta_n(\mathbf{u}_n). \tag{B.1.4}$$

By Conditions (B_1) and by directly computing the mean and variance for each component, it can be shown that

$$||\tilde{\mathbf{z}}_n - \mathbf{z}_n|| = o_p(1).$$

Then by the Cauchy-Schwartz inequality, we have

$$|\delta_n(\mathbf{u}_n)| = o_p(\sqrt{n}\alpha_n)\|\mathbf{u}_n\|. \tag{B.1.5}$$

Using Condition (B_2) and the same argument as for (A.1.7), we obtain that $\sum_{k=1}^{K} \eta_{n,k} v_k + \mathbf{z}_n' \mathbf{u}_n = O_p(\sqrt{p_n})\|\mathbf{u}_n\|$. This combined with (B.1.4) and (B.1.5) leads to

$$S_n(\mathbf{u}_n, \mathbf{v}) = \sum_{k=1}^{K} \omega_k B_n^{(k)} + O_p(n\alpha_n^2)\|\mathbf{u}_n\|. \tag{B.1.6}$$

Using Condition (C) and taking iterative expectation, we establish that

$$\begin{aligned} E[B_n^{(k)}] &= nE\Big\{\int_0^{s_{1k}} [G(b^*_{\tau_k} + x) - G(b^*_{\tau_k})]dx\Big\}. \\ &= \frac{n}{2} g(b^*_{\tau_k}) E s_{1k}^2 (1 + o(1)). \end{aligned}$$

Put $\boldsymbol{\mu}_n = E(\boldsymbol{\nabla} f^*_{n1})$ and $\boldsymbol{\Gamma}_n = E[(\boldsymbol{\nabla} f^*_{n1})^{\otimes 2}]$. By Conditions (B_1), straightforward calculation leads to $E s_{1k}^2 = \alpha_n^2 (v_k^2 + \mathbf{u}_n' \boldsymbol{\Gamma}_n \mathbf{u}_n + 2 v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o(1))$, and hence

$$E[B_n^{(k)}] = \frac{1}{2} g(b^*_{\tau_k}) n \alpha_n^2 (v_k^2 + \mathbf{u}_n' \boldsymbol{\Gamma}_n \mathbf{u}_n + 2 v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o(1)).$$

Simple algebra yields that

$$\begin{aligned} \mathrm{Var}(B_n^{(k)}) &= \sum_{i=1}^{n} E\Big\{\int_0^{s_{ik}} [I(\varepsilon_i \le b^*_{\tau_k} + x) - I(\varepsilon_i \le b^*_{\tau_k}) \\ &\quad -(G(b^*_{\tau_k} + x) - G(b^*_{\tau_k})]dx\Big\}^2 \\ &\le 4 \sum_{i=1}^{n} E(s_{ik}^2) = O(n\alpha_n^2)(v_k^2 + \mathbf{u}_n' \boldsymbol{\Gamma}_n \mathbf{u}_n + 2 v_k \boldsymbol{\mu}_n' \mathbf{u}_n). \end{aligned}$$

Therefore,

$$B_n^{(k)} = \frac{1}{2} g(b^*_{\tau_k}) n \alpha_n^2 (v_k^2 + \mathbf{u}_n' \boldsymbol{\Gamma}_n \mathbf{u}_n + 2 v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o_p(1)). \tag{B.1.7}$$

Combining (B.1.6) and (B.1.7) yields that

$$\begin{aligned} S_n(\mathbf{u}_n, \mathbf{v}) &= \frac{1}{2} n\alpha_n^2 \sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)(1 + o_p(1)) \\ &+ O_p(n\alpha_n^2)\|\mathbf{u}_n\|. \end{aligned} \tag{B.1.8}$$

Finally, we consider the Taylor expansion for the penalty term $P_{\lambda_n}(\mathbf{u}_n)$ in (B.1.2). Using $p_{\lambda_n}(0) = 0$ and Condition (A_4), we establish that

$$\begin{aligned} P_{\lambda_n}(\mathbf{u}_n) &\geq n \sum_{j=1}^{s_n} \big\{ p_{\lambda_n}(|\beta^*_{nj} + \alpha_n u_{nj}|) - p_{\lambda_n}(|\beta^*_{nj}|) \big\} \\ &= \sum_{j=1}^{s_n} \big[n\alpha_n p'_{\lambda_n}(|\beta^*_{nj}|)\mathrm{sgn}(\beta^*_{nj}) u_{nj} + \frac{1}{2} n\alpha_n^2 p''_{\lambda_n}(|\beta^*_{nj}|) u_{nj}^2 (1 + o(1)) \big]. \end{aligned}$$

Then by the Hölder inequality and Conditions (A_2)-(A_3), we have

$$\begin{aligned} P_{\lambda_n}(\mathbf{u}_n) &\geq -(\sqrt{s_n} n\alpha_n a_n \|\mathbf{u}_n\| + \frac{1}{2} n\alpha_n^2 b_n \|\mathbf{u}_n\|^2 (1 + o(1))) \\ &\geq -(n\alpha_n^2 \|\mathbf{u}_n\| + o_p(n\alpha_n^2)). \end{aligned} \tag{B.1.9}$$

It follows from (B.1.2), (B.1.8) and (B.1.9) that $D_n^{SC}(\mathbf{u}_n, \mathbf{v})$ is dominated by the positive quadratic term $\frac{1}{2} n\alpha_n^2 \sum_{k=1}^{K} \omega_k g(b^*_\tau)(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)$ as long as $\|\mathbf{u}_n\|$ and $\|\mathbf{v}\|$ are large enough. Therefore, $(B.1.1)$ holds, and the proof is completed.

Lemma B.1.1. *Suppose Conditions (A_1)-(A_4), (B_1)-(B_3) and (C) hold. If $\lambda_n \to 0$, $\sqrt{n_p}\lambda_n \to \infty$, and $p_n^3/n \to 0$, as $n \to \infty$, then with probability tending to 1, for any given $\boldsymbol{\beta}_{n1}$ satisfying $\|\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}^*_{n1}\| = O_p(n_p^{-\frac{1}{2}})$, $\|\mathbf{b} - \mathbf{b}^*\| = O_p(n_p^{-\frac{1}{2}})$ and any constant C, we have*

$$Q_n^{SC}((\boldsymbol{\beta}_{n1}', \mathbf{0}')', \mathbf{b}) = \min_{\|\boldsymbol{\beta}_{n2}\| \leq C n_p^{-\frac{1}{2}}} Q_n^{SC}((\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')', \mathbf{b}).$$

Proof. Let $\sqrt{n_p}(\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}^*_{n1}) = \mathbf{u}_{n1}$, $\sqrt{n_p}(\boldsymbol{\beta}_{n2} - \boldsymbol{\beta}^*_{n2}) = \mathbf{u}_{n2}$, $\mathbf{u}_n = (\mathbf{u}'_{n1}, \mathbf{u}'_{n2})'$, and $\sqrt{n_p}(\mathbf{b} - \mathbf{b}^*) = \mathbf{v}$. By the definition of $Q_n^{SC}(\boldsymbol{\beta}_n, \mathbf{b})$ and $S_n(\mathbf{u}_n, \mathbf{v})$, we have

$$\begin{aligned} & Q_n^{SC}((\boldsymbol{\beta}'_{n1}, \mathbf{0}')', \mathbf{b}) - Q_n^{SC}((\boldsymbol{\beta}'_{n1}, \boldsymbol{\beta}'_{n2})', \mathbf{b}) \\ = \;& S_n((\mathbf{u}'_{n1}, \mathbf{0}'))', \mathbf{v}) - S_n((\mathbf{u}'_{n1}, \mathbf{u}'_{n2})', \mathbf{v}) - n \sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|). \end{aligned}$$

Using (B.1.8), we obtain that

$$\begin{aligned} S_n((\mathbf{u}'_{n1}, \mathbf{0}')', \mathbf{v}) &= \frac{1}{2} p_n \sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'_{n1} \boldsymbol{\Gamma}_{n11} \mathbf{u}_n + 2 v_k \boldsymbol{\mu}'_{n1} \mathbf{u}_{n1})(1 + o_p(1)) \\ &+ O_p(p_n)\|\mathbf{u}_{n1}\|, \end{aligned}$$

and

$$\begin{aligned} S_n((\mathbf{u}'_{n1}, \mathbf{u}'_{n2})', \mathbf{v}) &= \frac{1}{2} p_n \sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'_{n} \boldsymbol{\Gamma}_{n} \mathbf{u}_n + 2 v_k \boldsymbol{\mu}'_{n} \mathbf{u}_{n})(1 + o_p(1)) \\ &+ O_p(p_n)\|\mathbf{u}_{n}\|. \end{aligned}$$

By Condition (B_2), we have $\mathbf{u}'_n \mathbf{G}_n \mathbf{u}_n \le \|\mathbf{G}_n\| \|\mathbf{u}_n\|^2 = O_p(\sqrt{p_n})$. Note that $\|\mathbf{u}_{n1}\| = O_p(1)$, $\|\mathbf{v}\| = O_p(1)$, and $0 < \|\mathbf{u}_{n2}\| \le C$. It follows that $S_n((\mathbf{u}'_{n1}, \mathbf{0}')', \mathbf{v}) = O_p(p_n^{3/2})$ and $S_n((\mathbf{u}'_{n1}, \mathbf{u}'_{n2})', \mathbf{v}) = O_p(p_n^{3/2})$. Using $p_{\lambda_n}(0) = 0$ and the mean value theorem, we arrive at

$$\begin{aligned} n \sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|) &= n \sum_{j=s_n+1}^{p_n} p'_{\lambda_n}(|\beta^\dagger_{nj}|)|\beta^\dagger_{nj}| \\ &= n\lambda_n \sum_{j=s_n+1}^{p_n} \frac{p'_{\lambda_n}(|\beta^\dagger_{nj}|)}{\lambda_n} |\beta^\dagger_{nj}| \\ &\ge p_n^2 \sqrt{\frac{n}{p_n^3}} \sqrt{n_p} \lambda_n (\liminf_{n\to+\infty} \liminf_{\theta\to 0+} p'_{\lambda_n}(\theta)/\lambda_n) \sum_{j=s_n+1}^{p_n} |\beta^\dagger_{nj}|, \end{aligned}$$

where $0 < \beta^\dagger_{nj} < |\beta_j|$ for $j = s_n + 1, \ldots, p_n$. Since $\sqrt{n_p}\lambda_n \to \infty$ and $p_n^3/n \to 0$, $p_n^2 \sqrt{\frac{n}{p_n^3}} \sqrt{n_p}\lambda_n$ is of higher order than $O_p(p_n^{3/2})$. This together with Condition (A_1) means that $Q_n^{SC}((\boldsymbol{\beta}'_{n1}, \mathbf{0}')', \mathbf{b}) - Q_n^{SC}((\boldsymbol{\beta}'_{n1}, \boldsymbol{\beta}'_{n2})', \mathbf{b})$ is dominated by the term

$-n \sum_{j=s_n+1}^{p_n} p_{\lambda_n}(|\beta_{nj}|)$ for larger n, which is less than zero. This result of the lemma holds.

Proof of Theorem 3.2.2.

(i) follows Lemma (B.1.1).

(ii) Let $\mathbf{u}_n = \alpha_n^{-1}(\boldsymbol{\beta}_n - \boldsymbol{\beta}_n^*)$. Partition the vectors $\mathbf{u}_n = (\mathbf{u}_{n1}', \mathbf{u}_{n2}')'$, $\mathbf{z} = (\mathbf{z}_1', \mathbf{z}_2')'$, and $\boldsymbol{\nabla} f_{ni}^* = ((\boldsymbol{\nabla} f_{ni1}^*)', (\boldsymbol{\nabla} f_{ni2}^*)')'$, in the same way as $\boldsymbol{\beta}_n = (\boldsymbol{\beta}_{n1}', \boldsymbol{\beta}_{n2}')'$. Partition $\mathbf{G}_n$ as a 2×2 block matrix $\mathbf{G}_n = (\mathbf{G}_{nij})$ (for $i, j = 1, 2$), where $\mathbf{G}_{n11}$ was defined before, and let

$$P_{\lambda_n}(\mathbf{u}_{n1}) = n \sum_{j=1}^{s_n} (p_{\lambda_n}(|\beta_{nj}^* + \alpha_n u_{nj}|) - p_{\lambda_n}(|\beta_{nj}^*|)).$$

Then by Taylor's expansion and Condition (B_3), the jth component of $\partial P_{\lambda_n}(\mathbf{u}_{n1})/\partial \mathbf{u}_{n1}$ is

$$n\alpha_n p_{\lambda_n}'(|\beta_{nj}^*|)\mathrm{sgn}(\beta_{nj}^*)u_{nj} + \frac{1}{2} n\alpha_n^2 p_{\lambda_n}''(|\beta_{nj}^*|)u_{nj}^2(1 + o(1)).$$

By (B.1.2), we have

$$D_n^{SC}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) \;=\; S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) + P_{\lambda_n}(\mathbf{u}_{n1}).$$

Put $t_{ik}(\mathbf{u}_{n1}, \mathbf{u}_{n2}, v_k) = \alpha_n v_k + f(\mathbf{x}_i, \boldsymbol{\beta}_n^* + \alpha_n \mathbf{u}_n) - f(\mathbf{x}_i, \boldsymbol{\beta}_n^*)$. Then by (B.1.3), we have

$$S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) = \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \big\{\rho_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\mathbf{u}_{n1}, \mathbf{0}, v_k)) - \rho_{\tau_k}(\varepsilon_i - b_{\tau_k}^*)\big\}.$$

Therefore, the minimizer $(\hat{\mathbf{u}}_{n1}, \hat{\mathbf{v}})$ of $D_n^{SC}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v})$ satisfies the score equations:

$$\begin{aligned} & n^{-1} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \psi_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) \boldsymbol{\nabla} f_{ni1}^*(1 + o_p(1)) \\ & = \mathbf{b}_n + \alpha_n \boldsymbol{\Sigma}_{\lambda_n} \hat{\mathbf{u}}_{n1}(1 + o_p(1)) \end{aligned} \tag{B.1.10}$$

and

$$\omega_k \sum_{i=1}^{n} \psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k} - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) = 0. \tag{B.1.11}$$

Note that $\psi_\tau(u) = \tau - I(u < 0)$. We rewrite

$$\begin{aligned} & n^{-1} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k} - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) \boldsymbol{\nabla} f^*_{ni1} \\ = \; & -n^{-1/2} \mathbf{z}_{n1} + \sum_{k=1}^{K} \omega_k B^{(k)}_{n2}, \end{aligned}$$

where $\mathbf{z}_{n1} = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\nabla} f^*_{ni1} \sum_{k=1}^{K} \omega_k [I(\varepsilon_i < b^*_{\tau_k}) - \tau_k]$ and

$$\begin{aligned} B^{(k)}_{n2} &= n^{-1} \sum_{i=1}^{n} [\psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k} - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) - \psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k})] \boldsymbol{\nabla} f^*_{ni1} \\ &= n^{-1} \sum_{i=1}^{n} [I(\varepsilon_i < b^*_{\tau_k}) - I(\varepsilon_i < b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))] \boldsymbol{\nabla} f^*_{ni1}. \end{aligned}$$

Decompose $B^{(k)}_{n2}$ into $B^{(k)}_{n2} = B^{(k)}_{n21} + B^{(k)}_{n22}(\hat{\mathbf{u}}_{n1}, \hat{v}_k)$, where

$$B^{(k)}_{n21} = n^{-1} \sum_{i=1}^{n} [G(b^*_{\tau_k}) - G(b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))] \boldsymbol{\nabla} f^*_{ni1},$$

$$\begin{aligned} B^{(k)}_{n22}(\hat{\mathbf{u}}_{n1}, \hat{v}_k) &= n^{-1} \sum_{i=1}^{n} \big\{ [I(\varepsilon_i < b^*_{\tau_k}) - I(\varepsilon_i < b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))] \\ &\quad - [G(b^*_{\tau_k}) - G(b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k))] \big\} \boldsymbol{\nabla} f^*_{ni1}. \end{aligned}$$

By the mean value theorem,

$$\begin{aligned} B^{(k)}_{n21} &= -n^{-1} \sum_{i=1}^{n} g(b^{**}_{\tau_k}) t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k) \boldsymbol{\nabla} f^*_{ni1} \\ &= -n^{-1} \sum_{i=1}^{n} g(b^*_{\tau_k}) t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k) \boldsymbol{\nabla} f^*_{ni1} \\ &\quad -n^{-1} \sum_{i=1}^{n} [g(b^{**}_{\tau_k}) - g(b^*_{\tau_k})] t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k) \boldsymbol{\nabla} f^*_{ni1}, \end{aligned}$$

where $b^{**}_{\tau_k}$ is between $b^*_{\tau_k}$ and $b^*_{\tau_k} + t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)$. The first term above is

$$-\alpha_n g(b^*_{\tau_k})(\mathbf{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k)(1 + o_p(1)),$$

and the second term is dominated by the first term. Therefore,

$$B^{(k)}_{n21} = -\alpha_n g(b^*_{\tau_k})(\mathbf{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k)(1 + o_p(1)).$$

By directly calculating the mean and variance, we obtain that $B^{(k)}_{n22}(\mathbf{u}_{n1}, v_k) = o_p(\alpha_n)$. For any given large number $M > 0$, using Bickel's (1975) chaining approach, we can show as in Jiang et al. (2001) that

$$\sup_{\|\mathbf{u}\| \le M} |B^{(k)}_{n22}(\mathbf{u}_{n1}, v_k)| = o_p(\alpha_n).$$

Since $\hat{\mathbf{u}}_{n1} = O_p(1)$ and $\hat{v}_k = O_p(1)$, $B^{(k)}_{n22}(\hat{\mathbf{u}}_{n1}, \hat{v}_k) = o_p(\alpha_n)$. Hence,

$$B^{(k)}_{n2} = -\alpha_n g(b^*_{\tau_k})(\mathbf{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k)(1 + o_p(1)).$$

This combined with (B.1.10) leads to

$$-(n^{-1/2}\mathbf{z}_{n1} + \mathbf{b}_n) = \alpha_n\{\sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(\mathbf{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k) + \mathbf{\Sigma}_{\lambda_n}\hat{\mathbf{u}}_{n1}\}(1 + o_p(1)). \tag{B.1.12}$$

Similarly, the score equation (B.1.11) can be simplified as

$$n^{-1/2}\eta_{n,k} + \alpha_n \omega_k g(b^*_{\tau_k})(\hat{v}_k + \boldsymbol{\mu}'_{n1}\hat{\mathbf{u}}_{n1}(1 + o_p(1))) = 0. \tag{B.1.13}$$

Solving (B.1.12) and (B.1.13), we obtain that

$$\begin{aligned} & \alpha_n\big(\mathbf{G}_{n11} + \mathbf{\Sigma}_{\lambda_n}/\boldsymbol{\omega}'\mathbf{g}\big)\hat{\mathbf{u}}_{n1} + \mathbf{b}_n/\boldsymbol{\omega}'\mathbf{g} \\ = \; & -n^{-1/2}(\mathbf{z}_{n1} - \boldsymbol{\mu}_{n1}\sum_{k=1}^{K}\eta_{n,k})/\boldsymbol{\omega}'\mathbf{g} + o_p(n^{-1/2}), \end{aligned}$$

where $(\mathbf{z}_{n1} - \boldsymbol{\mu}_{n1}\sum_{k=1}^{K}\eta_{n,k})/\boldsymbol{\omega}'\mathbf{g}$ is normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\boldsymbol{\omega})\mathbf{G}_{n11}$. Note that $\hat{\mathbf{u}}_{n1} = \alpha_n^{-1}(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}^*_{n1})$. It follows that

$$\begin{aligned} & \sqrt{n}\mathbf{A}_n\mathbf{G}_{n11}^{-\frac{1}{2}}\big(\mathbf{G}_{n11} + \boldsymbol{\Sigma}_{\lambda_n}/\boldsymbol{\omega}'\mathbf{g}\big) \\ \times \quad & [(\hat{\boldsymbol{\beta}}_{n1} - \boldsymbol{\beta}^*_{n1}) + \big(\mathbf{G}_{n11} + \boldsymbol{\Sigma}_{\lambda_n}/\boldsymbol{\omega}'\mathbf{g}\big)^{-1}\mathbf{b}_n/\boldsymbol{\omega}'\mathbf{g}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B}). \end{aligned}$$

Proof of Theorem 3.2.3. This can be proven using an argument similar to that for Theorem 3.2.1. Let $\sqrt{n_p}(\boldsymbol{\beta}_n - \boldsymbol{\beta}^*_n) = \mathbf{u}_n$ and $\sqrt{n_p}(\mathbf{b} - \mathbf{b}^*) = \mathbf{v}$. Partition the vectors $\mathbf{u}_n = (\mathbf{u}'_{n1}, \mathbf{u}'_{n2})'$ according to $\boldsymbol{\beta}_n = (\boldsymbol{\beta}'_{n1}, \boldsymbol{\beta}'_{n2})'$. We will show that for any $\delta > 0$,

$$P\Big\{\inf_{(\mathbf{u}_n,\mathbf{v})\in\mathcal{C}_n} Q_n^{AL}\big(\boldsymbol{\beta}^*_n + n_p^{-1/2}\mathbf{u}_n, \mathbf{b}^* + n_p^{-1/2}\mathbf{v}\big) > Q_n^{AL}(\boldsymbol{\beta}^*_n, \mathbf{b}^*)\Big\} \geq 1 - \delta, \quad \text{(B.1.14)}$$

which implies that there is a local minimum $\hat{\boldsymbol{\beta}}_n^{AL}$ in the ball $\{(\boldsymbol{\beta}^*_n + n_p^{-1/2}\mathbf{u}_n, \mathbf{b}^* + n_p^{-1/2}\mathbf{v}) : \|\mathbf{u}_n\| \leq C, \|\mathbf{v}\| \leq C\}$ such that $\|\hat{\boldsymbol{\beta}}_n^{AL} - \boldsymbol{\beta}^*_n\| = O_p(n_p^{-1/2})$.

Let $P_{h_n}(\mathbf{u}_n) = \sum_{j=1}^{p_n} nh_n|\tilde{\beta}_{nj}|^{-\gamma}(|\beta^*_{nj} + n_p^{-1/2}u_{nj}| - |\beta^*_{nj}|)$ and

$$P_{h_n}(\mathbf{u}_{n1}) = \sum_{j=1}^{s_n} nh_n|\tilde{\beta}_{nj}|^{-\gamma}(|\beta^*_{nj} + n_p^{-1/2}u_{nj}| - |\beta^*_{nj}|).$$

Define

$$D_n^{AL}(\mathbf{u}_n, \mathbf{v}) = Q_n^{AL}(\boldsymbol{\beta}^*_n + n_p^{-1/2}\mathbf{u}_n, \mathbf{b}^* + n_p^{-1/2}\mathbf{v}) - Q_n^{AL}(\boldsymbol{\beta}^*_n, \mathbf{b}^*).$$

Then

$$D_n^{AL}(\mathbf{u}_n, \mathbf{v}) = S_n(\mathbf{u}_n, \mathbf{v}) + P_{h_n}(\mathbf{u}_n). \quad \text{(B.1.15)}$$

Similar to (B.1.8),

$$\begin{aligned} S_n(\mathbf{u}_n, \mathbf{v}) &= \frac{1}{2}p_n\sum_{k=1}^{K}\omega_k g(b^*_{\tau_k})(v_k^2 + \mathbf{u}'_n\boldsymbol{\Gamma}_n\mathbf{u}_n + 2v_k\boldsymbol{\mu}'_n\mathbf{u}_n)(1 + o_p(1)) \\ &+ O_p(p_n)\|\mathbf{u}_n\|. \end{aligned}$$

Note that $\tilde{\beta}_{nj}$ is the $n_p^{-1/2}$ consistent estimator for β_{nj}^* such that $\tilde{\beta}_{nj} = \beta_{nj}^*(1+o_p(1))$. Then $|\tilde{\beta}_{nj}|^\gamma = |\beta_{nj}^*|^\gamma(1+o_p(1))$. Therefore, $P_{h_n}(\mathbf{u}_n) > P_{h_n}(\mathbf{u}_{n1})$ and

$$\begin{aligned} P_{h_n}(\mathbf{u}_{n1}) &\rightarrow \sum_{j=1}^{s_n} \frac{nh_n}{|\tilde{\beta}_{nj}|^\gamma} n_p^{-1/2} u_{nj} sgn(\beta_{nj}^*) \\ &\geq -\frac{n^{1/2}h_n}{\min\limits_{1\leq j\leq s_n} |\beta_{nj}^*|^\gamma (1+o_p(1))} p_n \|\mathbf{u}_n\|. \end{aligned}$$

By Condition (B_4) and $n^{1/2}h_n \rightarrow 0$, it is easy to see that D_n^{AL} is dominated by the positive quadratic term

$$\frac{1}{2} p_n \sum_{k=1}^{K} \omega_k g(b_{\tau_k}^*)(v_k^2 + \mathbf{u}_n' \mathbf{\Gamma}_n \mathbf{u}_n + 2v_k \boldsymbol{\mu}_n' \mathbf{u}_n)$$

as long as $\|\mathbf{u}_n\|$ and $\|\mathbf{v}\|$ are allowed to be large enough. This means (B.1.14) holds.

Proof of Theorem 3.2.4.

(i) Following the same argument as for Lemma B.1.1, we can complete the proof for sparsity.

(ii) By (B.1.15), we have

$$\begin{aligned} D_n^{AL}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) &= S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) + P_{h_n}(\mathbf{u}_{n1}) \\ &= S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) + \sum_{j=1}^{s_n} \frac{nh_n}{|\tilde{\beta}_{nj}|^\gamma} n_p^{-1/2} u_{nj} \mathrm{sgn}(\beta_{nj}^*)(1+o(1)). \end{aligned}$$

Let $t_{ik}(\mathbf{u}_{n1}, \mathbf{u}_{n2}, v_k) = n_p^{-1/2} v_k + f(\mathbf{x}_i, \boldsymbol{\beta}_n^* + n_p^{-1/2}\mathbf{u}_n) - f(\mathbf{x}_i, \boldsymbol{\beta}_n^*)$. Then

$$S_n((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v}) = \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \big\{ \rho_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\mathbf{u}_{n1}, \mathbf{0}, v_k)) - \rho_{\tau_k}(\varepsilon_i - b_{\tau_k}^*) \big\}.$$

Therefore, the minimizer $(\hat{\mathbf{u}}_{n1}, \hat{v}_k)$ of $D_n^{AL}((\mathbf{u}_{n1}', \mathbf{0}')', \mathbf{v})$ satisfies the score equations:

$$\begin{aligned} & n^{-1} \sum_{k=1}^{K} \omega_k \sum_{i=1}^{n} \psi_{\tau_k}(\varepsilon_i - b_{\tau_k}^* - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) \boldsymbol{\nabla} f_{ni1}^*(1+o_p(1)) \\ &= h_n \mathbf{d}_n (1+o_p(1)) \end{aligned}$$

and $\omega_k \sum_{i=1}^{n} \psi_{\tau_k}(\varepsilon_i - b^*_{\tau_k} - t_{ik}(\hat{\mathbf{u}}_{n1}, \mathbf{0}, \hat{v}_k)) = 0$.Then similar to (B.1.12) and (B.1.13), the above score equations can be simplified as

$$-(n^{-1/2}\mathbf{z}_{n1} + h_n \mathbf{d}_n) = n_p^{-1/2} \sum_{k=1}^{K} \omega_k g(b^*_{\tau_k})(\mathbf{\Gamma}_{n11}\hat{\mathbf{u}}_{n1} + \boldsymbol{\mu}_{n1}\hat{v}_k)(1 + o_p(1))$$

and $n^{-1/2}\eta_{n,k} + n_p^{-1/2}\omega_k g(b^*_{\tau_k})(\hat{v}_k + \boldsymbol{\mu}'_{n1}\hat{\mathbf{u}}_{n1}(1 + o_p(1))) = 0$. Solving the above equations, we obtain that

$$\begin{aligned} & n_p^{-1/2}\mathbf{G}_{n11}\hat{\mathbf{u}}_{n1}(1 + o_p(1)) + h_n \mathbf{d}_n/\boldsymbol{\omega}'\mathbf{g} \\ = \quad & -n^{-1/2}(\mathbf{z}_{n1} - \boldsymbol{\mu}_{n1} \sum_{k=1}^{K} \eta_{n,k})/\boldsymbol{\omega}'\mathbf{g} + o_p(n^{-1/2}). \end{aligned}$$

Note that $n_p^{-1/2}\hat{\mathbf{u}}_{n1}^{AL} = \hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}^*_{n1}$. It follows that

$$\sqrt{n}\mathbf{A}_n \mathbf{G}_{n11}^{\frac{1}{2}}[(\hat{\boldsymbol{\beta}}_{n1}^{AL} - \boldsymbol{\beta}^*_{n1}) + h_n \mathbf{G}_{n11}^{-1}\mathbf{d}_n/\boldsymbol{\omega}'\mathbf{g}] \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\mathbf{B}).$$

Appendix C

Proofs of Theorems in Chapter 4

The arguments for the theorems in Chapter 4 are different from those used nonlinear regression models, and the previous arguments cannot be used to deal with the simultaneous estimation of the AR and ARCH parameters. We will need the theory about generalized functions of random variables and stochastic limit operations with partial sums of these generalized functions of random variables.

C.1 Generalized functions

For convenience, let us introduce at first the theory of generalized functions, which can be found in Phillips (1991b) and Phillips (1995).

C.1.1 Generalized functions of random variables and generalized limit theory

Phillips' idea is to treat non-smooth objective criteria like $\rho_\tau(.)$ that appears in QR estimation as generalized function and uses generalized Taylor's series expansions to represent their local behaviors.

Although $\rho'_\tau(\cdot)$, has no meaning as an ordinary derivative for $\rho_\tau(\cdot)$, it can be interpreted in terms of the derivative of generalized function $\rho_\tau(\cdot)$ by using the "regular sequence" approach given in Lighthill (1958).

Definition C.1.1. A regular sequence for any generalized function $f(x)$ is a sequence $f^m(x)$ of good functions (i.e., functions which are everywhere differentiable any number of times and such that it and its derivatives are $O(|x|^{-N})$ as $|x| \to \infty$ for all N, the set of such functions is denoted as GF) converging weakly to $f(x)$, denoted by $f^m \Rightarrow f$, in the sense that

$$\int_{-\infty}^{\infty} f(x)F(x)dx = \lim_{m\to\infty} \int_{-\infty}^{\infty} f^m(x)F(x)dx \tag{C.1.1}$$

exists for any $F \in GF$.

Since the sequence $f^m(\cdot)$ is measurable, $f^m(u_t)$ has a meaning as ordinary random variable on the probability space where u_t is defined. From (C.1.1), it follows that, if $pdf(u) \in GF$ is the density of u_t, then the expectation of the generalized function $f(\cdot)$ of u_t is defined by

$$E[f(u_t)] = \lim_{m\to\infty} E(f^m(u_t)) = \lim_{m\to\infty} \int_{-\infty}^{\infty} f^m(u)pdf(u)du, \tag{C.1.2}$$

which also means $E[f^m(u_t)] \to E[f(u_t)]$.

Next, we introduce a weak law of large numbers (WLLN) or strong law of large numbers (SLLN) for partial sums of the generalized function of random variables $f(u_t)$.

Definition C.1.2. A WLLN and SLLN for $f(u_t)$, that is,

$$T^{-1}\sum_{t=1}^{T} f(u_t) \to_p E[f(u_t)], \tag{C.1.3}$$

is defined by the corresponding weak and strong laws for partial sums of the regular sequence $f^m(u_t)$ of ordinary random variables, that is, by

$$T^{-1}\sum_{t=1}^{T} f^m(u_t) \to_p E[f^m(u_t)], \ \forall m, \tag{C.1.4}$$

and the limit that appears on the right side of (C.1.4) is given by (C.1.2).

Lemma C.1.1. *(Lemma 3.1 Phillips 1995) (SLLN for Ordinary Random Variables as Generalized Functions of Random Variables). Suppose u_t is strictly stationary and ergodic and $f(u_t)$ is an ordinary (measurable) function of u_t. Then, (C.1.3) holds in the sense of ordinary random sequences if and only f it holds in the sense of generalized functions of random sequences, that is, if and only if (C.1.4) holds.*

C.1.2 Ordinary functions as generalized functions

The following examples illustrate ordinary functions as generalized functions.

Example 1. (Phllips) The discontinuous function sgn(x), which is 1 for $x > 0$, 0 for $x = 0$ and -1 for $x < 0$, satisfies the condition of Definition 7 (Lighthill, 1958, p.21) as a generalized function and can be defined by the following regular sequence:

$$sgn^m(u) = \int_{-\infty}^{\infty} sgn(v)S(m(v-u))me^{-v^2/m^2}dv,$$

where the function $S(.)$ is a "smudge function" defined in Definition 7 (Lighthill, 1958, p.21) whose role in $sgn^m(u)$ is to smudge out $sgn(v)$ when v is outside the interval $(u - m^{-1}, u + m^{-1})$.

The regular sequence $sgn^m(u_t)$ has the property $E[(sgn^m(u_t))^2] \to E[sgn^2(u_t)]$ (see for example Phillips, 1995, p. 923). It follow that

$$var(sgn^m(u_t)) \to var(sgn(u_t)) \tag{C.1.5}$$

from (C.1.2).

Example 2. (Phllips) The sign function $sgn(x)$, as a generalized function, has its derivative $sgn'(x) = 2\delta(x)$, where $\delta(.)$ is the Dirac delta generalized function with the property that $\int_{-\infty}^{\infty} \delta(x)F(x)dx = F(0)$ for any continuous function $F(x)$. With this property, it follows that $\delta^m(u) = (m/\pi)^{1/2}e^{-mu^2}$ is a regular sequence

for $\delta(u)$. Note that $\rho''_\tau(u) = \delta(u)$. Therefore, $\delta^m(u)$ is also a regular sequence for $\rho''_\tau(u)$.

Let $\rho_\tau(u) = \frac{1}{2}|u| + \frac{2\tau-1}{2}u$ and $\psi_\tau(u) = \frac{1}{2}sgn(u) + \frac{2\tau-1}{2} = \tau - I(u < 0)$. From Definition 6 (Lighthill, 1958, p.18), $\rho_\tau(u)$ can be defined by the sequences

$$\rho^m_\tau(u) = \frac{1}{2}sgn^m(u)u + \frac{2\tau-1}{2}u \tag{C.1.6}$$

and $\rho'_\tau(u)$ can be defined by the sequence

$$\psi^m_\tau(u) = \frac{1}{2}sgn^m(u) + \frac{2\tau-1}{2}. \tag{C.1.7}$$

Also, we know $\rho'_\tau(u)$ and $\rho''_\tau(u)$ are defined by the sequence $\rho^{m\prime}_\tau(u)$ and $\rho^{m\prime\prime}_\tau(u)$, respectively. Therefore, $\rho^{m\prime}_\tau(u)$ and $\psi^m_\tau(u)$ are equivalent regular sequences of $\rho'_\tau(u)$, and $\rho^{m\prime\prime}_\tau(u)$ and $\delta^m(u)$ are equivalent regular sequences of $\rho''_\tau(u)$. Combining (C.1.5)-(C.1.7) produces that $\text{var}(\psi^m_\tau(u_t)) \to var(\psi_\tau(u_t))$ and

$$var(\rho^{m\prime}_\tau(u_t))) \to var(\psi_\tau(u_t)). \tag{C.1.8}$$

C.2 Proofs of Theorems

In this appendix, we give rigorous proofs of our theorems. To facilitate the formulation of proofs, we introduce the following notations:

$$\mu^m_k = E[\rho^{m\prime}_{\tau_k}(\log(|u_t|) - c^*_{\tau_k})],$$

$$\sigma^m(\boldsymbol{\omega}) = var[\sum_{k=1}^{K}\omega_k\rho^{m\prime}_{\tau_k}(\log(|u_t|) - c^*_{\tau_k})],$$

$$\mathbf{r}^m_n \equiv -\frac{1}{\sqrt{n}}\sum_{t=s'+1}^{n}\mathbf{J}_t\sum_{k=1}^{K}\omega_k\rho^{m\prime}_{\tau_k}(\log(|u_t|) - c^*_{\tau_k}),$$

$$\mathbf{z}^m_n \equiv -\frac{1}{\sqrt{n}}\sum_{t=s'+1}^{n}\mathbf{D}_{ht2}\sum_{k=1}^{K}\omega_k\rho^{m\prime}_{\tau_k}(\log(|u_t|) - c^*_{\tau_k}), \tag{C.2.9}$$

$$q_{n,k}^m \equiv -\frac{1}{\sqrt{n}} \sum_{t=s'+1}^{n} \rho_{\tau_k}^{m\prime}(\log(|u_t|) - c_{\tau_k}^*)$$

$$\mathbf{q}_n^m = (q_{n,1}^m, \ldots, q_{n,K}^m)'$$

and

Proof of Theorem 4.3.1. Let $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$ and $\sqrt{n}(b_\tau - b_\tau^*) = v$. Put $l(v, \mathbf{u}) = (h_t(\boldsymbol{\beta}^* + n^{-1/2}\mathbf{u}))^{-1}\varepsilon_t - (b_\tau^* + n^{-1/2}v)$. Define

$$S_n = S_n(v, \mathbf{u}) = \sum_{t=s+1}^{n} \{\rho_\tau(l(v, \mathbf{u})) - \rho_\tau((h_t(\boldsymbol{\beta}^*))^{-1}\varepsilon_t - b_\tau^*)\}.$$

Then minimizing (4.3.4) is equivalent to minimizing S_n. However, $S_n(v, \mathbf{u})$, as a generalized process, is defined by the following regular sequence of process

$$S_n^m(v, \mathbf{u}) = \sum_{t=s+1}^{n} \{\rho_\tau^m(l(v, \mathbf{u})) - \rho_\tau^m((h_t(\boldsymbol{\beta}^*))^{-1}\varepsilon_t - b_\tau^*)\},$$

where $\rho_\tau^m(\cdot)$ is the regular sequence defined in (C.1.6).

Denote by $h_t = h_t(\boldsymbol{\beta}^*)$, $\boldsymbol{\nabla} h_t = -\frac{\mathbf{Z}_t}{h_t^2}$, $\boldsymbol{\nabla}^2 h_t = \frac{2\mathbf{Z}_t\mathbf{Z}_t'}{h_t^3}$ and

$$\boldsymbol{\nabla}^2\rho_\tau^m = \frac{1}{n}\rho_\tau^{m\prime\prime}(u_t - b_\tau^*)\begin{pmatrix} 1 & \varepsilon_t\boldsymbol{\nabla}h_t' \\ \varepsilon_t\boldsymbol{\nabla}h_t & \varepsilon_t^2\boldsymbol{\nabla}h_t\boldsymbol{\nabla}h_t' \end{pmatrix} + \frac{1}{n}\rho_\tau^{m\prime}(u_t - b_\tau^*)\begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \varepsilon_t\boldsymbol{\nabla}^2 h_t \end{pmatrix},$$

which is the Hessian matrix of $\rho_\tau^m(l(v, \mathbf{u}))$ with respect to $(v, \mathbf{u}')'$. By Taylor's expansions, we have

$$\begin{aligned} S_n^m(v, \mathbf{u}) &= \sum_{t=s+1}^{n} \{\rho_\tau^{m\prime}(u_t - b_\tau^*)(n^{-\frac{1}{2}}\varepsilon_t\boldsymbol{\nabla}h_t'\mathbf{u} - n^{-\frac{1}{2}}v) \\ &+ \frac{1}{2n}[\rho_\tau^{m\prime\prime}(u_t - b_\tau^*)(v^2 + 2\varepsilon_t\mathbf{u}'\boldsymbol{\nabla}h_t v) \\ &+ \mathbf{u}'(\varepsilon_t^2\rho_\tau^{m\prime\prime}(u_t - b_\tau^*)\boldsymbol{\nabla}h_t\boldsymbol{\nabla}h_t' + \varepsilon_t\rho_\tau^{m\prime}(u_t - b_\tau^*)\boldsymbol{\nabla}^2 h_t)\mathbf{u} \\ &\times (1 + o_p(1))]\}. \end{aligned}$$

Rewrite $S_n^m(v, \mathbf{u})$ as

$$S_n^m(v, \mathbf{u}) = q_{n,0}^m v + (\boldsymbol{\gamma}_n^m)'\mathbf{u} + B_n^m, \tag{C.2.10}$$

where

$$q_{n,0}^m = -\frac{1}{\sqrt{n}} \sum_{t=s+1}^{n} \rho_\tau^{m\prime}(u_t - b_\tau^*),$$

$$\boldsymbol{\gamma}_n^m = -\frac{1}{\sqrt{n}} \sum_{t=s+1}^{n} u_t \rho_\tau^{m\prime}(u_t - b_\tau^*) \frac{\mathbf{Z}_t}{h_t}$$

and

$$\begin{aligned} B_n^m &= \frac{1}{2n} \sum_{t=s+1}^{n} \rho_\tau^{m\prime\prime}(u_t - b_\tau^*) v^2 - \mathbf{u}' \{ \frac{1}{n} \sum_{t=s+1}^{n} u_t \rho_\tau^{m\prime\prime}(u_t - b_\tau^*) \frac{\mathbf{Z}_t}{h_t} \} v \\ &+ \frac{1}{2} \mathbf{u}' \{ \frac{1}{n} \sum_{t=s+1}^{n} (u_t^2 \rho_\tau^{m\prime\prime}(u_t - b_\tau^*) + 2u_t \rho_\tau^{m\prime}(u_t - b_\tau^*)) \frac{\mathbf{Z}_t \mathbf{Z}_t'}{h_t^2} \} \mathbf{u}(1 + o_p(1)). \end{aligned}$$

Denote by $a_1^m = E(u_t \rho_\tau^{m\prime}(u_t - b_\tau^*))$, $b_1^m = -E(u_t \rho_\tau^{m\prime\prime}(u_t - b_\tau^*))$, $b_2^m = E(u_t^2 \rho_\tau^{m\prime\prime}(u_t - b_\tau^*))$, and $c_2^m = E(\rho_\tau^{m\prime\prime}(u_t - b_\tau^*))$. Then, by the Chebyshev weak law of large number, we have

$$B_n^m \to_p B^m = \frac{1}{2} c_2^m v^2 + b_1^m \mathbf{u}' \boldsymbol{\mu} v + \frac{1}{2}(b_2^m + 2a_1^m) \mathbf{u}' \mathbf{G}_2 \mathbf{u}(1 + o_p(1)). \tag{C.2.11}$$

Using the Cramér-Wald device and the martingale CLT, we obtain that

$$q_{n,0}^m v + (\boldsymbol{\gamma}_n^m)' \mathbf{u} \to_d q_0^m v + (\boldsymbol{\gamma}^m)' \mathbf{u}, \tag{C.2.12}$$

where q_0^m and $\boldsymbol{\gamma}^m$ are normally distributed. Then, combination of (C.2.10)-(C.2.12) leads to

Statement (i): $S_n^m(v, \mathbf{u}) \to S^m(v, \mathbf{u})$ for $\forall$ m, where

$$\begin{aligned} S^m(v, \mathbf{u}) &\equiv q_0^m v + (\boldsymbol{\gamma}^m)' \mathbf{u} + \frac{1}{2} c_2^m v^2 + b_1^m \mathbf{u}' \boldsymbol{\mu} v \\ &+ \frac{1}{2}(b_2^m + 2a_1^m) \mathbf{u}' \mathbf{G}_2 \mathbf{u}(1 + o_p(1)). \end{aligned}$$

Denote by $q_{n,0} = -n^{-1/2} \sum_{t=s+1}^{n} \psi_\tau(u_t - b_\tau^*)$ and $\boldsymbol{\gamma}_n = -n^{-1/2} \sum_{t=s+1}^{n} u_t \psi_\tau(u_t - b_\tau^*) \frac{\mathbf{Z}_t}{h_t}$. Let γ be the limit variable of $\boldsymbol{\gamma}_n$ and q_0 be the limit variable of $q_{n,0}$ with $q_0 \sim \mathcal{N}(0, \tau(1-\tau))$. Then we have

Statement (ii): $S^m(v, \mathbf{u}) \to S(v, \mathbf{u})$as $m \to \infty$, where

$$S(v, \mathbf{u}) = q_0 v + \boldsymbol{\gamma}'\mathbf{u} + \frac{1}{2} f(b_\tau^*) v^2 + \frac{1}{2}(b_2 + 2a_1)\mathbf{u}'\mathbf{G}_2\mathbf{u} + b_1\mathbf{u}'\boldsymbol{\mu} v. \tag{C.2.13}$$

[The proof of the statement is stated briefly as follows. First, it is easy to show $q_0^m \to q_0$ and $\boldsymbol{\gamma}^m \to \boldsymbol{\gamma}$. Second, from Definition 6 (Lighthill, 1958) and (C.1.1)-(C.1.2), we have $a_1^m \to a_1$, $b_1^m \to b_1$, $b_2^m \to b_2$, $c_2^m \to E(\delta(u_t - b_\tau^*)) = f(b_\tau^*)$. Therefore, Statement (ii) holds.]

By Lemma C.1.1 and the above two statements, $S(v, \mathbf{u})$ is also the limit of $S_n(v, \mathbf{u})$. Therefore, we establish that the weak convergence of $S_n(v, \mathbf{u}) \to_d S(v, \mathbf{u})$ as generalized process (Phillips, 1996, p. 941).

Let $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{u}}$ be the minimizers of $S_n(\mathbf{u}, \mathbf{v})$ and $S(\mathbf{u}, \mathbf{v})$ for $\mathbf{u}$, respectively. Since $S(\mathbf{u}, \mathbf{v})$ is a quadratic form of $(\mathbf{u}', \mathbf{v}')'$, $\hat{\mathbf{u}}$ is unique. Simple algebra gives that

$$\hat{\mathbf{u}} = -\{(b_2 + 2a_1)\mathbf{G}_2 - \frac{b_1^2 \boldsymbol{\mu}\boldsymbol{\mu}'}{f(b_\tau^*)}\}^{-1}(\boldsymbol{\gamma} - \frac{b_1 \boldsymbol{\mu}}{f(b_\tau^*)} q_0),$$

where $(\boldsymbol{\gamma} - \frac{b_1 \boldsymbol{\mu}}{f(b_\tau^*)} q_0)$ is a random variable with mean nonzero.

Since the minimization operator is continuous under the infimum topology, by the continuous mapping theorem [see for example Theorem 25.7 in Billingsley (1995)],

$$\hat{\mathbf{u}}_n \xrightarrow{\mathcal{D}} \hat{\mathbf{u}}. \tag{C.2.14}$$

Note $\hat{\mathbf{u}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*)$. Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}^*) = -\{(b_2 + 2a_1)\mathbf{G}_2 - \frac{b_1^2 \boldsymbol{\mu}\boldsymbol{\mu}'}{f(b_\tau^*)}\}^{-1}(\boldsymbol{\gamma}_n - \frac{b_1 \boldsymbol{\mu}}{f(b_\tau^*)} q_{n,0}) + o_p(1).$$

Proof of Theorem 4.3.2. Let $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$, $\sqrt{n}(c_{\tau_k} - c_{\tau_k}^*) = v_k$, and $\mathbf{v} = (v_1, \ldots, v_K)'$. Put $l_0(\varepsilon_t, v_k, \mathbf{u}) = \log|\varepsilon_t| - \log(h_t(\boldsymbol{\beta}^* + n^{-1/2}u)) - (c_{\tau_k}^* + n^{-1/2}v_k)$

and $l_1(u_t, v_k, \mathbf{u}) = (\log(|u_t|) - c^*_{\tau_k}) - \frac{1}{\sqrt{n}}(\frac{\mathbf{Z}'_t}{h_t}\mathbf{u} + v_k)$. Then minimizing the objective in (4.3.6) is equivalent to minimizing

$$S_n(\mathbf{v}, \mathbf{u}) = \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \{\rho_{\tau_k}(l_0(\varepsilon_t, v_k, \mathbf{u}) - \rho_{\tau_k}(\log(|u_t|) - c^*_{\tau_k})\}.$$

Define

$$S^*_n(\mathbf{v}, \mathbf{u}) = \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \{\rho_{\tau_k}(l_1(u_t, v_k, \mathbf{u})) - \rho_{\tau_k}(\log|u_t| - c^*_{\tau_k})\},$$

$$S^{**}_n(\mathbf{v}, \mathbf{u}) = \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \{\rho_{\tau_k}(l_1(u_t, v_k, \mathbf{u}) - \frac{1}{n}\mathbf{u}'\mathbf{H}_t\mathbf{u}) - \rho_{\tau_k}(\log|u_t| - c^*_{\tau_k})\}.$$

We will show that S^*_n converges to a quadratic form of $\mathbf{u}$ and $\mathbf{v}$ through the following 3 steps:

Step (i). We first prove $S^*_n \longrightarrow_d S$ on $C(\mathbf{R}^{p+1})$.

By the identity(Knight 1998)

$$|r - s| - |r| = -s(I(r > 0) - I(r < 0) + 2\int_0^s [I(r \le x) - I(r \le 0)]dx,$$

we have

$$\rho_\tau(r - s) - \rho_\tau(r) = s(I(r < 0) - \tau) + \int_0^s [I(r \le x) - I(r \le 0)]dx.$$

Thus, we rewritten

$$S^*_n = \sum_{k=1}^{K} \omega_k q_{n,k} v_k + \mathbf{z}'_n\mathbf{u} + \sum_{k=1}^{K} \omega_k B^{(k)}_n, \tag{C.2.15}$$

where

$$q_{n,k} = \frac{1}{\sqrt{n}} \sum_{t=s+1}^{n} (I(\log(|u_t| < c^*_{\tau_k}) - \tau_k),$$

$$\begin{aligned} \mathbf{z}_n &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \frac{\mathbf{Z}_t}{h_t}(I(\log|u_t| < c^*_{\tau_k}) - \tau_k) \\ &= -n^{\frac{1}{2}} \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \frac{\mathbf{Z}_t}{h_t}\psi_\tau(\log|u_t| - c^*_{\tau_k}), \end{aligned}$$

$$B_n^{(k)} \equiv \sum_{t=s+1}^{n} \int_0^{\frac{\frac{\mathbf{z}_t'}{h_t}\mathbf{u}+v_k}{\sqrt{n}}} I(\log|u_t| \le c_{\tau_k}^* + x) - I(\log|u_t| \le c_{\tau_k}^*)dx.$$

Let $q_n = (q_{n,1}, \ldots, q_{n,K})'$. Using the Cramér-Wold device and the multivariate central limit theorem, we establish that $(\mathbf{q}_n', \mathbf{z}_n')' \xrightarrow{\mathcal{D}} (\mathbf{q}', \mathbf{z}')'$, and hence

$$\sum_{k=1}^{K} \omega_k q_{n,k} v_k + \mathbf{z}_n'\mathbf{u} \to_d \sum_{k=1}^{K} \omega_k q_k v_k + \mathbf{z}'\mathbf{u}, \tag{C.2.16}$$

where $\mathbf{q} = (q_1, \ldots, q_k)' \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \mathbf{A})$ and $\mathbf{z} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\mathbf{G}_2)$ with $(\mathbf{q}', \mathbf{z}')'$ being jointly normal with mean zero and covariance matrix $Cov(\mathbf{q}, \mathbf{z}) = \mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}'$.

By taking iterative expectation, we obtain that

$$E[B_n^{(k)}] \;=\; E\Big\{ \sum_{t=s+1}^{n} \int_0^{\frac{\frac{\mathbf{z}_t'}{h_t}\mathbf{u}+v_k}{\sqrt{n}}} [G(c_{\tau_k}^* + x) - G(c_{\tau_k}^*)]dx \Big\}.$$

Using the mean value theorem and Condition (b2), we arrive at

$$E[B_n^{(k)}] = \frac{1}{2} g(c_{\tau_k}^*)(\mathbf{u}'\mathbf{G}_2\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u} + v_k^2) + o_p(1).$$

Denote by $m_k(u_t, x) = I(\log|u_t| \le c_{\tau_k}^* + x) - I(\log|u_t| \le c_{\tau_k}^*)$. It is straightforward to show that

$$\begin{aligned}
Var(B_n^{(k)}) \;&=\; \sum_{t=s+1}^{n} E\{ \int_0^{\frac{\frac{\mathbf{z}_t'}{h_t}\mathbf{u}+v_k}{\sqrt{n}}} [m_k(u_t, x) - (G(c_{\tau_k}^* + x) - G(c_{\tau_k}^*))]dx \}^2 \\
&\le\; \sum_{t=s+1}^{n} E\{ | \int_0^{\frac{\frac{\mathbf{z}_t'}{h_t}\mathbf{u}+v_k}{\sqrt{n}}} [m_k(u_t, x) - (G(c_{\tau_k}^* + x) - G(c_{\tau_k}^*))]dx | \} \\
&\times\; 2|\frac{\frac{\mathbf{z}_t'}{h_t}\mathbf{u} + v_k}{\sqrt{n}}| \\
&\le\; 4E[B_n^{(k)}] \frac{\max_{s+1 \le t \le n} |\frac{\mathbf{z}_t'}{h_t}\mathbf{u} + v_k|}{\sqrt{n}} \\
&\to\; 0
\end{aligned}$$

from the condition $\sup_{\mathbf{u}} |n^{-\frac{1}{2}} \frac{\mathbf{Z}_t'}{h_t}\mathbf{u}| = o_p(1)$ which holds by the assumption (a0). Hence, by Chebychev's inequality, we have

$$B_n^{(k)} = \frac{1}{2} g(c^*_{\tau_k})(\mathbf{u}'\mathbf{G}_2\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u} + v_k^2) + o_p(1). \tag{C.2.17}$$

From (C.2.15)-(C.2.17), it follows that

$$S_n^* \rightarrow_d S = \sum_{k=1}^{K} \omega_k q_k v_k + \mathbf{z}^T\mathbf{u} + \frac{1}{2}\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k})(\mathbf{u}'\mathbf{G}_2\mathbf{u} + 2v_k\boldsymbol{\mu}'\mathbf{u} + v_k^2) + o_p(1). \tag{C.2.18}$$

Step (ii). Following the same argument as in Davis (1997), we can show that S_n^{**} has the same limit, that is, $S_n^{**} - S_n^* \longrightarrow_p 0$.

Step (iii). We show that S_n has the same limit as S_n^{**}. By Taylor's expansion, we can rewrite $S_n(\mathbf{u}, \mathbf{v})$ as

$$S_n(\mathbf{v}, \mathbf{u}) = \sum_{k=1}^{K} \omega_k \sum_{t=s+1}^{n} \{\rho_{\tau_k}(l_1(u_t, v_k, \mathbf{u}) - \frac{1}{n}\mathbf{u}'\mathbf{H}_t(\tilde{\boldsymbol{\beta}})\mathbf{u}) - \rho_{\tau_k}(\log|u_t| - c^*_{\tau_k})\},$$

where $\tilde{\boldsymbol{\beta}}$ is between $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^* + n^{-1/2}u$. Using the inequality

$$|\rho_\tau(r_1) - \rho_\tau(r_2)|/|r_1 - r_2| \leq \max(\tau, 1-\tau) < 1,$$

we establish that

$$\begin{aligned} |S_n - S_n^{**}| &\leq \sum_{t=s+1}^{n} \{\frac{1}{n}|\mathbf{u}^T(\mathbf{H}_t - \mathbf{H}_t(\tilde{\boldsymbol{\beta}}))\mathbf{u}|\} \\ &\rightarrow_p 0. \end{aligned}$$

Therefore, combining Steps (i)-(iii) leads to $S_n \longrightarrow_d S$. Let $\hat{\mathbf{u}}_n$ and $\hat{\mathbf{u}}$ be the minimizers of $S_n(\mathbf{u}, \mathbf{v})$ and $S(\mathbf{u}, \mathbf{v})$ for $\mathbf{u}$, respectively. Since $S(\mathbf{u}, \mathbf{v})$ is a quadratic form of $(\mathbf{u}', \mathbf{v}')'$, $\hat{\mathbf{u}}$ is unique. Simple algebra gives that

$$\hat{\mathbf{u}} = -[\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k})]^{-1}\boldsymbol{\Gamma}^{-1}(\mathbf{z} - \boldsymbol{\mu}(\sum_{k=1}^{K} \omega_k q_k)).$$

By the same arguments for (C.2.13), $\hat{\mathbf{u}}_n \xrightarrow{\mathcal{D}} \hat{\mathbf{u}}$. Note that $\hat{\mathbf{u}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*)$ and $(\sum_{k=1}^K \omega_k g(c^*_{\tau_k}))^{-1}(\mathbf{z} - \boldsymbol{\mu}(\sum_{k=1}^K \omega_k q_k))$ is normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\omega)\boldsymbol{\Gamma}$. Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}^*) \to_d \mathcal{N}\Big(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{\Gamma}^{-1}\Big).$$

Proof of Theorem 4.3.3. The proofs are basically the same as for Theorem 2.3.1 with $\hat{\boldsymbol{\omega}}$ replaced by $\boldsymbol{\omega}_{opt}(1 + o_p(1))$.

Proof of Theorem 4.3.4. Let $\sqrt{n}(c_{\tau_k} - c^*_{\tau_k}) = v_k$, $\mathbf{v} = (v'_1, \ldots, v_K)'$, $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$, $\sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*) = \boldsymbol{\delta}_n$. Put $l(v_k, \boldsymbol{\delta}_n, \mathbf{u}) = \log|\varepsilon_t(\boldsymbol{\alpha}^* + n^{-\frac{1}{2}}\boldsymbol{\delta}_n)| - \log(h_t(\boldsymbol{\alpha}^* + n^{-\frac{1}{2}}\boldsymbol{\delta}_n, \boldsymbol{\beta}^* + n^{-\frac{1}{2}}\mathbf{u})) - (c^*_{\tau_k} + n^{-1/2}v_k)$. Then minimizing the objective in (4.3.9) is equivalent to minimizing

$$S_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) \;=\; \sum_{k=1}^K \omega_k \sum_{t=s'+1}^n \{\rho_{\tau_k}(l(v_k, \boldsymbol{\delta}_n, \mathbf{u})) - \rho_{\tau_k}(\log|u_t| - c^*_{\tau_k})\}.$$

However, $S_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$, as a generalized process, is defined by the following regular sequence of process

$$S^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) \;=\; \sum_{k=1}^K \omega_k \sum_{t=s'+1}^n \{\rho^m_{\tau_k}(l(v_k, \boldsymbol{\delta}_n, \mathbf{u})) - \rho^m_{\tau_k}(\log|u_t| - c^*_{\tau_k})\}.$$

Denote by $\mathbf{H}_t = \begin{pmatrix} \mathbf{H}_{t11} & \mathbf{H}_{t12} \\ \mathbf{H}_{t21} & \mathbf{H}_{t22} \end{pmatrix}$, where $\mathbf{H}_t = -\frac{\partial^2 \log(h_t(\boldsymbol{\alpha}, \boldsymbol{\beta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}|_{\boldsymbol{\theta}^*}$ is a 2×2 block matrix and $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$. Let $\mathbf{D}_\varepsilon = E(\mathbf{D}_{\varepsilon t})$, $\mathbf{H} = E(\mathbf{H}_t) = \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{21} & \mathbf{H}_{22}, \end{pmatrix}$.

Taking Taylor's expansion for $S^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$, we obtain that

$$S^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) = A^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) + B^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) + C^m_n(\boldsymbol{\delta}_n, \mathbf{u}) + o_p(1), \qquad \text{(C.2.19)}$$

where

$$A^m_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) = -\frac{1}{\sqrt{n}} \sum_{k=1}^K \omega_k \sum_{t=s'+1}^n \rho^{m\prime}_{\tau_k}(\log(|u_t|) - c^*_{\tau_k})(\mathbf{J}'_t \boldsymbol{\delta}_n + \frac{\mathbf{Z}'_t}{h_t}\mathbf{u} + v_k),$$

$$B_n^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) = \frac{1}{2n}\sum_{k=1}^{K}\omega_k\sum_{t=s'+1}^{n}\rho_{\tau_k}^{m\prime\prime}(\log(|u_t|)-c_{\tau_k}^*)(v_k^2+\boldsymbol{\delta}_n'\mathbf{J}_t\mathbf{J}_t\boldsymbol{\delta}_n + \mathbf{u}'\frac{\mathbf{Z}_t\mathbf{Z}_t'}{h_t^2}\mathbf{u}+2v_k\mathbf{J}_t'\boldsymbol{\delta}_n+2v_k\frac{\mathbf{Z}_t'}{h_t}\mathbf{u}+2\boldsymbol{\delta}_n\mathbf{J}_t\frac{\mathbf{Z}_t'}{h_t}\mathbf{u}),$$

$$C_n^m(\boldsymbol{\delta}_n,\mathbf{u}) = \frac{1}{2n}\sum_{k=1}^{K}\omega_k\sum_{t=s'+1}^{n}\rho_{\tau_k}^{m\prime}(\log(|u_t|)-c_{\tau_k}^*)(\boldsymbol{\delta}_n'(\mathbf{D}_{\varepsilon t}+\mathbf{H}_{t11})\boldsymbol{\delta}_n + 2\boldsymbol{\delta}_n'\mathbf{H}_{t12}\mathbf{u}+\mathbf{u}'\mathbf{H}_{t22}\mathbf{u}),$$

and Lagrange remainder $o_p(1)$ holds from the assumption (c1) and the boundedness of $\rho_\tau^{m\prime\prime}(.)$ and $\rho_\tau^{m\prime\prime}(.)$.

Rewrite $A_n^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u})$ as

$$A_n^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) = (\mathbf{r}_n^m)'\boldsymbol{\delta}_n + (\mathbf{z}_n^m)'\mathbf{u} + \sum_{k=1}^{K}\omega_k q_{n,k}^m v_k.$$

Using the Cramer-Wald device and the martingale CLT, we have

$$A_n^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) \to_d A^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) = (\mathbf{r}^m)'\boldsymbol{\delta}_n + (\mathbf{z}^m)'\mathbf{u} + \sum_{k=1}^{K}\omega_k q_k^m v_k, \qquad \text{(C.2.20)}$$

where $\mathbf{r}^m \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{\mu}_r^m,\boldsymbol{\Sigma}_r^m)$, $\mathbf{z}^m \xrightarrow{\mathcal{D}} \mathcal{N}(\boldsymbol{\mu}_z^m,\boldsymbol{\Sigma}_z^m)$ and $q_k^m \xrightarrow{\mathcal{D}} \mathcal{N}(\mu_{q,k}^m,\sigma_k^m)$ with $\boldsymbol{\mu}_r^m = \lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{t=s'+1}^{n}\mathbf{J}_t\sum_{k=1}^{K}\omega_k\mu_k^m$, $\boldsymbol{\mu}_z^m = \lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{t=s'+1}^{n}\frac{\mathbf{Z}_t}{h_t}\sum_{k=1}^{K}\omega_k\mu_k^m$, $\mu_{q,k}^m = \lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{t=s'+1}^{n}\omega_k\mu_k^m$, $\sigma_k^m = \text{var}(\psi_{\tau_k}^m(\log(|u_t|)-c_{\tau_k}^*))$, $\boldsymbol{\Sigma}_r^m = \sigma^m(\boldsymbol{\omega})\boldsymbol{\Omega}$, $\boldsymbol{\Omega} = E(\mathbf{J}_t\mathbf{J}_t')$ and $\boldsymbol{\Sigma}_z^m = \sigma^m(\boldsymbol{\omega})\mathbf{G}_2$.

Denote by $\boldsymbol{\Psi} = E(\frac{\mathbf{Z}_t}{h_t}\mathbf{J}_t')$ and $\gamma_k^m = E(\rho_{\tau_k}^{m\prime\prime}(\log(|u_t|)-c_{\tau_k}^*))$. Applying the Chebyshev weak law of large number, we have

$$B_n^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) \to_p B^m(\mathbf{v},\boldsymbol{\delta}_n,\mathbf{u}) = \frac{1}{2}\sum_{k=1}^{K}\omega_k\gamma_k^m m(v_k,\boldsymbol{\delta}_n,\mathbf{u}) \qquad \text{(C.2.21)}$$

with

$$m(v_k,\boldsymbol{\delta}_n,\mathbf{u}) = [v_k^2+\boldsymbol{\delta}_n'\boldsymbol{\Omega}\boldsymbol{\delta}_n+\mathbf{u}'\mathbf{G}_2\mathbf{u}+2v_k\boldsymbol{\mu}_a'\boldsymbol{\delta}_n+2v_k\boldsymbol{\mu}'\mathbf{u}+2\boldsymbol{\delta}_n'\boldsymbol{\Psi}'\mathbf{u}]$$

and

$$C^m(\boldsymbol{\delta}_n,\mathbf{u}) \to_p C^m(\boldsymbol{\delta}_n,\mathbf{u}) = \frac{1}{2}\sum_{k=1}^{K}\omega_k\mu_k^m(\boldsymbol{\delta}_n'(\mathbf{E}+\mathbf{H}_{11})\boldsymbol{\delta}_n+\mathbf{u}_n'\mathbf{H}_{22}\mathbf{u}+2\boldsymbol{\delta}_n'\mathbf{H}_{12}\mathbf{u}). \qquad \text{(C.2.22)}$$

Combining (C.2.20)-(C.2.21), we establish that

$$S_n^m \to_d S^m(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) = A^m(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) + B^m(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) + C^m(\boldsymbol{\delta}_n, \mathbf{u}), \tag{C.2.23}$$

Let $\mathbf{q}^m = (q_1^m, \ldots, q_K^m)'$ and $\mathbf{q} = (q_1, \ldots, q_K)'$. Denote by

$$S(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) = \mathbf{r}'\boldsymbol{\delta}_n + \mathbf{z}'\mathbf{u} + \sum_{k=1}^{K} \omega_k q_k v_k + \frac{1}{2}\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k}) m(v_k, \boldsymbol{\delta}_n, \mathbf{u}), \tag{C.2.24}$$

where $\mathbf{q}^m \xrightarrow{\mathcal{D}} \mathbf{q}$, $(\mathbf{q}', \mathbf{r}', \mathbf{z}')'$ being jointly normal with $\text{cov}(\mathbf{q}, \mathbf{r}) = \mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}_a'$, $\text{cov}(\mathbf{q}, \mathbf{z}) = \mathbf{A}\boldsymbol{\omega}\boldsymbol{\mu}'$, $\text{cov}(\mathbf{r}, \mathbf{z}) = \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\text{cov}(\mathbf{J}_t, \frac{\mathbf{Z}_t}{h_t})$ and $\mathbf{r} \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \boldsymbol{\omega}'\mathbf{A}\boldsymbol{\omega}\boldsymbol{\Omega})$. We'll prove

$$S^m(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) \to S(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) \text{ as } m \to \infty. \tag{C.2.25}$$

Note that $\psi_{\tau_k}(u)$ can be defined by the sequence $\rho^{m\prime}_{\tau_k}(u)$ and $\delta(u)$ can be defined by the sequence $\rho^{m\prime\prime}_{\tau_k}(u)$. From (C.1.1)-(C.1.2), we have

$$\mu_k^m \to E(\psi_{\tau_k}(\log(|u_t|) - c^*_{\tau_k})) = 0 \tag{C.2.26}$$

and

$$\gamma_1^m \to E(\delta(\log(|u_t|) < c^*_{\tau_k})) = g(c^*_{\tau_k}) \tag{C.2.27}$$

as $m \to \infty$,

Also, we have $\sigma^m(\boldsymbol{\omega}) \to \sigma(\boldsymbol{\omega})$ from (C.1.8). Thus,

$$A^m(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u}) \to \mathbf{r}'\boldsymbol{\delta}_n + \mathbf{z}'\mathbf{u} + \sum_{k=1}^{K} \omega_k q_k v_k, \text{ as } m \to \infty. \tag{C.2.28}$$

It follows that (C.2.25) holds from (C.2.21)- (C.2.23) and (C.2.26)-(C.2.28). Therefore, by Lemma C.1.1, $S(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$ is also the limit of $S_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$.

Finally, by the same argument as for (C.2.14), the minimizer $\hat{\mathbf{u}}_n$ of $S_n(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$ converges in distribution to the minimizer $\hat{\mathbf{u}}$ of $S(\mathbf{v}, \boldsymbol{\delta}_n, \mathbf{u})$ such that

$$\hat{\mathbf{u}} = -\frac{\boldsymbol{\Gamma}^{-1}}{\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k})}(\mathbf{z} - \boldsymbol{\mu}\sum_{k=1}^{K} \omega_k q_k) - \boldsymbol{\Gamma}^{-1}(\boldsymbol{\Psi} - \boldsymbol{\mu}\boldsymbol{\mu}_a')\boldsymbol{\delta}_n,$$

where $\boldsymbol{\delta}_n = \sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*)$.

Denote by $\mathbf{C}' = \boldsymbol{\Psi} - \boldsymbol{\mu}\boldsymbol{\mu}_a'$ which equals to $\text{cov}(\frac{\mathbf{z}_t}{h_t}, \mathbf{J}_t)$ and $\mathbf{N} = (\sum_{k=1}^K \omega_k g(c^*_{\tau_k}))^{-1}\boldsymbol{\Gamma}^{-1}(\mathbf{z} - \boldsymbol{\mu}\sum_{k=1}^K \omega_k q_k)$ which is a normal random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2(\omega)\boldsymbol{\Gamma}^{-1}$. Then, we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{N} - \boldsymbol{\Gamma}^{-1}\mathbf{C}'\sqrt{n}(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*) + o_p(1).$$

Proof of Theorem 4.3.5. Let $\sqrt{n}(c_{\tau_k} - c^*_{\tau_k}) = v_k$, $\mathbf{v} = (v_1', v_2', \ldots, v_K')'$, $\sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \mathbf{u}$, and $\sqrt{n}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = \boldsymbol{\delta}$. Put

$$l(v_k, \boldsymbol{\delta}, \mathbf{u}) = \log|\varepsilon_t(\boldsymbol{\alpha}^* + n^{-\frac{1}{2}}\boldsymbol{\delta})| - \log(h_t(\boldsymbol{\alpha}^* + n^{-\frac{1}{2}}\boldsymbol{\delta}, \boldsymbol{\beta}^* + n^{-\frac{1}{2}}\mathbf{u})) - (c^*_{\tau_k} + n^{-1/2}v_k).$$

Then minimizing the objective in (4.3.10) is equivalent to minimizing

$$S_n(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) = \sum_{k=1}^K \omega_k \sum_{t=s'+1}^n \{\rho_\tau(l(v_k, \boldsymbol{\delta}, \mathbf{u}) - \rho_\tau(\log|u_t| - c^*_{\tau_k})\}.$$

Define

$$S_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) = \sum_{k=1}^K \omega_k \sum_{t=s'+1}^n \{\rho_\tau^m(l(v_k, \boldsymbol{\delta}, \mathbf{u}) - \rho_\tau^m(\log|u_t| - c^*_{\tau_k})\}.$$

Taking Taylor's expansion for $S_n^m(\mathbf{v}, \boldsymbol{\eta}, \mathbf{u})$, we obtain that

$$S_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) = (\mathbf{r}_n^m)'\boldsymbol{\delta} + (\mathbf{z}_n^m)'\mathbf{u} + \sum_{k=1}^K \omega_k q_k v_k + B_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) + C_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}),$$

where $B_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u})$ and $C_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u})$ are defined in (C.2.19). Then, similar to Theorem 4.3.4, we can show that

$$S_n^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) \to S^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) \quad \forall m,$$

and $S^m(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) \to S(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u})$, as $m \to \infty$, where $S(\mathbf{v}, \boldsymbol{\eta}, \mathbf{u})$ is defined in (C.2.24) with $\boldsymbol{\delta}_n$ being replaced of $\boldsymbol{\delta}$.

Therefore, by Lemma C.1.1, $S(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u})$ is also the limits of S_n, that is

$$S_n(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}) \to S(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u}). \tag{C.2.29}$$

Finally, since S_n is a convex function of $(\mathbf{v}', \boldsymbol{\delta}', \mathbf{u}')'$, the minimizers of $S(\mathbf{v}, \boldsymbol{\delta}, \mathbf{u})$ for $(\boldsymbol{\delta}', \mathbf{u})'$ are

$$\begin{pmatrix} \hat{\boldsymbol{\delta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \{\sum_{k=1}^{K} \omega_k g(c^*_{\tau_k})\}^{-1} \begin{pmatrix} \boldsymbol{\Pi} & \mathbf{C} \\ \mathbf{C}' & \boldsymbol{\Gamma} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \omega_k q_k \\ \mathbf{z} - \boldsymbol{\mu} \sum_{k=1}^{K} \omega_k q_k \end{pmatrix} + o_p(1).$$

Equivalently,

$$\begin{aligned} \hat{\boldsymbol{\delta}} &= -(\boldsymbol{\omega}'\mathbf{g})^{-1}[(\boldsymbol{\Pi}^{-1} + \boldsymbol{\Pi}^{-1}\mathbf{C}\mathbf{D}^{-1}\mathbf{C}'\boldsymbol{\Pi}^{-1})(\mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \omega_k q_k) \\ &\quad -\boldsymbol{\Pi}^{-1}\mathbf{C}\mathbf{D}^{-1}(\mathbf{z} - \boldsymbol{\mu} \sum_{k=1}^{K} \omega_k q_k)] \end{aligned}$$

and

$$\hat{\mathbf{u}} = -(\boldsymbol{\omega}'\mathbf{g})^{-1}[\mathbf{D}^{-1}(\mathbf{z} - \boldsymbol{\mu} \sum_{k=1}^{K} \omega_k q_k) - \mathbf{D}^{-1}\mathbf{C}'\boldsymbol{\Pi}^{-1}(\mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \omega_k q_k)],$$

where $(\boldsymbol{\omega}'\mathbf{g})^{-1}(\mathbf{r} - \boldsymbol{\mu}_a \sum_{k=1}^{K} \eta_k) \sim \mathcal{N}(\mathbf{0}, \sigma^2(\boldsymbol{\omega})\boldsymbol{\Pi})$. Therefore, with the same argument as (C.2.14), we have

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}^*) = \hat{\boldsymbol{\delta}} + o_p(1)$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_4 - \boldsymbol{\beta}^*) = \hat{\mathbf{u}} + o_p(1).$$

This completes the proof of the theorem.

Bibliography

[1] Bassett, G. W. and R. W. Koenker. (1992). A Note on Recent Proposals for Computing L_1 Estimates. *Computational Statistics & Data Analysis.* **14**, 207-211.

[2] Bera, A. K. and Higgins, M. L. (1993). On ARCH models: properties, estimation and testing. *Journal of Economic Survey.* **7**, 305-366.

[3] Bickel, P. J. (1975). One-step Huber estimates in linear models. *J. Amer. Statist. Assoc.* **70**, 428-433.

[4] Bickel, P.J. and Lehmann, E.L. (1976). Descriptive statistics for non-parametric models. III. Dispersion. *Ann. Statist.* **4**, 1139C1158.

[5] Bickel, P. J.(1978). Tests for heteroscedasticity, nonlinearity. *Annals of Statistics.* **6**, 266-291.

[6] Billingsley, P. (1995). *Probability and Measure*, 3rd ed. New York: Wiley.

[7] Bollerslev, T., Chou, R. and Kroner, K. (1992). ARCH modeling in finance. *Journal of Econometrics.* **50**, 5-59.

[8] Bollerslev, T., Engle, R. F. and Nelson, D. (1994). ARCH Models. Handbook of Econometrics, **4**, North-Holland.

[9] Breiman, L. (1995). Better subset regression using the non-negative garotte. *Technometrics.* **37**, 373-384.

[10] Carroll, R. J. and Ruppert. (1988). *Transformations and Weighting in Regression.* New York.

[11] Chaudhuri, P.,Doksum, K. and Samarov A. (1997) On average derivative quantile regression. *Annals of Statistics.* **25**,715-744

[12] Davis, R. A. and Dunsmuir, W. T. M. (1997). Least absolute deviation estimation for regression with ARMA errors. *J. Theor. Prob.* **10**, 481-497.

[13] Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, 2000, to the American Mathematical Society on Math Challenges of the 21st Century. Available at http://www.inma.ucl.ac.be/ francois/these/papers/entry-Donoho-2000.html.

[14] Dupačová, J. and Wets, R. (1988). Asymptotic behavior of statistical estimators and of optimal solutions of stochastic optimization problems. *Ann. Statist.* **16**, 1517-1549.

[15] Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the U.K. inflation. *Econometrica.*, **50**, 987-1008.

[16] Fan, J. and Jiang, J. (2002). Generalized likelihood ratio tests for additive models, tentatively accepted by *Jour. Amer. Statist. Assoc.*

[17] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

[18] Fan, J. and Li, R. (2006). Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III, European Mathematical Society, Zurich, 595-622.

[19] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Royal Statist. Soc.*, Ser. B. **70**, 849-911.

[21] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928-961.

[21] Fan, J. and Yao, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* New York: Springer-Verlag.

[22] Fu, W.J. (1998). Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphics Statistics.* **7**, 397-416.

[23] Hui, Y. V. and Jiang, J. (2005). Robust modelling of DTARCH models. *Econometrics Journal.* **8**, 143-158.

[24] Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *Ann. Statist.* **1**, 799-821.

[25] Huber, P. J. (1988). Robust regression: asymptotics, conjectures and monte carlo. *Ann. Statist.* **1**, 799-821.

[26] Jennrich, R. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.* **40**, 633-643.

[27] Jiang, J., Zhao, Q. and Hui, Y. V. (2001). Robust modlling of ARCH models. *Journal of Forecasting.* **20**, 111-133.

[28] Jurečková, J. and Procházka, B. (1994). Regression quantiles and trimmed least squares estimator in nonlinear regression model. *J. Nonparametr. Statist.* **3**, 201-222.

[29] Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *Ann. Statist.* **26**, 755-770.

[30] Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *Ann. Statist.* **28**, 1356-1378.

[31] Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

[32] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica.* **46**, 33-50.

[33] Koenker, R., Ng, P. and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika.* **81**, 673-680.

[34] Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics.* **71**, 265-283.

[35] Koenker, R. and Zhao, Q. (1996). Conditional quantile estimation and inference for ARCH models. *Econometric Theory.* **12**, 793-813.

[36] Kutner, M. H. et, al. (2005). *Applied Linear Statistical Models, 5th ed.* McGraw-Hill/Irwin.

[37] Lam, C. and Fan, J. (2008). Profile-Kernel likelihood inference with diverging number of parameters. *The Annals of Statistics.* **36**, 2232-2260.

[38] Li, C. W. and Li, W. K. (1996). On a double-threshold autoregressive heteroscedastic time series model. *J. Appl. Econometrics.* **11**, 253–274.

[39] Li, Y., Liu, Y. and Zhu, J. (2007). Quantile regression in reproducing kernel hilbert spaces. *J. Amer. Statist. Assoc.* **102**, 255-268.

[41] Li, Y. and Zhu, J. (2008). The l_1 norm quantile regression. *Journal of Computational and Graphical Statistics.* **17**, 163-185.

[41] Lighthill. (1958) Introduction To Fourier Analysis & Generalized functions.

[42] Oberhofer, W. (1982). The consistency of nonlinear regression minimizing the L_1-norm. *Ann. Statist.* **10**, 316-319.

[43] Osborne, M. R. and G. A.Watson. (1971). On an Algorithm for Discrete Nonlinear L1 Approximation. *Computer Journal.* **14**, 184-188.

[44] Peng, L. and Yao, Q. (2003). Least absolute deviation estimation for ARCH and GARCH models. *Biometrika.* **90**, 967-975.

[45] Phillips,p.C.B. (1991b). A shortcut to LAD estimator asymptotics. it Econometric Theory. **7**, 450-463.

[46] Phillips,p.C.B. (1995). Roubust nonstationary regression. it Econometric Theory. **11**, 912-951.

[47] Portnoy, S. (1984). Asymptotic behavior of M-estimatiors of p regression parameters when p^2/n is large. I. consistency. *Ann. Statist.* **12**, 1298-1309.

[48] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366.

[49] Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics.* **32**, 143-155.

[50] Powell, J. (1991). Estimation of Monotonic Regression Models Under Quantile Restriction, in: *Nonparametric and Semiparametric Methods in Economics and Statistics*, Barnett et al., eds., Cambridge University Press.

[51] Richard A. and Dean W. (2007) *Applied Multivariate Statistical Analysis.* Prentice Hall

[52] Tibshirani, R. J. (1996). Regression shrinkage and selection via lasso. *J. Royal Statist. Soc.*, Ser. B. **58**, 267-288.

[53] Tibshirani, R. J. (1997). The Lasso method for variable selection in the Cox model. *Statist. Medicine* **16**, 385-395.

[54] Tong, H. and Lim, K. S. (1980). Threshold autoregressive, limit cycles and cyclical data. *J. Royal Statist. Soc.*, Ser. B, **42**, 245-292.

[55] Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Jour. Amer. Statist. Assoc.*, **84**, 231–240.

[56] Vanderbei., Meketon. and Freedman. (1986). A Modification of Karmarkar's Linear Programming Algorithm. *Algorithmica.* **1**, 395-407.

[58] Wang, H. Li, G. and Jiang, G. (2007). Robust Regression Shrinkage and Consistent Variable Selection Through the LAD-Lasso. *Journal of Business & Economic Statistics.* **25**, 347-355.

[58] Wang, H. Li, R. and Tsai, C-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika.* **94**, 553-568.

[59] Wang, J. (1995). Asymptotic normality of L_1-estimators in nonlinear regression. *Journal of Multivariate Analysis.* **54**, 227-238.

[61] Wu, C. J. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9**, 501-513.

[61] Yuan, M. and Lin, Y. (2007). On the nonnegative Garrote estimator. *Jour. Roy. Statist. Soc.*, Ser. B. **69**, 143-161.

[62] Zou, H. (2006). The adaptive lasso and its oracle properties. *Jour. Amer. Statist. Assoc.* **101**, 1418-1429.

[63] Zou, H. and Yuan, M. (2008a). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **3**, 1108-1126.

[64] Zou, H. and Yuan, M. (2008b). Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics and Data Analysis.* **52**, 5296-5304.