The Pennsylvania State University
The Graduate School

EFFICIENT PARAMETER ESTIMATION METHODS USING QUANTILE REGRESSION IN HETEROSCEDASTIC MODELS

A Dissertation in
Statistics
by
Zhanxiong Xu

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2017

The dissertation of Zhanxiong Xu was reviewed and approved* by the following:

Zhibiao Zhao
Associate Professor of Statistics
Dissertation Advisor, Chair of Committee

Runze Li
Verne M. Willaman Professor of Statistics

Lingzhou Xue
Assistant Professor of Statistics

Tao Yao
Associate Professor of Industrial and Manufacturing Engineering

Aleksandra Slavkovic
Professor of Statistics
Associate Head for Graduate Studies

*Signatures are on file in the Graduate School.

Abstract

The quantile regression method, first introduced by Koenker and Bassett Jr. (1978), provides a comprehensive toolkit of performing statistical inference for a class of statistical models and has become an important surrogate for the conventional least squares method. Specifically, quantile regression offers several versatile approaches to produce highly efficient estimates, regardless whether the error distribution is homoscedastic or not.

This dissertation is concerned with developing some efficient estimation methods for both the regression parameter and the dispersion parameter under the parametric nonlinear heteroscedastic model. The proposed methods have their roots in quantile regression and rely heavily on large-sample properties of the estimates.

In Chapter 2, we estimate the parameters by solving the “double-weighted composite quantile regression (DWCQR)” optimization problem. We establish central limit theorems for both estimates, based on which we recommend an objective way of choosing the optimal weights for both the quantile losses and the heteroscedasticity. It is shown by theoretical calculation that the resulting estimates are typically more efficient than those obtained from other methods, and their asymptotic variances converge to the Cramér-Rao lower bounds as the number of quantile positions tends to infinity. An adaptive estimation procedure is reported at the end of this chapter.

The computational aspects of the DWCQR problem are discussed in Chapter 3. Although the DWCQR problem in general does not admit numerical solutions that are guaranteed to converge, we attempted to provide an algorithm that combines the MM algorithm (Hunter and Lange (2000)) and the linear programming. The proposed MMLP algorithm overall works well and successfully confirms the nice theoretical properties of the DWCQR estimates using the optimal weights. The Monte Carlo study demonstrates that the DWCQR method outperforms the conventional estimation methods for the models under investigation.

In Chapter 4, for simplicity, we restrict the regression function to be linear and consider an alternative efficient estimation approach, which is based on a

preliminary estimate $\hat{\alpha}_n$ of the dispersion parameter. We first derive the Bahadur representation of the regression quantile $\hat{\beta}(\tau)$ for fixed τ. It is then interesting to note that the effect of the $\hat{\alpha}_n$ propagates in the asymptotic representation of $\hat{\beta}(\tau)$. Such asymptotic bias brought by $\hat{\alpha}_n$ can be eliminated by averaging regression quantiles across different quantile positions with a set of carefully chosen weights. In the meantime, it can be shown that these weights can be simultaneously adjusted so that the resulting estimate is also asymptotically efficient. The chapter is concluded by Monte Carlo studies.

In the appendix, by surveying some classical examples and representative proofs in the quantile regression literature, we illustrate an important proof technique known as the "chaining arguments".

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Literature Review

The concept of quantile regression introduced in the seminal paper of Koenker and Bassett Jr. (1978), has become a widely used and accepted technique in many areas of theoretical and applied statistics and econometrics. The first monograph on this topic has been published by Koenker (2005), covering a wide scope of well established foundations and actual research frontiers. In this dissertation, by effectively employing the quantile regression tools, we will study the parameter estimation problem for nonlinear heteroscedastic models in depth.

The nonlinear heteroscedastic model (2.1) has been discussed systematically by Carroll and Ruppert (1988) using the least squares approach, for a similar treatment, see also Bates and Watts (1988). Welsh et al. (1994) applied regression quantiles (Koenker and Bassett Jr. (1978)) to estimate regression and dispersion parameters in nonlinear heteroscedastic models. If the regression part is restricted to be linear, there is more literature on statistical inference based on regression quantiles, see for example Koenker and Bassett Jr. (1982), Koenker and Zhao (1996), Zhou and Portnoy (1998). By contrast to the aforementioned papers for which the dispersion parts are modelled parametrically, Welsh (1996), Zhao (2001) considered the case that the dispersion part is left unspecified and estimated nonparametrically.

It is well-known that residual-based methods are employed extensively in regression diagnosis and as an intermediate step to obtain robustified estimates of regression parameters. For a comprehensive introduction to how to wisely summarize, transform and exploit residuals under the least squares context, see Bickel (1978) and Chapters 2, 3 of Carroll and Ruppert (1988). Ruppert and Carroll (1980) constructed trimmed least squares least estimator by discarding a small portion of observations whose residuals based on some preliminary estimate are extreme, in this way the robustness of the original least squares estimator gets improved. Under the general M-estimation setting, Bickel (1975) established large-sample theory for the one-step estimates based on residuals. Since the weights must be estimated in advance before applying any weighted regression procedure, it is natural that the residual-based methods would play a vital role in heteroscedastic models such as (2.1), among others we refer readers to Koenker and Zhao (1996), Zhou and Portnoy (1998) and Xiao and Koenker (2009), for which the preliminary estimates are obtained via regression quantiles.

One of the remarkable advantages of quantile-based regression method is that it allows us to incorporate information across different quantile positions. Based on regression quantiles and regression rank scores (Gutenbrunner and Jurečková (1992)), three broad classes of statistics, which are termed as linear rank statistics, the first type L-statistics and the second type L-statistics, can be derived. As manifested in Jurečková (1977) and Jurečková (1985), these classes of statistics are natural generalizations in linear models to their counterparts in location models. Among these statistics, of our primary interest is the L-statistics (of the first type) which takes the form

$$T_n^\nu = \int \hat{\beta}_n(\tau)\,\mathrm{d}\nu(\tau), \tag{1.1}$$

where $\hat{\beta}_n(\tau), \tau \in (0,1)$ are regression quantiles and ν is a finite signed measure on the open unit interval $(0,1)$ that has a compact support. In practice, this integral representation form usually reduces to a weighted average of some regression quantiles at finite distinctive quantile positions $\tau_1, \ldots, \tau_K$. Under the framework of quantile regression, the L-estimator (1.1) has been extensively studied since 1980s. For representative work, see Koenker and Portnoy (1987), Portnoy and Koenker (1989), Koenker and Zhao (1994), and Koenker and Xiao (2002). It is demonstrated in Zhao and Xiao (2014) that the discrete weighted average version of (1.1) is efficient by carefully choosing the weights under various settings. In this dissertation the similar idea will be further extended to the more general model (2.1).

Compared to L-estimators based on regression quantiles, the composite quantile regression (CQR) estimators seem to be less intuitive but could have better efficient performance than L-estimators. The possibility of estimation in the linear homoscedastic case CQR was studied by Koenker (1984) and attributed to Robert Hogg. A notable application of CQR to variable selection problem was conducted by Zou and Yuan (2008). Kai et al. (2010) applied the CQR technique to local polynomial regression and showed its attractive asymptotic efficiency property.

We finally remark the development on asymptotic theory related to L-estimator in quantile regression and nonlinear quantile regression. There are in general two routes to establish large-sample properties under the context of quantile regression: before the advent of the innovative work by Pollard (1991), the derivation of asymptotic normality of regression quantiles relies heavily on empirical process and stochastic equicontinuity argument (Bickel (1975)), see Ruppert and Carroll (1980), and Koenker and Portnoy (1987), among others. Such argument is mathematically technical and usually requires some sort of preliminary consistency result. By proving the elegant *convexity lemma*, Pollard (1989) circumvented the key difficulty in the asymptotic proof and made the arguments more accessible to many potential users. Knight (1998) elaborated this technique by introducing the celebrated *Knight's identity*. Recently, Hjort and Pollard (2011) reviewed this topic systematically and illustrated its applications to many other statistical problems. We believe that nowadays the second route has become the main workhorse in proving consistency and asymptotic normality in quantile regression literature, and the spirit of this route is adopted in our own proof in Chapter 2. Regarding the asymptotic theory in nonlinear quantile regression, Oberhofer (1982) considered the consistency and Wang (1995) derived the asymptotic normality of the least absolute deviations (LAD) estimator under the assumption of i.i.d. errors, respectively, see also Jurečková and Procházka (1994), Koenker (2005). He and Shao (1996) provided very general treatments for other classes of M-estimators under this context. Recently, Oberhofer and Haupt (2016) studied the asymptotic properties of the nonlinear quantile regression under general assumptions on the error process, which is allowed to be heterogeneous and mixing.

1.2 Contribution and Organization

The main contribution of this dissertation is to systematically study the CQR method in nonlinear heteroscedastic models. Nonlinear models and heteroscedastic models rarely appeared together in the quantile regression literature. Initially, the main goal of the dissertation is to propose some optimal weighting mechanism under the CQR setting, under the guidance of the derived large-sample behavior of the estimates. In the course of completing Chapter 2, we found that the numerical studies of CQR estimators could be considerably challenging. Therefore, on one hand, we have made some attempt on estimating the parameters by combining the MM algorithm (Hunter and Lange (2000)) and the classical linear programming technique, which forms the content of Chapter 3, whereas on the other hand, we resorted to the classical L-estimators to alleviate the computational difficulty, which results in another efficient estimation approach and the content of Chapter 4. In the quantile regression literature, it is well-known that the establishment of the Bahadur representation of the constructed estimates is standard and can be highly technical, to account for this topic, we conclude this dissertation by providing a comprehensive survey on some important prove technique, known as the chaining arguments, that deciphers the essential technicality in the process of building the Bahadur representation.

Chapter 2

Efficient Composite Quantile Regression for Nonlinear Heteroscedastic Models

2.1 Introduction

There is vast literature on parameter estimation and variable selection for the classical linear regression model $Y = X^T\beta + \varepsilon$. The classical linear regression imposes the restrictive assumption of homoscedastic errors, i.e., the error has the same distribution regardless of the covariates. From the modeling perspective, it is hard to justify that the error is independent of the covariates whereas the mean relationship strongly depends on the covariates. For example, Engle (1982) examined the United Kingdom inflation during the time period 1958-1977 and found that it is more reasonable to allow the conditional variances to vary over time, leading to the widely used autoregressive conditional heteroscedastic (ARCH) models. As shown in Section 2.3.1, ignoring such heteroscedasticity could result in substantial loss of efficiency.

In this chapter we study the following nonlinear heteroscedastic regression model

$$Y_i = g(X_i, \beta) + s(X_i, \alpha)\varepsilon_i, \quad i = 1, \dots, n. \tag{2.1}$$

where $(X_i, Y_i, \varepsilon_i)$ is the triplet of covariates, response, and error, and $g(X_i, \beta)$ and $s(X_i, \alpha) > 0$ are two known functions with unknown parameters $\beta \in \mathbb{R}^p$ and $\alpha \in$

$\mathbb{R}^q$. If $g(X_i, \beta) = X_i^T\beta$ and $s(X_i, \alpha) \equiv 1$, then (2.1) reduces to the classical homoscedastic linear model. Model (2.1) allows for heteroscedasticity in the error with the scale $s(X_i, \alpha) > 0$ depending on the covariates X_i.

Our goal is seeking for efficient estimation for the parameters (β, α).

The rest of this chapter is organized as follows. In Section 2.2 we present the double weighted composite quantile regression method and study its asymptotic properties. Following the asymptotic results established in Section 2.2, we propose an objective way to determine the optimal weights in Section 2.3 and discuss the theoretical possible efficiency gain by doing so. An adaptive estimation procedure with estimated optimal weights is reported in Section 2.4. Proofs are collected in Section 2.5.

We now introduce some notation. For matrix $A = (a_{ij})$, denote its Frobenius norm by $\|A\| = (\sum_{i,j} a_{ij}^2)^{1/2}$, which reduces to the Euclidean L^2 norm if A is a vector. For $d > 0$ and a random variable Z, we write $Z \in \mathcal{L}^d$ if $\mathbb{E}(|Z|^d) < \infty$. A^T stands for the transpose of A. The indicator function of a set A is denoted by $\mathbf{1}_{x\in A}$, or $\mathbf{1}(x \in A)$.

2.2 Double-weighted Composite Quantile Regression

For $\tau \in (0, 1)$, denote by $Q_{Y_i}(\tau|X_i)$ the conditional τ-th quantile of Y_i given X_i and by $Q_\varepsilon(\tau)$ the τ-th quantile of ε_i. Suppose ε_i is independent of X_i. Then

$$Q_{Y_i}(\tau|X_i) = g(X_i, \beta) + s(X_i, \alpha)Q_{\varepsilon_i}(\tau) \tag{2.2}$$

which can be used to construct quantile regression estimation (Koenker (2005)).

Denote by $\rho_\tau(z) = z(\tau - \mathbf{1}_{z<0})$ the quantile loss at a quantile position $\tau \in (0, 1)$. In particular, $\rho_{0.5}(z) = |z|/2$ is the median quantile loss or the least absolute deviation (LAD). If ε_i has median zero (i.e., $Q_\varepsilon(0.5) = 0$), then $Q_{Y_i}(0.5|X_i) = g(X_i, \beta)$ and we can estimate β by

$$\hat{\beta}_{\text{LAD}} = \arg\min_b \sum_{i=1}^n |Y_i - g(X_i, b)|$$

However, since $s(X_i, \alpha)Q_\varepsilon(0.5) = 0$ for all α, α is non-identifiable from this LAD regression, although it may be estimated through regression of residuals, which is the theme of Chapter 4.

Unlike the least squares regression that employs information from the conditional mean, quantile regression can provide a more informative picture of the distribution by specifying different quantile positions $\tau \in (0, 1)$. In (2.2), since (β, α) does not depend on τ, we can combine the information across quantiles to estimate (β, α). Throughout this paper we consider combining information across k uniformly spaced quantile positions $\tau_j = j/(k+1), j = 1, \ldots, k$.

From (2.2), $Q_{Y_i}(\tau_j|X_i) = g(X_i, \beta) + s(X_i, \alpha)Q_{\varepsilon_i}(\tau_j), j = 1, \ldots, k$. There are quantile-independent unknown parameters $(\beta^T, \alpha^T)^T$, and the unknown quantiles $Q_\varepsilon(\tau_1), \ldots, Q_\varepsilon(\tau_k)$. To estimate $\beta, \alpha, Q_\varepsilon(\tau_1), \ldots, Q_\varepsilon(\tau_k)$ jointly, we consider the weighted quantile loss

$$L(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{w_j}{h(X_i)} \rho_{\tau_j}\{Y_i - g(X_i, b) - s(X_i, a)q_j\}, \quad \theta = (b^T, a^T, q_1, \ldots, q_k)^T \tag{2.3}$$

where $w = (w_1, \ldots, w_k)^T$ with $w_j \geq 0$ are weights for the quantile loss ρ_{τ_j} and $h(X_i) > 0$ is a general weight function for the heteroscedasticity due to $s(X_i, \alpha)\varepsilon_i$. Since the proposed method incorporates two weights and multiple quantile losses, we call it *double-weighted composite quantile regression* (DWCQR).

For the homoscedastic model $Y = X^T\beta + \varepsilon$, Koenker (1984) studied parameter estimation via weighted composite quantile regression (WCQR), Zou and Yuan (2008) studied parameter estimation and variable selection using unweighted composite quantile regression (UCQR) with weights $w_1 = \cdots = w_k = 1$, Bradic et al. (2011) further extended their methodology to weighted case. In these works, the model has homoscedastic errors so that $h(X_i) = 1$. As shown in Section 2.3.1 below, in the presence of heteroscedastic errors, their methods could lead to substantial loss of efficiency.

To study the asymptotic behavior of our DWCQR estimator, we set up the following framework:

Condition 1. (i) For each $i \in \mathbb{Z}$, ε_i is independent of $X_i, \varepsilon_{i-1}, X_{i-1}, \varepsilon_{i-2}, \ldots$. (ii) $\{(X_i, \varepsilon_i)\}_{i\in\mathbb{Z}}$ is a strictly stationary ergodic process.

The dependence structure in Condition 1 offers a very flexible framework to study asymptotic theory. In particular, as shown in Section 2.5, Condition 1(i) allows us to construct appropriate martingales, which can greatly facilitate asymptotic theory.

We illustrate Condition 1 by several common examples:

Example 1 (i.i.d. data). If $\{(X_i, \varepsilon_i)\}_{i\in\mathbb{Z}}$ are i.i.d. and the error ε_i is independent of the covariate X_i, then Condition 1 holds.

Example 2 (Causal ergodic processes). Let $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ be i.i.d. random variables. In (2.1), if $X_i = (Y_{i-1}, \ldots, Y_{i-p})$, then we have the nonlinear ARCH model

$$Y_i = g(Y_{i-1}, \ldots, Y_{i-p}, \beta) + s(Y_{i-1}, \ldots, Y_{i-p}, \alpha)\varepsilon_i \tag{2.4}$$

This model includes many widely used time series models, including threshold autoregressive models, exponential autoregressive models, Engle's ARCH models. Under conditions in Wu (2005), (2.4) admits a unique stationary solution $Y_i = G(\varepsilon_i, \varepsilon_{i-1}, \ldots)$ for some function G. Thus $X_i = (Y_{i-1}, \ldots, Y_{i-p})$ is a function of $\varepsilon_{i-1}, \varepsilon_{i-2}, \ldots$, and Condition 1 holds.

Example 3 (Mixing processes). To facilitate asymptotic theory, a popular condition is the strong mixing condition with mixing coefficients $\{\alpha_j\}_{j\in\mathbb{N}}$ satisfying $\sum_{j=1}^{\infty} \alpha_j^{1-2\delta} < \infty$ for some $\delta > 2$ (Fan and Yao (2003)). It is well known that strong mixing processes are ergodic so that Condition 1(ii) holds. Due to the martingale structure of Condition 1(i), we do not require the usual condition $\sum_{j=1}^{\infty} \alpha_j^{1-2\delta} < \infty$. It is worth pointing out many non-mixing processes also satisfy Condition 1. For the autoregressive model $Y_i = \rho Y_{i-1} + \varepsilon_i$ with $\rho \in (0, 1/2]$ and ε_i being Bernoulli variables $\mathbb{P}(\varepsilon_i = 1) = 1 - \mathbb{P}(\varepsilon_i = 0) = q \in (0, 1)$, the stationary solution is not strong mixing (Andrews (1984)). By contrast, Condition 1 holds.

Denote by $\dot{g}(X_i, \beta) = \partial g(X_i, \beta)/\partial\beta$ and $\ddot{g}(X_i, \beta) = \partial^2 g(X_i, \beta)/\partial\beta^2$ the gradient vector and Hessian matrix of $g(X_i, \beta)$ with respect to β. Similarly, define $\dot{s}(X_i, \alpha)$ and $\ddot{s}(X_i, \alpha)$.

Condition 2. Let X, ε be distributed as X_i, ε_i. Denote by $F_\varepsilon, f_\varepsilon$ the distribution function and density function of ε respectively. (i) $\sup_u[f_\varepsilon(u) + |f'_\varepsilon(u)|] < \infty$ and $f_\varepsilon(u) > 0$ on $\{u : 0 < F_\varepsilon(u) < 1\}$. (ii) $s(X, \alpha) > c$ and $h(X) > c$ for some constant $c > 0$, $\mathbb{E}\left[\frac{s(X,\alpha)}{h(X)}\right] < \infty$. (iii) For $D = (\dot{g}(X, \beta)^T, \dot{s}(X, \alpha)^T, s(X, \alpha))^T$,

$\mathbb{E}(DD^T)$ is positive definite. (iv) $w_1, \ldots, w_k > 0$ are strictly positive. (v) For some constants $\epsilon > 0$ and $\delta > 2$,

$$\begin{aligned} V_i \coloneqq & \sup_{\|\vartheta_\beta\|<\epsilon} [\|\dot{g}(X_i, \beta + \vartheta_\beta)\| + \|\ddot{g}(X_i, \beta + \vartheta_\beta)\|] \\ & + \sup_{\|\vartheta_\alpha\|<\epsilon} [\|s(X_i, \alpha + \vartheta_\alpha)\| + \|\dot{s}(X_i, \alpha + \vartheta_\alpha)\| + \|\ddot{s}(X_i, \alpha + \vartheta_\alpha)\|] \in \mathcal{L}^\delta \end{aligned} \tag{2.5}$$

We briefly comment on Condition 2. (i) is a common assumption in quantile regression. (ii) is imposed for technical convenience and can be weakened. (iii) rules out singular design matrix (see Lemma 2.4). (iv) is necessary to avoid an ill-posed problem. Otherwise, if $w_j = 0$ for some j, then q_j in (2.3) can take arbitrary value; in that case, we can remove the quantile position τ_j from (2.3) and use the remaining quantiles with positive weights in the analysis. (v) is used to control the remainder terms in Taylor's expansions.

Theorem 2.1. *Let the true value $\theta_0 = (\beta^T, \alpha^T, Q_\varepsilon(\tau_1), \ldots, Q_\varepsilon(\tau_k))^T$. Under Conditions 1–2, $L(\theta)$ in (2.3) has a local minimizer $\hat{\theta} = (\hat{\beta}^T, \hat{\alpha}^T, \hat{q}_1, \ldots, \hat{q}_k)^T$ such that $\|\hat{\theta} - \theta_0\| = O_P(n^{-1/2})$. Define $e_j = (0, \ldots, 1, \ldots, 0)^T \in \mathbb{R}^k$ with 1 on the j-th element and 0 elsewhere, and*

$$M_j = \begin{pmatrix} \dot{g}(X, \beta) \\ Q_\varepsilon(\tau_j)\dot{s}(X, \alpha) \\ e_j s(X, \alpha) \end{pmatrix} \in \mathbb{R}^{p+q+k}, \quad j = 1, \ldots, k.$$

Then the asymptotic normality holds:

$$\sqrt{n}\left[\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \beta \\ \alpha \end{pmatrix}\right] \Rightarrow \mathcal{N}(0, \Sigma(w, h)), \quad \Sigma(w, h) = \{M^{-1}\Omega M^{-1}\}_{(p+q)\times(p+q)} \tag{2.6}$$

where the subscript means the $(p + q) \times (p + q)$ leading submatrix of $M^{-1}\Omega M^{-1}$, with

$$M = \sum_{j=1}^{k} w_j f_\varepsilon(Q_\varepsilon(\tau_j))\mathbb{E}\left[\frac{M_j M_j^T}{h(X)s(X, \alpha)}\right] \tag{2.7}$$

$$\Omega = \sum_{j=1}^{k}\sum_{j'=1}^{k} w_j w_{j'}[\min(\tau_j, \tau_{j'}) - \tau_j\tau_{j'}]\mathbb{E}\left[\frac{M_j M_{j'}^T}{h(X)^2}\right] \tag{2.8}$$

By Theorem 2.1, with different choices of the weights $w_1, \ldots, w_k$ and $h(X_1), \ldots, h(X_n)$ in (2.3), the resulting estimator has different limiting covariance matrices $\Sigma(w, h)$ in (2.6). In Section 2.3 we study optimal choice of w and h to make $\Sigma(w, h)$ as “small" as possible.

2.3 Optimal DWCQR

Definition 2.1. For two matrices M_1 and M_2, we write $M_1 \geq M_2$ if $M_1 - M_2$ is non-negative definite. Let $K(z)$ be a matrix-valued function. We say that $K(z)$ attains the minimum at z^* , denoted by $z^* = \arg\min K(z)$, if $K(z) \geq K(z^*)$ for all z.

Minimum in Definition 2.1 also implies minimum in trace or determinant. However, some matrix-valued functions (e.g., the diagonal matrix $diag(z^2, (z-1)^2)$) may not have a minimum. Due to its complicated structure, it is generally infeasible to study the problem $\min_{w,h} \Sigma(w, h)$. Here we shall solve this problem under the symmetric density assumption.

Condition 3. The density function $f_\varepsilon(u)$ of ε_i is symmetric, i.e., $f_\varepsilon(u) = f_\varepsilon(-u)$, and we use the symmetric weights $w_j = w_{k+1-j}$, $j = 1, \ldots, k$.

Theorem 2.2. *Assume Conditions 1–3. Then, for the local minimizer in Theorem 2.1,*

(i) $\hat\beta$ *satisfies CLT:* $\sqrt{n}(\hat\beta - \beta) \Rightarrow \mathcal{N}(0, I_\beta(w)J_\beta(h))$, *where* $J_\beta(h) = M_\beta^{-1}\Omega_\beta M_\beta^{-1}$,

$$I_\beta(w) = \frac{\sum_{j=1}^{k}\sum_{j'=1}^{k} w_j w_{j'}[\min(\tau_j, \tau_{j'}) - \tau_j\tau_{j'}]}{[\sum_{j=1}^{k} w_j f_\varepsilon(Q_\varepsilon(\tau_j))]^2} \tag{2.9}$$

$$M_\beta = \mathbb{E}\left[\frac{\dot g(X,\beta)\dot g(X,\beta)^T}{h(X)s(X,\alpha)}\right] - \frac{\mathbb{E}[\dot g(X,\beta)/h(X)]\mathbb{E}[\dot g(X,\beta)^T/h(X)]}{\mathbb{E}[s(X,\alpha)/h(X)]}$$

$$\Omega_\beta = \mathbb{E}(U_\beta U_\beta^T) \text{ with } U_\beta = \frac{\dot g(X,\beta)}{h(X)} - \frac{s(X,\alpha)}{h(X)}\frac{\mathbb{E}[\dot g(X,\beta)/h(X)]}{\mathbb{E}[s(X,\alpha)/h(X)]}$$

(ii) $\hat{\alpha}$ satisfies CLT: $\sqrt{n}(\hat{\alpha}-\alpha) \Rightarrow \mathcal{N}(0, I_\alpha(w) J_\alpha(h))$, where $J_\alpha(h) = M_\alpha^{-1}\Omega_\alpha M_\alpha^{-1}$,

$$I_\alpha(w) = \frac{\sum_{j=1}^k \sum_{j'=1}^k w_j w_{j'} Q_\varepsilon(\tau_j) Q_\varepsilon(\tau_{j'})[\min(\tau_j, \tau_{j'}) - \tau_j \tau_{j'}]}{[\sum_{j=1}^k w_j Q_\varepsilon(\tau_j)^2 f_\varepsilon(Q_\varepsilon(\tau_j))]^2} \tag{2.10}$$

$$M_\alpha = \mathbb{E}\left[\frac{\dot{s}(X,\alpha)\dot{s}(X,\alpha)^T}{h(X)s(X,\alpha)}\right] - \frac{\mathbb{E}[\dot{s}(X,\alpha)/h(X)]\mathbb{E}[\dot{s}(X,\alpha)^T/h(X)]}{\mathbb{E}[s(X,\alpha)/h(X)]}$$

$$\Omega_\alpha = \mathbb{E}(U_\alpha U_\alpha^T) \textit{ with } U_\alpha = \frac{\dot{s}(X,\alpha)}{h(X)} - \frac{s(X,\alpha)}{h(X)} \frac{\mathbb{E}[\dot{s}(X,\alpha)/h(X)]}{\mathbb{E}[s(X,\alpha)/h(X)]}$$

By Theorem 2.2, the asymptotic covariance matrix of $\hat{\beta}$ has two independent components: the $I_\beta(w)$ that depends on the quantile weight w and the $J_\beta(h)$ that depends on the heteroscedasticity weight $h(X_i)$. To achieve optimal performance, we can choose h to minimize (Definition 2.1) the matrix $J_\beta(h)$ and choose w to minimize $I_\beta(w)$. Similarly, the most efficient $\hat{\alpha}$ is obtained by choosing (w, h) minimizing $J_\alpha(h)$ and $I_\alpha(w)$. See Sections 2.3.1 – 2.3.3.

2.3.1 Optimal choice of $h(X_i)$

Theorem 2.3 concerns with the optimal choice of the heteroscedasticity weight function h.

Theorem 2.3. *Recall $J_\beta(h)$ and $J_\alpha(h)$ in Theorem 2.2. Then, in the sense of Definition 2.1,*

$$\arg\min_h J_\beta(h) = \arg\min_h J_\alpha(h) = s(\cdot, \alpha)$$

That is, both $J_\beta(h)$ and $J_\alpha(h)$ are minimized at $h(X) = s(X,\alpha)$. Furthermore,

$$J_\beta(s(\cdot,\alpha)) = \left\{\mathbb{E}\left[\frac{\dot{g}(X,\beta)\dot{g}(X,\beta)^T}{s(X,\alpha)^2}\right] - \mathbb{E}\left[\frac{\dot{g}(X,\beta)}{s(X,\alpha)}\right]\mathbb{E}\left[\frac{\dot{g}(X,\beta)^T}{s(X,\alpha)}\right]\right\}^{-1}$$

$$J_\alpha(s(\cdot,\alpha)) = \left\{\mathbb{E}\left[\frac{\dot{s}(X,\alpha)\dot{s}(X,\alpha)^T}{s(X,\alpha)^2}\right] - \mathbb{E}\left[\frac{\dot{s}(X,\alpha)}{s(X,\alpha)}\right]\mathbb{E}\left[\frac{\dot{s}(X,\alpha)^T}{s(X,\alpha)}\right]\right\}^{-1}$$

Theorem 2.3 agrees with our intuition that the optimal choice of $h(X_i)$ is exactly the heteroscedasticity function $s(X_i,\alpha)$. Intuitively, using the denominator $h(X_i) = s(X_i,\alpha)$ in (2.3) can completely remove the heteroscedasticity of errors $s(X_i,\alpha)\varepsilon_i$, leading to optimal efficiency. If we extend the composite quantile regression method in Koenker (1984), Zou and Yuan (2008), and Bradic et al. (2011) in the simple linear regression setting $Y = X^T\beta + \varepsilon$ to the nonlinear heteroscedastic model (2.1) but without using the weight $h(X_i)$, we have the single (as opposed to double in (2.3)) weighted composite quantile regression

$$(\hat{\beta}_{\mathrm{WCQR}}, \hat{\alpha}_{\mathrm{WCQR}}, \hat{q}_1, \ldots, \hat{q}_k) = \underset{b,a,q_1,\ldots,q_k}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{k} w_j \rho_{\tau_j}\{Y_i - g(X_i, b) - s(X_i, a)q_j\}$$

Under Conditions 1–3, $\hat{\beta}_{\mathrm{WCQR}}$ satisfies the asymptotic normality in Theorem 2.2(i) with $J_\beta(h)$ therein replaced by $J_\beta(1)$, where 1 stands for the constant function $h(X) \equiv 1$. By Theorem 2.3, $J_\beta(1) \geq J_\beta(s(\cdot, \alpha))$. The same assertion holds for $\hat{\alpha}_{\mathrm{WCQR}}$.

To compare the efficiency of the optimal heteroscedasticity weight $h(X) = s(X, \alpha)$ to that of the unweighted case $h(X) \equiv 1$, we consider univariate X with $\mathbb{E}(X) = 0$, and linear function $g(X, \beta) = \beta X$. Then (the analysis for $J_\alpha(h)$ is similar)

$$J_\beta(1) = \frac{\mathbb{E}(X^2)}{\{\mathbb{E}[X^2/s(X,\alpha)]\}^2} \text{ and } J_\beta(s(\cdot,\alpha)) = \frac{1}{\mathbb{E}[X^2/s^2(X,\alpha)] - \{\mathbb{E}[X/s(X,\alpha)]\}^2}$$

Define the relative efficiency gain (REG) of $J_\beta(s(\cdot, \alpha))$ over $J_\beta(1)$ as

$$\mathrm{REG} = 1 - J_\beta(s(\cdot,\alpha))/J_\beta(1) \tag{2.11}$$

By Theorem 2.3, $0 \leq \mathrm{REG} \leq 1$. Figure 2.1 plots REG for $s(X, \alpha) = \exp(\alpha X)$ for $\alpha \in [0, 1]$ and different distributions of X. The distribution of X is scaled properly so that $\mathrm{Var}(X) = 1$. It reveals that: (i) There could be potentially substantial efficiency gain; (ii) The efficiency gain depends on the distribution of X; and (iii) As α increases, $s(\cdot, \alpha)$ becomes further away from the constant function, and the efficiency gain increases quickly.

2.3.2 Optimial choice of w for $I_\beta(w)$ and asymptotic efficiency

Recall $I_\beta(w)$ in (2.9) of Theorem 2.2(i). The simplest choice of w is the equal weight $w = u_k := (1, \ldots, 1)^T$. For simple homoscedastic linear model $Y = X^T\beta + \varepsilon$,

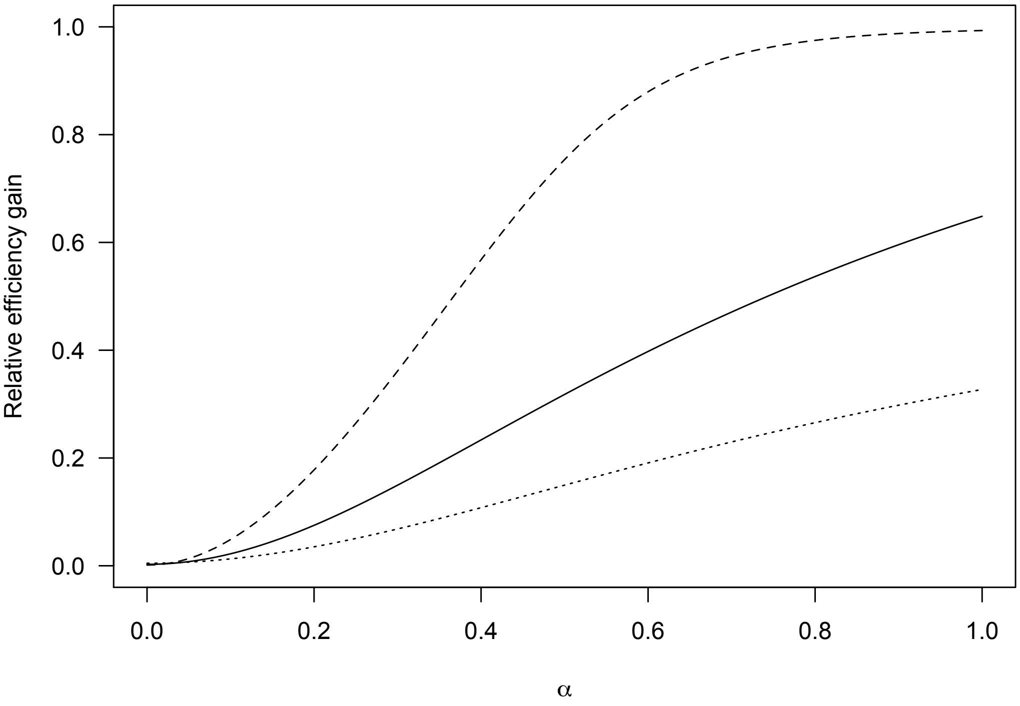

Figure 2.1. *Relative efficiency gain (REG)* [*Equation* (2.11)] *as a function of* α *in* $s(X,\alpha) = \exp(\alpha X)$*. Solid, dotted, and dashed curves correspond to standard normal* $\mathcal{N}(0,1)$*, uniform distribution on* $[-\sqrt{3},\sqrt{3}]$*, and Laplace distribution with density* $\exp(-\sqrt{2}|x|)/\sqrt{2}, x \in \mathbb{R}$*, respectively, for the covariate* X*.*

Zou and Yuan (2008) used this unweighted composite quantile regression (UCQR) approach. Then

$$I_\beta(u_k) = \frac{\sum_{j=1}^{k}\sum_{j'=1}^{k}[\min(\tau_j,\tau_{j'}) - \tau_j\tau_{j'}]}{[\sum_{j=1}^{k} f_\varepsilon(Q_\varepsilon(\tau_j))]^2} \tag{2.12}$$

Naturally, a more efficient approach is possible by choosing the optimal weight via minimizing $I_\beta(w)$:

$$w_\beta^* = \arg\min_w I_\beta(w), \quad \text{subject to } w_1,\ldots,w_k \geq 0, w_j = w_{k+1-j}, j = 1,2,\ldots,\lfloor k/2 \rfloor. \tag{2.13}$$

Due to the non-negative and symmetric weight constraints, there is generally no closed form solution for w_β^*, which makes it difficult to study w_β^* and $I_\beta(w_\beta^*)$. To understand the best possible performance of the proposed estimator, we proceed in two steps:

(I) First, we study the *unconstrained* minimization problem $\min_w I_\beta(w)$, see Theorem 2.4.

(II) Second, we study conditions to ensure the non-negative and symmetric constraints, see Theorem 2.5.

Condition 4. We say that a function $\ell(\tau) : (0,1) \to \mathbb{R}$ satisfies the efficiency regularity condition if $[\ell^2(\tau) + \ell^2(1-\tau)]/\tau + \tau^2 \int_\tau^{1-\tau} |\ell''(t)|^2 \,\mathrm{d}t \to 0$ as $\tau \to 0$.

Theorem 2.4. *For $I_\beta(w)$ in* (2.9)*, consider the unconstrained minimization problem $\min_w I_\beta(w)$.*

(i) The minimizer is $w_\beta^\circ = c\Gamma^{-1}\lambda$ for any constant $c > 0$, where

$$\Gamma = [\min(\tau_j, \tau_{j'}) - \tau_j \tau_{j'}] \in \mathbb{R}^{k\times k}, \quad \lambda = (f_\varepsilon(Q_\varepsilon(\tau_1)), \dots, f_\varepsilon(Q_\varepsilon(\tau_k)))^T \quad (2.14)$$

(ii) Let $\rho(\tau) = f_\varepsilon(Q_\varepsilon(\tau))$. If $\rho(\tau)$ satisfies the efficiency regularity Condition 4, then

$$\lim_{k\to\infty} I_\beta(w_\beta^\circ) = \frac{1}{\mathcal{F}(f_\varepsilon)}, \quad \textit{where } \mathcal{F}(f_\varepsilon) = \int_{\mathbb{R}} \frac{[f_\varepsilon'(u)]^2}{f_\varepsilon(u)} \,\mathrm{d}u. \quad (2.15)$$

Note that $\mathcal{F}(f_\varepsilon)$ in (2.15) is the Fisher information of f_ε. By Theorem 2.4(ii), as the number of quantiles $k \to \infty$, the *unconstrained* minimum $I_\beta(w^\circ)$ converges to the inverse of Fisher information of f_ε, which is the maximum likelihood estimation efficiency.

Under the assumption that f_ε is symmetric, it can be easily verified that w_β° is symmetric (see (2.57)). It remains to discuss under which condition that w_β° is also non-negative. For homoscedastic linear model $Y = X^T\beta + \varepsilon$, Koenker (1984) obtained the same optimal weight but failed to address the essential issue of non-negative constraint. Theorem 2.5 below presents a characterization of $f_\varepsilon(u)$ for w_β° to satisfy the non-negative constraint.

Theorem 2.5. *A sufficient condition for the unconstrained minimizer w_β° in Theorem 2.4(i) to satisfy the non-negative constraint is that* $\log f_\varepsilon(u)$ *is concave. Conversely, if w_β° is non-negative for every $k \in \mathbb{N}$, then* $\log f_\varepsilon(u)$ *is concave.*

From Theorems 2.4, 2.5, for densities whose logarithm are concave, the optimal DWCQR estimator of β can asymptotically achieve the optimal efficiency. The log-concave density assumption is satisfied for, e.g., standard normal density, logistic density $e^{-u}/(1+e^{-u})^2$, Laplace density $\exp(-|u|/2)$, and the Gumbel (extreme-value) density $\exp(-u-\exp(-u))$, etc. For these log-concave densities, the constrained minimizer has the closed form $w_\beta^* = w_\beta^\circ = c\Gamma^{-1}\lambda$ and $I_\beta(w_\beta^*)$ asymptotically attains the inverse Fisher information bound. On the other hand, Theorem 2.5 asserts that the log-concavity is also a necessary condition, and thus $1/\mathcal{F}(f_\varepsilon)$ bound in Theorem 2.4(ii) is not attainable if this assumption is violated; examples include Student's t and normal mixture distributions.

Without log-concavity, (2.13) can be solved numerically. Table 2.1 tabulates $I_\beta(w_\beta^*)$ for some commonly used densities (both log-concave and non-log-concave) for different k. To appreciate the results in Theorems 2.4, 2.5, we include the Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$ in the last column of Table 2.1. For LS regression, the asymptotic variance is proportional to $\mathrm{Var}(\varepsilon)$. For LAD regression, the asymptotic variance is proportional to $1/[4f_\varepsilon^2(Q_\varepsilon(0.5))]$. For comparison purpose, we also include them in Table 2.1. Furthermore, we include the UCQR case $I_\beta(u_k)$ (cf. Equation (2.12)) in the column UCQR.

Table 2.1 reveals some interesting phenomena. First, for all the eight densities considered, $I_\beta(w_\beta^*)$ stabilizes quickly at some limit. However, the limits are different for the two categories of densities. For the four log-concave densities, the limit is $1/\mathcal{F}(f_\varepsilon)$. For the four non-log-concave densities, the limit is larger than $1/\mathcal{F}(f_\varepsilon)$ (for Student's t_2, they are close), but it is unclear how to derive an explicit expression; we leave this as an open problem. Second, for both categories of densities, $I_\beta(w_\beta^*)$ with $k = 9$ or even 5 performs almost as good as $I_\beta(w_\beta^*)$ with $k = 39$. Thus, in practice, we recommend using $k = 9$. Third, the proposed method can significantly outperform the LS, LAD and Zou and Yuan (2008)'s UCQR.

Table 2.1. *Efficiency comparison of LS* (Var(ε))*, LAD* ($1/[4f_\varepsilon^2(Q_\varepsilon(0.5))]$)*, and the proposed method with $k = 5, 9, 19, 29, 39$ quantiles. The column UCQR is $I_\beta(u_9)$ in* (2.12) *with $k = 9$. The last column is the Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$. Smaller number means better efficiency. The densities in the upper part are log-concave, while that in the lower part are not log-concave.*

				$I_\beta(w^*_\beta)$ with $k =$					$1/\mathcal{F}(f_\varepsilon)$
Distribution of ε	LS	LAD	UCQR	5	9	19	29	39	
$\mathcal{N}(0, 1)$	1.00	1.57	1.07	1.09	1.04	1.02	1.01	1.01	1.00
logistic	3.29	4.00	3.03	3.09	3.03	3.01	3.00	3.00	3.00
Laplace	2.00	1.00	1.32	1.00	1.00	1.00	1.00	1.00	1.00
Gumbel	1.64	2.08	1.37	1.22	1.12	1.06	1.04	1.03	1.00
Student's t_1	∞	2.47	3.26	2.39	2.31	2.29	2.28	2.28	2.00
Student's t_2	∞	2.00	1.91	1.77	1.74	1.72	1.72	1.72	1.67
$0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1)$	5.00	85.76	4.27	2.52	2.23	2.09	2.05	2.03	1.38
$0.95\mathcal{N}(0, 1) + 0.05\mathcal{N}(0, 100)$	5.95	1.72	1.25	1.25	1.22	1.25	1.22	1.22	1.08

2.3.3 Optimal choice of w for $I_\alpha(w)$

Similar to (2.13), for optimal estimation efficiency of $\hat{\alpha}$, we can minimize $I_\alpha(w)$ (2.10):

$$w^*_\alpha = \arg\min_w I_\alpha(w), \quad \text{subject to } w_1, \ldots, w_k \geq 0, w_j = w_{k-j+1}, j = 1, \ldots, \lfloor k/2 \rfloor. \tag{2.16}$$

As in Section 2.3.2, we first consider the unconstrained minimization problem. This can shed some light on the best possible performance of the estimator. On the other hand, we can study conditions to ensure the non-negative constraint.

However, it is more complicated to study $I_\alpha(w)$ than $I_\beta(w)$. To see this, if $Q_\varepsilon(\tau^*) = 0$ for some τ^*, then (2.2) gives $Q_{Y_i}(\tau^*|X_i) = g(X_i, \beta)$, and thus τ^* does not contribute to the estimation of α and the corresponding weight has no effect. Under Condition 3 and Condition 2(i), $Q_\varepsilon(\tau) = 0$ if and only if $\tau = 0.5$. Thus, we need to distinguish whether k is odd or even, which determines whether $0.5 \in \{\tau_j = j/(k + 1), j = 1, \ldots, k\}$.

Theorem 2.6. *For $I_\alpha(w)$ in* (2.10)*, consider the unconstrained minimization problem $\min_w I_\alpha(w)$. Assume Condition 3 and Condition 2(i). Let $w^\circ_\alpha = \arg\min_w I_\alpha(w)$.*

(i) For even k ($Q_\varepsilon(\tau_j) \neq 0$ for all j), $w_\alpha^\circ = c\Xi^{-1}\zeta$ for any constant $c > 0$, and

$$\begin{aligned} \Xi &= \Big[Q_\varepsilon(\tau_j)Q_\varepsilon(\tau_{j'})[\min(\tau_j, \tau_{j'}) - \tau_j\tau_{j'}]\Big] \in \mathbb{R}^{k\times k} \\ \zeta &= (Q_\varepsilon(\tau_1)^2 f_\varepsilon(Q_\varepsilon(\tau_1)), \ldots, Q_\varepsilon(\tau_k)^2 f_\varepsilon(Q_\varepsilon(\tau_k)))^T \in \mathbb{R}^k \end{aligned} \tag{2.17}$$

For odd k, $w_\alpha^\circ = (w_{\alpha,1}^\circ, \ldots, w_{\alpha,k^*-1}^\circ, w_{k^*}, w_{\alpha,k^*+1}^\circ, \ldots, w_{\alpha,k}^\circ)^T$ with $k^* = (k+1)/2$, arbitrary w_k^*, and $(w_{\alpha,1}^\circ, \ldots, w_{\alpha,k^*-1}^\circ, w_{\alpha,k^*+1}^\circ, \ldots, w_{\alpha,k}^\circ)^T = c\Xi^{*-1}\zeta^*$. Here Ξ^{*-1} is formed by removing k^*-th row and k^*-th column of Ξ, and ζ^* is formed by removing k^*-th entry of ζ. For either case, the symmetric constraint in (2.16) *is automatically satisfied.*

(ii) If $s(\tau) \coloneqq Q_\varepsilon(\tau)f_\varepsilon(Q_\varepsilon(\tau))$ satisfies the efficiency regularity Condition 4, then

$$\lim_{k\to\infty} I_\alpha(w_\alpha^\circ) = \frac{1}{\mathcal{G}(f_\varepsilon)}, \quad \textit{where } \mathcal{G}(f_\varepsilon) = \int_{\mathbb{R}} \frac{[f_\varepsilon(u) + uf_\varepsilon'(u)]^2}{f_\varepsilon(u)}\,\mathrm{d}u$$

From Theorems 2.4 and 2.6, the optimal weights w_β^* and w_α^* for $\hat{\beta}$ and $\hat{\alpha}$ are different, and we can achieve different estimation efficiency for $\hat{\beta}$ and $\hat{\alpha}$. This can be interpreted by the quantile representation (2.2), where β does not depend on τ whereas $s(X_i, \alpha)$ has the coefficient $Q_\varepsilon(\tau)$ and thus introduces more dependence on the quantiles.

Similar to Theorem 2.5, we can derive necessary and sufficient condition for w_α° in Theorem 2.6(i) to satisfy the non-negative constraint. However, the condition is more complicated and has no clear interpretation, and thus we omit it. Theorem 2.6(ii) intends to demonstrate the optimal performance when ignoring the non-negative constraint. In practice, we can numerically solve the constrained problem (2.16).

2.3.4 Choice of w for estimating both β and α

From Sections 2.3.2 – 2.3.3, if we are interested in estimating β only, the optimal weight w_β^* in (2.13) leads to asymptotically efficient estimator of β for log-concave densities. Similarly, if we are interested in estimating α only, the weight w_α^* in (2.16) is the optimal choice of w.

In general, the weights w^*_β and w^*_α are different, and thus we cannot simultaneously achieve optimal efficiency for both β and α. If we are interested in estimating both β and α, we can run (2.3) twice, one using w^*_β to estimate β and the other using w^*_α to estimate α.

To estimate both β and α, an alternative approach is to run (2.3) with w of the form

$$w = \omega w^*_\beta + (1-\omega) w^*_\alpha, \quad w \in [0,1].$$

Different ω assigns different importance to β and α, and its choice is up to the goal of practitioners. For example, to achieve a balanced performance of β and α, we may use

$$\omega = \frac{\mathrm{trace}[J_\beta(h)]}{\mathrm{trace}[J_\beta(h)] + \mathrm{trace}[J_\alpha(h)]}$$

where $J_\beta(h)$ and $J_\alpha(h)$ are defined in Theorem 2.2. On the other hand, if we wish to put more emphasize on the accuracy of β than α, then we can use relatively large ω. As mentioned above, this approach leads to sub-optimal estimates of β and α. Thus, if computation is not a serious concern, we recommend the above two-separate-regression approach, which is taken in the Monte Carlo studies in Chapter 3.

2.4 Adaptive Estimation with Estimated Weights

By the discussion in Section 2.3.1, in the DWCQR (2.3), the theoretical optimal choice of $h(X_i)$ is $h(X_i) = s(X_i, \alpha)$ for estimating both β and α. Denote by $w^* = (w^*_1, \ldots, w^*_k)$ the optimal choice of $w = (w_1, \ldots, w_k)$. From Sections 2.3.2–2.3.3, under Conditions 1–3,

(i) To estimate β, we use $w^* = w^*_\beta$ determined by (2.13).

(ii) To estimate α, we use $w^* = w^*_\alpha$ determined by (2.16).

With the above choices of optimal weights, (2.3) becomes the following loss function

$$\bar{L}(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{w_j^*}{s(X_i,\alpha)}\rho_{\tau_j}\{Y_i - g(X_i,b) - s(X_i,a)q_j\}, \quad \theta = (b^T, a^T, q_1, \ldots, q_k)^T \tag{2.18}$$

By Sections 2.2 and 2.3, minimizing $\bar{L}(\theta)$ gives optimal estimates of β or α, depending on different choices of w^* in (i) or (ii) above. In practice w^* and $s(X,\alpha)$ are unknown and thus (2.3) is an infeasible function. Nevertheless, by minimizing the infeasible loss function (2.18), the resulting infeasible estimator can serve as a standard against which we can measure other estimators.

2.4.1 Adaptive estimation: A general theory

We say an estimator is adaptive if it can attain the same asymptotic efficiency of the infeasible estimator in (2.18); that is, it works as well as if the optimal weights were known.

Denote by $\hat{w}^*$ and $\hat{\alpha}^0$ some consistent estimates of w^* and α, respectively, see Section 2.4.2 below. Then a practical version of (2.18) is

$$\tilde{L}(\theta) = \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{\hat{w}_j^*}{s(X_i,\hat{\alpha}^0)}\rho_{\tau_j}\{Y_i - g(X_i,b) - s(X_i,a)q_j\}, \quad \theta = (b^T, a^T, q_1, \ldots, q_k)^T \tag{2.19}$$

Under mild conditions, Theorem 2.7 establishes $\sqrt{n}$-equivalence between regressions (2.18) and (2.19).

Condition 5. In (2.19), $\|\hat{\alpha}^0 - \alpha\| = O_P(n^{-1/2})$ and $\hat{w}_j^* - w_j^* = o_P(1)$, $j = 1, \ldots, k$.

Theorem 2.7. *Assume Conditions 1–3 and 5, then there exists a local minimizer of $\tilde{L}(\theta)$, denote by $\tilde{\theta}$, and a local minimizer of $\bar{L}(\theta)$, denote by $\bar{\theta}$, such that $\|\sqrt{n}(\tilde{\theta} - \bar{\theta})\| = o_P(1)$, therefore $\tilde{\theta}$ has the same asymptotic distribution as $\bar{\theta}$. Consequently, $\tilde{\theta}$ is adaptive.*

2.4.2 An adaptive procedure

From Theorem 2.7, in order to achieve adaptiveness, we need to find $\hat{w}^*$ and $\hat{\alpha}^0$ that satisfy Condition 5. We propose the following procedure:

(W1) Use (2.3) with $w_j \equiv 1$ and $h(X_i) \equiv 1$ to obtain preliminary estimates $\hat{\beta}^0$ and $\hat{\alpha}^0$.

(W2) Compute the estimated noises:

$$\hat{\varepsilon}_i = \frac{Y_i - g(X_i, \hat{\beta}^0)}{s(X_i, \hat{\alpha}^0)}, \quad i = 1, \dots, n. \tag{2.20}$$

(W3) Estimate $f_\varepsilon(u)$ through the nonparametric kernel density estimator:

$$\tilde{f}_\varepsilon(u) = \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{u - \hat{\varepsilon}_i}{b_n}\right) \tag{2.21}$$

for a bandwidth $b_n > 0$ and kernel function $K(\cdot)$.

(W4) Estimate $Q_\varepsilon(\tau)$ by the sample τ-th quantile of $\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$, denoted by $\tilde{Q}_\varepsilon(\tau)$.

(W5) To ensure the symmetry in Condition 3, we symmetrize $\tilde{f}_\varepsilon$ and $\tilde{Q}_\varepsilon(\tau)$ through

$$\hat{f}_\varepsilon(u) = \frac{\tilde{f}_\varepsilon(u) + \tilde{f}_\varepsilon(-u)}{2} \quad \text{and} \quad \hat{Q}_\varepsilon(\tau) = \frac{\tilde{Q}_\varepsilon(\tau) - \tilde{Q}_\varepsilon(1-\tau)}{2}$$

(W6) Plug $\hat{Q}_\varepsilon(\tau)$ and $\hat{f}_\varepsilon(\hat{Q}_\varepsilon(\tau))$ into $I_\beta(w)$ (resp. $I_\alpha(w)$) in Theorem 2.2 and use (2.13) (resp. (2.16)) to obtain the estimated optimal weight $\hat{w}^* = \hat{w}^*_\beta$ (resp. $\hat{w}^*_\alpha$).

Condition 6. (i) $V_i\varepsilon_i \in \mathcal{L}^\delta$, where V_i and δ are defined in Condition 2(v). (ii) The bandwidth $b_n \propto n^{-1/5}$. (iii) The kernel $K(\cdot)$ is Lipschitz continuous and $\int_{\mathbb{R}} K(x)\,\mathrm{d}x = 1$.

Theorem 2.8. *Assume Condtions 1–3 and 6, then $\hat{\alpha}^0$ in step (W1) and $\hat{w}^*$ in step (W6) satisfy Condition 5. Therefore, the above procedure is adaptive.*

In (2.21), we follow Silverman (1986) to use the rule-of-thumb bandwidth b_n:

$$b_n = 0.9n^{-1/5} \min\left\{\mathrm{sd}(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n), \frac{\mathrm{IQR}(\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)}{1.34}\right\}$$

Here, “sd" and “IQR" are the sample standard deviation and the sample interquartile range.

2.5 Technical Proofs

In linear quantile regression, the convexity of the loss function leads to technical simplicity, e.g., point-wise convergence of convex functions implies uniform convergence. Due to the nonlinearity of $g(X_i, b)$ and $s(X_i, a)$ in (2.3), the loss function is no longer convex in the parameter $\theta = (b^T, a^T, q_1, \ldots, q_k)^T$, which adds significant technical complexities. In Section 2.5.1, we present a general result on uniform convergence that may be of independent interest.

Throughout $c, c_1, c_2, \ldots$ are generic finite constants. Without causing confusion, we use 0 to denote both the number zero and the zero vector, depending on specific context.

Lemma 2.1 states some elementary results under mild finite moments conditions. For example, if $U_i = 1$, Lemma 2.1(iv) asserts that it suffices to have $\mathbb{E}(|V_i|^\delta) < \infty$ for some $\delta > 2$ to ensure $n^{-3/2}\sum_{i=1}^n |V_i|^3 = o_P(1)$, although the latter holds trivially under condition $\mathbb{E}(|V_i|^3) < \infty$.

Lemma 2.1. *Let $\{(U_i, V_i)\}_{i\in\mathbb{N}}$ be a stationary process with $V_i \in \mathcal{L}^2$ and $U_iV_i \in \mathcal{L}^\delta$ for some $\delta > 2$. Then (i) $\max_{1\le i\le n}|U_iV_i| = O_P(n^{1/\delta})$; (ii) $\sum_{i=1}^n |U_iV_i^2| = O_P(n)$; (iii) $n^{-3/2}\sum_{i=1}^n |U_iV_i^3| = O_P(n^{1/\delta-1/2})$; and (iv) $n^{-3/2}\sum_{i=1}^n |U_i^2V_i^3| = O_P(n^{1/\delta-1/2})$.*

Proof. (i) Let $s_n \to \infty$. By Markov's inequality,

$$\mathbb{P}(\max_{1\le i\le n}|U_iV_i| > s_n n^{1/\delta}) \le \sum_{i=1}^n \mathbb{P}(|U_iV_i| > s_n n^{1/\delta}) \le n\frac{\mathbb{E}(|U_1V_1|^\delta)}{(s_n n^{1/\delta})^\delta} = O(s_n^{-\delta}) \to 0$$

Since the rate of $s_n \to \infty$ can be arbitrarily slow, we have $\max_{1\le i\le n}|U_iV_i| = O_P(n^{1/\delta})$.

(ii) From $V_i \in \mathcal{L}^2$ and $U_iV_i \in \mathcal{L}^\delta$ with $\delta > 2$, it follows from the Schwarz inequality that $U_iV_i^2 = (U_iV_i)V_i \in \mathcal{L}^1$.

(iii) From (i), $\sum_{i=1}^n |U_iV_i^3| \le \max_{1\le i\le n}|U_iV_i| \times \sum_{i=1}^n V_i^2 = O_P(n^{1/\delta})O_P(n)$.

(iv) From (i) and (ii),

$$\sum_{i=1}^n |U_i^2V_i^3| \le \max_{1\le i\le n}|U_iV_i| \times \sum_{i=1}^n |U_iV_i^2| = O_P(n^{1/\delta})O_P(n).$$

This completes the proof. □

2.5.1 Uniform convergence of stationary processes

Let $\{G_i(\cdot) : \vartheta \mapsto G_i(\vartheta) \in \mathbb{R} | \vartheta \in \mathbb{R}^m\}_{i\in\mathbb{N}}$ and $\{H_i(\cdot) : s \mapsto H_i(s) \in \mathbb{R} | s \in \mathbb{R}\}_{i\in\mathbb{N}}$ be sequences of random functions, and $\{W_i\}$ be a sequence of random variables. Define the random function

$$\mathcal{G}_n(\vartheta) = \sum_{i=1}^{n} \left\{ W_i \int_0^{G_i(\vartheta/\sqrt{n})} [H_i(s) - H_i(0)]\, \mathrm{d}s \right\}, \quad \vartheta \in \mathbb{R}^m. \tag{2.22}$$

Let $\dot{G}_i(\vartheta) = \dfrac{\partial G_i(\vartheta)}{\partial \vartheta} \in \mathbb{R}^m$ be the gradient vector and $\ddot{G}(\vartheta) = \dfrac{\partial^2 G_i(\vartheta)}{\partial \vartheta^2} \in \mathbb{R}^{m\times m}$ be the Hessian matrix. Our goal is to study uniform convergence of $\mathcal{G}_n(\vartheta)$ on compact sets.

Condition 7. In (2.22), $\{W_i, G_i(\cdot), H_i(\cdot)\}$ satisfy the following conditions:

(A1) For each $\vartheta \in \mathbb{R}^m$ and $s \in \mathbb{R}$, $\{(W_i, G_i(\vartheta), H_i(s))\}_{i\in\mathbb{N}}$ is strictly stationary.

(A2) For each i, $\{H_i(s) : s \in \mathbb{R}\}$ is independent of $\{(W_j, G_j(\vartheta), H_{j-1}(s)) : \vartheta \in \mathbb{R}^m, s \in \mathbb{R}\}_{j\le i}$.

(A3) $n^{-1}\sum_{i=1}^n W_i \dot{G}_i(0)\dot{G}_i(0)^T \to_P \mathbb{E}[W\dot{G}(0)\dot{G}(0)^T]$ with (G, W) distributed as (G_i, W_i).

(A4) $G_i(0) = 0, i \in \mathbb{N}$. For some stationary process $\{Z_i\}$ and constants $\epsilon > 0$ and $\delta > 2$,

$$\sup_{\|\vartheta\|\le\epsilon} [\|\dot{G}_i(\vartheta)\| + \|\ddot{G}_i(\vartheta)\|] \le Z_i, \quad Z_i \in \mathcal{L}^2,\ W_i Z_i \in \mathcal{L}^\delta. \tag{2.23}$$

(A5) For $\mathcal{H}(s) = \mathbb{E}[H_i(s)]$, $\sup_{s\in\mathbb{R}}[|\mathcal{H}'(s)| + |\mathcal{H}''(s)|] < \infty$.

(A6) For $\mathcal{J}(s) = \mathbb{E}\{[H_i(s) - H_i(0)]^2\}$, $\sup_{s\in\mathbb{R}} |\mathcal{J}'(s)| < \infty$.

(A7) There exists a constant $c < \infty$ such that $|H_i(s)| < c$ uniformly in s and i.

From the proof of Theorem 9, (A2) enables martingale construction (Lemma 2), and the law of large number condition (A3) is necessary. By ergodic theorem [Theorem 3.5.7 of Stout (1974)], if $\{(W_i, \dot{G}_i(0))\}_{i\in\mathbb{N}}$ is ergodic and $W_i\dot{G}_i(0)\dot{G}_i(0)^T \in \mathcal{L}^1$, then (A3) holds.

Let $\Theta \in \mathbb{R}^m$ be any given compact set. From Condition 7(A4), for large enough n,

$$\sup_{\vartheta\in\Theta}\left|G_i(\vartheta/\sqrt{n})\right| = \sup_{\vartheta\in\Theta}|G_i(\vartheta/\sqrt{n}) - G_i(0)| \le c_1 Z_i/\sqrt{n}, \tag{2.24}$$

where the constant $c_1 = \max_{\vartheta\in\Theta}\|\vartheta\| < \infty$. Similarly, for some constant c_2,

$$\sup_{\vartheta\in\Theta}\left|G_i(\vartheta/\sqrt{n}) - \frac{\dot{G}_i(0)^T\vartheta}{\sqrt{n}}\right| = \sup_{\vartheta\in\Theta}\left|G_i(\vartheta/\sqrt{n}) - G_i(0) - \frac{\dot{G}_i^T(0)\vartheta}{\sqrt{n}}\right| \le c_2\frac{Z_i}{n}. \tag{2.25}$$

Theorem 2.9. *Assume Condition* 7*. Recall $\mathcal{H}(s) = \mathbb{E}[H_i(s)]$ in Condition 7(A5). Then, for any compact set $\Theta \subset \mathbb{R}^m$, we have the uniform convergence*

$$\sup_{\vartheta\in\Theta}\left|\mathcal{G}_n(\vartheta) - \frac{\mathcal{H}'(0)}{2}\vartheta^T\mathbb{E}[W\dot{G}(0)\dot{G}(0)^T]\vartheta\right| = o_P(1).$$

Proof. We have the decomposition $\mathcal{G}_n(\vartheta) = \bar{\mathcal{G}}_n(\vartheta) + \mathcal{N}_n(\vartheta) + \mathcal{R}_n(\vartheta)$, where

$$\begin{aligned}
\bar{\mathcal{G}}_n(\vartheta) &= \frac{\mathcal{H}'(0)}{2}\sum_{i=1}^n W_i G_i(\vartheta/\sqrt{n})^2,\\
\mathcal{N}_n(\vartheta) &= \sum_{i=1}^n W_i\int_0^{G_i(\vartheta/\sqrt{n})}[\mathcal{H}(s) - \mathcal{H}(0) - s\mathcal{H}'(0)]\,\mathrm{d}s,\\
\mathcal{R}_n(\vartheta) &= \sum_{i=1}^n W_i\int_0^{G_i(\vartheta/\sqrt{n})}\{[H_i(s) - H_i(0)] - [\mathcal{H}(s) - \mathcal{H}(0)]\}\,\mathrm{d}s.
\end{aligned} \tag{2.26}$$

First, consider $\bar{\mathcal{G}}_n(\vartheta)$. We have

$$\sup_{\vartheta\in\Theta}\left|\bar{\mathcal{G}}_n(\vartheta) - \frac{\mathcal{H}'(0)}{2}\vartheta^T\left[\frac{1}{n}\sum_{i=1}^n W_i\dot{G}_i(0)\dot{G}_i(0)^T\right]\vartheta\right| \le \frac{|\mathcal{H}'(0)|}{2}\Lambda_n, \tag{2.27}$$

where $\Lambda_n = \sum_{i=1}^n\left\{|W_i|\sup_{\theta\in\Theta}\left|G_i(\vartheta/\sqrt{n})^2 - \left[\dot{G}_i(0)^T\vartheta/\sqrt{n}\right]^2\right|\right\}$. Observe the elementary inequality

$$|x^2 - y^2| = |(x-y)(x-y+2y)| \le \delta_1(\delta_1 + 2\delta_2), \text{ for } |x-y| \le \delta_1, |y| \le \delta_2.$$

Applying the above inequality with $x = G_i(\vartheta/\sqrt{n})$ and $y = \dot{G}_i(0)^T\vartheta/\sqrt{n}$ and using the bound in (2.25) for $x - y$ and the bound in (2.23), we obtain

$$\Lambda_n \le c_3 \sum_{i=1}^{n} |W_i| \frac{Z_i}{n} \left(\frac{Z_i}{n} + \frac{Z_i}{\sqrt{n}} \right) = O_P(n^{-1/2}), \tag{2.28}$$

for some constant $c_3 < \infty$. Here the last bound $O_P(n^{-1/2})$ follows from (2.23) in Condition 7(A4) and Lemma 1(ii). Thus, from (2.27), (2.28) and Condition 7(A3), we conclude that $\bar{\mathcal{G}}_n(\vartheta) = [\mathcal{H}'(0)/2]\vartheta^T \mathbb{E}[\dot{G}(0)\dot{G}(0)^T]\vartheta + o_P(1)$, uniformly.

Next, consider $\mathcal{N}_n(\vartheta)$. By Condition 7(A5), $|\mathcal{H}(s) - \mathcal{H}(0) - s\mathcal{H}'(0)| = O(s^2)$. By (2.24),

$$\sup_{\vartheta \in \Theta} |\mathcal{N}_n(\vartheta)| = O(1) \sum_{i=1}^{n} |W_i| \sup_{\vartheta \in \Theta} |G_i(\vartheta/\sqrt{n})|^3 = \frac{O(1)}{n^{3/2}} \sum_{i=1}^{n} |W_i Z_i^3| = o_P(1),$$

in view of Lemma 1(iii). Finally, by Lemma 2 below, the proof is completed. □

Lemma 2.2. *For $\mathcal{R}_n(\vartheta)$ defined in* (2.26)*, under the conditions in Theorem 2.9,* $\sup_{\vartheta \in \Theta} |\mathcal{R}_n(\vartheta)| = o_P(1)$*.*

Proof. We claim that $\mathcal{R}_n(\vartheta)$ is a martingale for each ϑ. Write $\mathcal{R}_n(\vartheta) = \sum_{i=1}^n r_i(\vartheta)$, where

$$r_i(\vartheta) = W_i \int_0^{G_i(\vartheta/\sqrt{n})} \{[H_i(s) - H_i(0)] - [\mathcal{H}(s) - \mathcal{H}(0)]\} \, \mathrm{d}s.$$

Let $\mathscr{F}_i$ be the σ-algebra generated by $\{W_{j+1}, G_{j+1}(\vartheta), H_j(s) : \vartheta \in \mathbb{R}^m, s \in \mathbb{R}, j \le i\}$. By Condition 7(A2), $\{H_i(s) : s \in \mathbb{R}\}$ is independent of $\mathscr{F}_{i-1}$. Thus $\mathbb{E}[H_i(s) - H_i(0) \mid \mathscr{F}_{i-1}] = \mathbb{E}[H_i(s) - H_i(0)] = \mathcal{H}(s) - \mathcal{H}(0)$ hence $\mathbb{E}[r_i(\vartheta) \mid \mathscr{F}_{i-1}] = 0$. Consequently, $\{r_i(\vartheta)\}_{i \in \mathbb{N}}$ is a martingale difference sequence with respect to the filtration $\{\mathscr{F}_i\}_{i \in \mathbb{N}}$ for each ϑ.

Assume without loss of generality that $\vartheta \in \mathbb{R}$ is a one-dimensional parameter and $\Theta = [0, 1]$. The multivariate case can be treated similarly. To bound $\mathcal{R}_n(\vartheta)$ uniformly on $\vartheta \in [0, 1]$, we proceed in four steps:

(i) Discretize the continuous martingale process $\mathcal{R}_n(\vartheta)$, $\vartheta \in [0, 1]$, at some grid points.

(ii) Truncate the discretized $\mathcal{R}_n(\vartheta)$ so that it is deterministically bounded.

(iii) Apply Freedman's exponential inequality for bounded martingales (Freedman (1975)).

(iv) Show that the effect of truncation is negligible.

Step (i): Introduce the grid points $\vartheta_j = j/n, j = 0, 1, \ldots, n$ on $[0, 1]$. Since any point $\vartheta \in [0, 1]$ is at most $1/n$ away from the set $\{\vartheta_1, \ldots, \vartheta_n\}$, we have the inequality

$$\sup_{\vartheta \in [0,1]} |\mathcal{R}_n(\vartheta)| \leq \max_{1 \leq j \leq n} |\mathcal{R}_n(\vartheta_j)| + \sup_{|\vartheta - \vartheta'| \leq 1/n} |\mathcal{R}_n(\vartheta) - \mathcal{R}_n(\vartheta')|. \tag{2.29}$$

By Condition 7(A7), $|[H_i(s) - H_i(0)| - [\mathcal{H}(s) - \mathcal{H}(0)]| \leq 4c$. Thus, by Condition 7(A4),

$$|r_i(\vartheta) - r_i(\vartheta')| \leq 4c|W_i||G_i(\vartheta/\sqrt{n}) - G_i(\vartheta'/\sqrt{n})| \leq 4c\frac{|\vartheta - \vartheta'|}{\sqrt{n}}|W_i Z_i|.$$

Therefore, by (2.29), we conclude that the discretization effect is negligible in view of

$$\sup_{|\vartheta - \vartheta'| \leq 1/n} |\mathcal{R}_n(\vartheta) - \mathcal{R}_n(\vartheta')| \leq \frac{4c}{n^{3/2}} \sum_{i=1}^{n} |W_i Z_i| = O_P(n^{-1/2}) = o_P(1). \tag{2.30}$$

Step (ii): Since $r_i(\vartheta_j)$ is not deterministically bounded (although bounded in probability), we cannot directly apply Freedman's exponential inequality which deals with bounded martingale differences. To address this issue, we use truncation. Let

$$A_i = \sup_{\vartheta \in [0,1]} |G_i(\vartheta/\sqrt{n})|, \quad i = 1, \ldots, n.$$

Recall $\delta > 2$ in Condition 7(A4). Consider the events

$$E_j = \left[|\mathcal{R}_n(\vartheta_j)| \geq \frac{1}{\log n}\right], j = 1, \ldots, n.$$
$$T = \left[\max_{1 \leq i \leq n} |W_i A_i| < n^{1/\delta - 1/2}(\log n)^2, \sum_{i=1}^{n} W_i^2 A_i^3 < n^{1/\delta - 1/2} \log n\right]. \tag{2.31}$$

Since $[\max_{1 \leq j \leq n} |\mathcal{R}_n(\vartheta_j)| \geq 1/\log n] = \bigcup_{j=1}^n E_j = [\bigcup_{j=1}^n (E_j \cap T)] \bigcup [(\bigcup_{j=1}^n E_j) \cap$

T^c], it follows that

$$P\left[\max_{1\le j\le n}|\mathcal{R}_n(\vartheta_j)|\ge\frac{1}{\log n}\right]\le\sum_{j=1}^n P(E_j\cap T)+P(T^c). \tag{2.32}$$

As will be shown below, the choice of T makes it easy to handle $P(E_j\cap T)$ and $P(T^c)$.

Step (iii): Consider the event $E_j\cap T$. By definitions of $r_i(\vartheta_j)$ and T, we have

$$|r_i(\vartheta_j)|\le 4c|W_i||A_i|\le 4c\max_{1\le i\le n}|W_iA_i| \tag{2.33}$$

uniformly in i,j. Recall $\mathcal{J}(s)=\mathbb{E}\{[H_i(s)-H_i(0)]^2\}$ in Condition 7(A6), since $\mathcal{J}(0)=0$ and $\mathcal{J}(s)$ has bounded derivative, we have $\mathcal{J}(s)=\mathcal{J}(s)-\mathcal{J}(0)=O(s)$, therefore,

$$\mathbb{E}\left(\{[H_i(s)-H_i(0)]-[\mathcal{H}(s)-\mathcal{H}(0)]\}^2\right)\le\mathbb{E}[(H_i(s)-H_i(0))^2]=\mathcal{J}(s)=O(s). \tag{2.34}$$

It then follows by (2.34), the independence between $\{H_i(s):s\in\mathbb{R}\}$ and $\mathscr{F}_{i-1}$, the Schwarz inequality, and the Tonelli's theorem that

$$\begin{aligned}\mathbb{E}[r_i^2(\vartheta_j)\mid\mathscr{F}_{i-1}]&\le W_i^2\mathbb{E}\left[\left(\int_{-A_i}^{A_i}|[H_i(s)-H_i(0)]-[\mathcal{H}(s)-\mathcal{H}(0)]|\,\mathrm{d}s\right)^2\middle|\mathscr{F}_{i-1}\right]\\&\le W_i^2\mathbb{E}\left[2A_i\int_{-A_i}^{A_i}\{[H_i(s)-H_i(0)]-[\mathcal{H}(s)-\mathcal{H}(0)]\}^2\,\mathrm{d}s\middle|\mathscr{F}_{i-1}\right]\\&=2W_i^2A_i\int_{-A_i}^{A_i}\mathbb{E}[\{[H_i(s)-H_i(0)]-[\mathcal{H}(s)-\mathcal{H}(0)]\}^2\mid\mathscr{F}_{i-1}]\,\mathrm{d}s\\&=2W_i^2A_i\int_{-A_i}^{A_i}\mathbb{E}[\{[H_i(s)-H_i(0)]-[\mathcal{H}(s)-\mathcal{H}(0)]\}^2]\,\mathrm{d}s\\&\le W_i^2A_i\int_{-A_i}^{A_i}O(s)\,\mathrm{d}s\le c_4W_i^2A_i^3\end{aligned} \tag{2.35}$$

uniformly in j, for some constant c_4. By (2.33) and (2.35), on the event (2.31), the martingale differences $r_i(\vartheta_j)$ are bounded by $4cn^{1/\delta-1/2}(\log n)^2$ and the sum of conditional variances $\sum_{i=1}^n\mathbb{E}[r_i^2(\vartheta_j)\mid\mathscr{F}_{i-1}]\le c_4\sum_{i=1}^n W_i^2A_i^3\le c_4n^{1/\delta-1/2}\log n$, uniformly in j. Thus, by Freedman's exponential inequality for bounded martingale differences,

$$P(E_j \cap T) \le 2\exp\left(-\frac{(1/\log n)^2}{2\{[4cn^{1/\delta-1/2}(\log n)^2](1/\log n) + c_4 n^{1/\delta-1/2}\log n\}}\right)$$
$$= 2\exp(-\lambda_n \log n), j = 1, 2, \ldots, n. \tag{2.36}$$

where $\lambda_n = \dfrac{n^{1/2-1/\delta}}{2(4c+c_4)(\log n)^4}$. Since $\delta > 2$, for large enough n, we have $\lambda_n > 2$. Thus, in (2.32),

$$\sum_{j=1}^{n} P(E_j \cap T) \le 2n\exp(-\lambda_n \log n) = O(1/n). \tag{2.37}$$

Step (iv): It remains to show that $P(T^c) \to 0$. Notice that

$$P(T^c) \le P\left[\max_{1\le i\le n} |W_i A_i| \ge n^{1/\delta-1/2}(\log n)^2\right] + P\left[\sum_{i=1}^{n} W_i^2 A_i^3 \ge n^{1/\delta-1/2}\log n\right]. \tag{2.38}$$

By (2.24), $A_i \le c_1 Z_i/\sqrt{n}$, therefore Condition 7(A4) and Lemma 2.1 imply that $\max_{1\le i\le n}|W_iA_i| \le c_1 n^{-1/2}\max_{1\le i\le n}|W_iZ_i| = O_P(n^{1/\delta-1/2})$ and $\sum_{i=1}^n W_i^2A_i^3 \le c_1^3 n^{-3/2}\sum_{i=1}^n W_i^2 Z_i^3 = O_P(n^{1/\delta-1/2})$. Therefore, the probabilities on the right hand side of (2.38) are of order $o(1)$, hence $P(T^c) \to 0$. This result together with (2.37) and (2.32) imply that $\max\limits_{1\le j\le n} |\mathcal{R}_n(\vartheta_j)| = o_P(1)$. The lemma then follows from (2.29) and (2.30). □

2.5.2 Proofs of Theorem 2.1, Theorem 2.2, Theorem 2.7

Lemma 2.3. *Let U be a column random vector. Suppose that the matrix $\mathbb{E}[UU^T]$ is positive definite, then for any random variable $\eta > 0$, $\mathbb{E}[UU^T/\eta]$ is also positive definite.*

Proof. Suppose $x^TE[UU^T/\eta]x = 0$ for some vector x, then $E[(x^TU/\sqrt{\eta})^2] = 0$, which implies that $P[x^TU/\sqrt{\eta} = 0] = 1$. Consequently, $P[x^TU = 0] = 1$ and $x^TE[UU^T]x = E[(x^TU)^2] = 0$, which implies that x is a zero vector by the positive definiteness of the matrix $E[UU^T]$. □

Lemma 2.4. *Under Condition 2(i)–(iv), the matrix M in Theorem 2.1 is positive definite.*

Proof. Recall D in Condition 2(iii). For $x = (x_\beta^T, x_\alpha^T, x_1, \ldots, x_k)^T$ with $x_\beta \in \mathbb{R}^p$, $x_\alpha \in \mathbb{R}^q$,

$$x^T M x = \sum_{j=1}^{k} w_j f(Q_\varepsilon(\tau_j))(x_\beta^T, Q_\varepsilon(\tau_j) x_\alpha^T, x_j) \mathbb{E}\left[\frac{DD^T}{h(X)s(X,\alpha)}\right] \begin{bmatrix} x_\beta \\ Q_\varepsilon(\tau_j) x_\alpha \\ x_j \end{bmatrix} \geq 0. \tag{2.39}$$

By Condition 2(iii) and Lemma 2.3, $\mathbb{E}\{DD^T/(h(X)s(X,\alpha))\}$ is positive definite. Since $w_j f_\varepsilon(Q_\varepsilon(\tau_j)) > 0$ (Condition 2(i) and (iv)), $x^T M x = 0$ implies that every summand in (2.39) is zero so that $(x_\beta^T, Q_\varepsilon(\tau_j) x_\alpha^T, x_j)^T$ is a zero vector for every j. It follows by Condition 2(i) that x must be a zero vector. □

Lemma 2.5. *Let $g_j : \mathbb{R}^p \to \mathbb{R}^m, j = 1, \ldots, k$ be k vector-valued functions, and $h_j : \mathbb{R} \to \mathbb{R}, j = 1, \ldots, k$ be k scalar functions such that $g_j(X_i)h_j(\varepsilon_i) \in \mathcal{L}^2$, $\mathbb{E}[h_j(\varepsilon_i)] = 0$, $j = 1, \ldots, k$, and the matrix*

$$C(g,h) = \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{E}[g_j(X_0) g_{j'}(X_0)^T] \mathbb{E}[h_j(\varepsilon_0) h_{j'}(\varepsilon_0)] \in \mathbb{R}^{m \times m}$$

is non-singular. Then, under Condition 1, the multivariate CLT holds

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \Rightarrow \mathcal{N}(0, C(g,h)), \text{ where } \xi_i = \sum_{j=1}^{k} g_j(X_i) h_j(\varepsilon_i). \tag{2.40}$$

Proof. By the Cramér-Wold device, without loss of generality, we consider $m = 1$. Let $\mathscr{F}_i$ be the algebra generated by $\{X_{j+1}, \varepsilon_j : j \leq i\}$. By Condition 1(i), ε_i is independent of $\mathscr{F}_{i-1}$. Thus, $\mathbb{E}[g_j(X_i)h_j(\varepsilon_i) \mid \mathscr{F}_{i-1}] = g_j(X_i)\mathbb{E}[h_j(\varepsilon_i)] = 0$ and $\mathbb{E}[\xi_i \mid \mathscr{F}_{i-1}] = 0$. Hence $\{\xi_i\}_{i \in \mathbb{N}}$ is a sequence of martingale differences with respect to $\{\mathscr{F}_i\}_{i \in \mathbb{N}}$. Since ε_i is independent of $\mathscr{F}_{i-1}$,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\xi_i^2 \mid \mathscr{F}_{i-1}] = \sum_{j=1}^{k} \sum_{j'=1}^{k} \mathbb{E}[h_j(\varepsilon_0) h_{j'}(\varepsilon_0)] \left[\frac{1}{n} \sum_{i=1}^{n} g_j(X_i) g_{j'}(X_i)\right] \to C(g,h)$$

with probability 1, by Condition 1(ii) and the ergodic theorem. Let $\varepsilon > 0$, it follows by $\xi_0 \in \mathcal{L}^2$ and the dominated convergence theorem that $n^{-1}\sum_{i=1}^{n} \mathbb{E}[\xi_i^2 \mathbf{1}(|\xi_i| \geq \varepsilon\sqrt{n})] = \mathbb{E}[\xi_0^2 \mathbf{1}(|\xi_0| > \varepsilon\sqrt{n})] \to 0$, thus the Lindeberg condition holds, the result then follows from the martingale CLT (Billingsley (1995), p.476). □

Proof of Theorem 2.1. To show that $L(\theta)$ has a local minimizer $\hat{\theta}$ satisfying $\|\hat{\theta} - \theta\| = O_P(n^{-1/2})$, it suffices to show that, for any given $\nu > 0$, there exists a constant C such that, for all large enough n,

$$P\left[\inf_{\|\vartheta\|=C} \chi(\vartheta) > 0\right] \geq 1 - \nu, \text{ where } \chi(\vartheta) = L(\theta_0 + \vartheta/\sqrt{n}) - L(\theta_0). \tag{2.41}$$

This means that, with probability at least $1-\nu$, $L(\theta_0+\vartheta/\sqrt{n}) > L(\theta_0)$ for $\|\vartheta\| = C$. Therefore, the continuous function $L(\theta)$ has a local minimizer on the compact ball $\{\theta_0 + \vartheta/\sqrt{n} : \|\vartheta\| \leq C\}$, and the minimizer, denoted by $\hat{\theta}$, must satisfy $\|\hat{\theta} - \theta_0\| = O_P(n^{-1/2})$.

Partition ϑ according to θ_0 as $\vartheta = (\vartheta_\beta^T, \vartheta_\alpha^T, \vartheta_1, \ldots, \vartheta_k)^T$. Define the random functions

$$D_{ij}(\vartheta) = \frac{g(X_i, \beta + \vartheta_\beta) - g(X_i, \beta)}{s(X_i, \alpha)} + \frac{s(X_i, \alpha + \vartheta_\alpha)}{s(X_i, \alpha)}[Q_\varepsilon(\tau_j) + \vartheta_j] - Q_\varepsilon(\tau_j) \tag{2.42}$$

for $i = 1, \ldots, n, j = 1, \ldots, k$. It is easy to see that

$$\frac{Y_i - g(X_i, \beta + \vartheta_\beta) - s(X_i, \alpha + \vartheta_\alpha)[Q_\varepsilon(\tau_j) + \vartheta_j]}{s(X_i, \alpha)} = \varepsilon_i - Q_\varepsilon(\tau_j) - D_{ij}(\vartheta).$$

From the above equation and the linearity $\rho_\tau(cz) = c\rho_\tau(z)$ for $c > 0$, for $\chi(\vartheta)$ in (2.41),

$$\chi(\vartheta) = \sum_{i=1}^{n}\sum_{j=1}^{k} \frac{w_j s(X_i, \alpha)}{h(X_i)} \left\{\rho_{\tau_j}\left(\varepsilon_i - Q_\varepsilon(\tau_j) - D_{ij}(\vartheta/\sqrt{n})\right) - \rho_{\tau_j}\left(\varepsilon_i - Q_\varepsilon(\tau_j)\right)\right\} \tag{2.43}$$

By Knight's identity (Knight (1998)) $\rho_\tau(x - y) - \rho_\tau(x) = -y(\tau - \mathbf{1}(x < 0)) +$

$\int_0^y (\mathbf{1}(x < s) - \mathbf{1}(x < 0))\,\mathrm{d}s$,

$$\chi(\vartheta) = -\sum_{j=1}^{k}\sum_{i=1}^{n} \frac{w_j s(X_i,\alpha)}{h(X_i)} D_{ij}(\vartheta/\sqrt{n})[\tau_j - \mathbf{1}(\varepsilon_i < Q_\varepsilon(\tau_j))] + R(\vartheta), \qquad (2.44)$$

where

$$R(\vartheta) = \sum_{j=1}^{k} w_j \left\{ \sum_{i=1}^{n} \frac{s(X_i,\alpha)}{h(X_i)} \int_0^{D_{ij}\left(\frac{\vartheta}{\sqrt{n}}\right)} [\mathbf{1}(\varepsilon_i < Q_\varepsilon(\tau_j) + s) - \mathbf{1}(\varepsilon_i < Q_\varepsilon(\tau_j))]\,\mathrm{d}s \right\}.$$

Applying Theorem 2.9, we have

$$\sup_{\|\vartheta\| \le C} \left| R(\vartheta) - \frac{1}{2}\vartheta^T M \vartheta \right| = o_P(1). \qquad (2.45)$$

where M is defined in Theorem 2.1.

Let $\dot{D}_{ij}(\vartheta)$ be the gradient of $D_{ij}(\vartheta)$. By the same martingale-discretization-truncation technique used in the proof of Lemma 2.2, it can be shown that

$$\sup_{\|\vartheta\| \le C} \left| \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{w_j s(X_i,\alpha)}{h(X_i)} \left[D_{ij}\left(\frac{\vartheta}{\sqrt{n}}\right) - D_{ij}(0) - \frac{\vartheta^T \dot{D}_{ij}(0)}{\sqrt{n}} \right] [\tau_j - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau_j)}] \right|$$
$$= o_P(1). \qquad (2.46)$$

Note that $D_{ij}(0) = 0$. From (2.44), (2.45) and (2.46),

$$\chi(\vartheta) = -\vartheta^T \gamma_n + \frac{1}{2}\vartheta^T M \vartheta + o_P(1) \qquad (2.47)$$

holds uniformly on $\|\vartheta\| \le C$, where

$$\gamma_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \text{ and } \xi_i = \sum_{j=1}^{k} \frac{w_j s(X_i,\alpha)}{h(X_i)} [\tau_j - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau_j)}] \dot{D}_{ij}(0). \qquad (2.48)$$

From the proof of Lemma 2.5, $\{\xi_i\}_{i \in \mathbb{N}}$ is a sequence martingale differences. By the orthogonality of martingale differences, $\mathbb{E}[\|\gamma_n\|^2] = n^{-1}\sum_{i=1}^{n} \mathbb{E}[\|\xi_i\|^2] = \mathbb{E}[\|\xi_0\|^2] < \infty$. The Schwarz inequality then gives that $|\vartheta^T \gamma_n| \le \|\vartheta\| \|\gamma_n\| = C O_P(1)$ on $\|\vartheta\| = C$. On the other hand, on $\|\vartheta\| = C$, $\vartheta^T M \vartheta \ge \rho_0 \|\vartheta\|^2 = \rho_0 C^2$, where ρ_0 is the smallest eigenvalue of M and $\rho_0 > 0$ (Lemma 2.4). Since the quadratic term grows faster than the linear term $CO_P(1)$, we can choose a large enough C so that (2.41) holds.

To prove the CLT for $\hat{\theta}$, let $\vartheta = \sqrt{n}(\theta - \theta_0)$. Since $\hat{\theta}$ is a local minimizer of the criterion function $L(\theta)$ in (2.3), the reparameterized parameter $\hat{\vartheta} = \sqrt{n}(\hat{\theta} - \theta_0)$ is a local minimizer of the reparameterized function $\chi(\vartheta)$ defined in (2.41). By (2.47), $\chi(\vartheta) = -\vartheta^T\gamma_n + \vartheta^T M\vartheta/2 + o_P(1)$ on the space of functions topologized by uniform convergence on compact sets. Also, by the above analysis, the local minimizer $\hat{\vartheta}$ satisfies $\|\hat{\vartheta}\| = O_P(1)$. Thus, by Theorem 2.7 of Kim and Pollard (1990) (see also the proof of Theorem 3 in Knight and Fu (2000)),

$$\hat{\vartheta} = \arg\min_{\vartheta}\left(-\vartheta^T\gamma_n + \frac{1}{2}\vartheta^T M\vartheta\right) + o_P(1) = M^{-1}\gamma_n + o_P(1). \tag{2.49}$$

Note that $\mathrm{Cov}(\tau - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau)}, \tau' - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau')}) = \min(\tau, \tau') - \tau\tau'$ for $\tau, \tau' \in (0, 1)$. Applying Lemma 2.5, we have $\gamma_n \Rightarrow \mathcal{N}(0, \Omega)$ with Ω defined in Theorem 1. Therefore, we conclude that $\hat{\vartheta} = \sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, M^{-1}\Omega M^{-1})$. □

Proof of Theorem 2.2. For notational convenience, write

$$U = \frac{\dot{g}(X, \beta)}{h(X)}, V = \frac{\dot{s}(X, \alpha)}{h(X)}, \eta = \frac{s(X, \alpha)}{h(X)}, c_j = w_j f_\varepsilon(Q_\varepsilon(\tau_j)), d_j = Q_\varepsilon(\tau_j).$$

For M in (2.7), we have the partitioned form:

$$M = \left[\begin{array}{cc|cccc}
\mathbb{E}\left[\frac{UU^T}{\eta}\right]\sum\limits_{j=1}^{k} c_j & \mathbb{E}\left[\frac{UV^T}{\eta}\right]\sum\limits_{j=1}^{k} c_j d_j & \mathbb{E}[U]c_1 & \mathbb{E}[U]c_2 & \cdots & \mathbb{E}[U]c_k \\
\mathbb{E}\left[\frac{VU^T}{\eta}\right]\sum\limits_{j=1}^{k} c_j d_j & \mathbb{E}\left[\frac{VV^T}{\eta}\right]\sum\limits_{j=1}^{k} c_j d_j^2 & \mathbb{E}[V]c_1 d_1 & \mathbb{E}[V]c_2 d_2 & \cdots & \mathbb{E}[V]c_k d_k \\
\hline
\mathbb{E}[U^T]c_1 & \mathbb{E}[V^T]c_1 d_1 & \mathbb{E}[\eta]c_1 & 0 & \cdots & 0 \\
\mathbb{E}[U^T]c_2 & \mathbb{E}[V^T]c_2 d_2 & 0 & \mathbb{E}[\eta]c_2 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\mathbb{E}[U^T]c_k & \mathbb{E}[V^T]c_k d_k & 0 & 0 & \cdots & \mathbb{E}[\eta]c_k
\end{array}\right]$$

$$:= \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}$$

where $A \in \mathbb{R}^{(p+q)\times(p+q)}$, $B \in \mathbb{R}^{(p+q)\times k}$, $D \in \mathbb{R}^{k\times k}$. By the block matrix inversion formula, with $E = A - BD^{-1}B^T$:

$$\begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}B^TE^{-1} & D^{-1} + D^{-1}B^TE^{-1}BD^{-1} \end{bmatrix}.$$

Therefore by (2.48), (2.49), it follows that

$$\sqrt{n}\left[\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} - \begin{pmatrix} \beta \\ \alpha \end{pmatrix}\right] = E^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{k} w_j[\tau_j - \mathbf{1}_{\varepsilon_i < Q_\varepsilon(\tau_j)}]U_j(X_i) + o_P(1), \quad (2.50)$$

where

$$U_j(X_i) = \begin{bmatrix} \dot{g}(X_i, \beta)/h(X_i) \\ Q_\varepsilon(\tau_j)\dot{s}(X_i, \alpha)/h(X_i) \end{bmatrix} - BD^{-1}e_j\frac{s(X_i, \alpha)}{h(X_i)}. \quad (2.51)$$

Condition 3 implies that $\sum_{j=1}^{k} w_j Q_\varepsilon(\tau_j) f_\varepsilon(Q_\varepsilon(\tau_j)) = 0$. It then can be shown that

$$E = \begin{bmatrix} M_\beta \sum_{j=1}^{k} w_j f_\varepsilon(Q_\varepsilon(\tau_j)) & 0 \\ 0 & M_\alpha \sum_{j=1}^{k} w_j Q_\varepsilon(\tau_j)^2 f_\varepsilon(Q_\varepsilon(\tau_j)) \end{bmatrix} \quad (2.52)$$

is a block diagonal matrix. The result then follows from (2.50)–(2.52) and Lemma 2.5. □

Proof of Theorem 2.7. Recall $\chi(\vartheta)$ in (2.41). With $\bar{L}(\theta)$ and $\tilde{L}(\theta)$ in (2.18) and (2.19), define

$$\bar{\chi}(\vartheta) = \bar{L}(\theta_0 + \vartheta/\sqrt{n}) - \bar{L}(\theta_0) \text{ and } \tilde{\chi}(\vartheta) = \tilde{L}(\theta_0 + \vartheta/\sqrt{n}) - \tilde{L}(\theta_0). \quad (2.53)$$

By the proof of Theorem 1, $\bar{\chi}(\vartheta)$ has the uniform quadratic approximation (2.47) on compact sets, which leads to the asymptotic Bahadur representation of $\sqrt{n}(\bar{\theta} - \theta_0)$ in (2.49). By the same argument, to prove the asymptotic $\sqrt{n}$-equivalence of $\tilde{\theta}$ and $\bar{\theta}$, it suffices to prove that $\tilde{\chi}(\vartheta) - \bar{\chi}(\vartheta) = o_P(1)$, uniformly on $\|\vartheta\| \le C$ for any $C > 0$.

Recall $D_{ij}(\vartheta)$ in (2.42). Let

$$T_{ij}(\vartheta) = \rho_{\tau_j}\left[\varepsilon_i - Q_\varepsilon(\tau_j) - D_{ij}(\vartheta/\sqrt{n})\right] - \rho_{\tau_j}\left[\varepsilon_i - Q_\varepsilon(\tau_j)\right] \quad (2.54)$$

It then can be seen that

$$\bar{\chi}(\vartheta) = \sum_{j=1}^{k} w_j^* \left[\sum_{i=1}^{n} T_{ij}(\vartheta) \right] \text{ and } \tilde{\chi}(\vartheta) = \sum_{j=1}^{k} \hat{w}_j^* \left[\sum_{i=1}^{n} \frac{s(X_i, \alpha)}{s(X_i, \hat{\alpha}^0)} T_{ij}(\vartheta) \right]$$

Applying the elementary identity $x/y - 1 = (x-y)^2/(xy) - (y-x-z)/x - z/x$ with $x = s(X_i, \alpha), y = s(X_i, \hat{\alpha}^0)$ and $z = (\hat{\alpha}^0 - \alpha)^T \dot{s}(X_i, \alpha)$, we have the decompsition

$$\tilde{\chi}(\vartheta) - \bar{\chi}(\vartheta) = \sum_{j=1}^{k} \left\{ (\hat{w}_j^* - w_j^*)\Delta_{1j}(\vartheta) + \hat{w}_j^* [\Delta_{2j}(\vartheta) - \Delta_{3j}(\vartheta) - \Delta_{4j}(\vartheta)] \right\},$$

where

$$\begin{aligned}
\Delta_{1j}(\vartheta) &= \sum_{i=1}^{n} T_{ij}(\vartheta), \\
\Delta_{2j}(\vartheta) &= \sum_{i=1}^{n} \frac{[s(X_i, \alpha) - s(X_i, \hat{\alpha}^0)]^2}{s(X_i, \alpha) s(X_i, \hat{\alpha}^0)} T_{ij}(\vartheta), \\
\Delta_{3j}(\vartheta) &= \sum_{i=1}^{n} \frac{s(X_i, \hat{\alpha}^0) - s(X_i, \alpha) - (\hat{\alpha}^0 - \alpha)^T \dot{s}(X_i, \alpha)}{s(X_i, \alpha)} T_{ij}(\vartheta), \\
\Delta_{4j}(\vartheta) &= (\hat{\alpha}^0 - \alpha)^T \sum_{i=1}^{n} \frac{\dot{s}(X_i, \alpha)}{s(X_i, \alpha)} T_{ij}(\vartheta).
\end{aligned}$$

Since k is fixed and $\hat{w}_j^* - w_j^* = o_P(1)$, it suffices to show that $\Delta_{1j}(\vartheta) = O_P(1)$, $\Delta_{2j}(\vartheta) = o_P(1)$, $\Delta_{3j}(\vartheta) = o_P(1)$, and $\Delta_{4j}(\vartheta) = o_P(1)$, uniformly on $\|\vartheta\| \leq C$. The same argument in (2.43)–(2.47) yields $\Delta_{1j}(\vartheta) = O_P(1)$ and $\Delta_{4j}(\vartheta) = O_P(\|\hat{\alpha}^0 - \alpha\|) = O_P(n^{-1/2})$, uniformly on $\|\vartheta\| \leq C$. It remains to prove the other two assertions.

First, consider $\Delta_{2j}(\vartheta)$. For $T_{ij}(\vartheta)$ in (2.42), since for any $x, y \in \mathbb{R}, \tau \in (0,1)$, $|\rho_\tau(x-y) - \rho_\tau(x)| \leq |y|$, it follows by Condition 2(v) and Theorem 9.19 in Rudin (1976) that

$$|T_{ij}(\vartheta)| \leq |D_{ij}(\vartheta/\sqrt{n})| = |D_{ij}(\vartheta/\sqrt{n}) - D_{ij}(0)| \leq c_1 V_i/\sqrt{n} \tag{2.55}$$

uniformly on $\|\vartheta\| \leq C$, for some constant $c_1 > 0$. Similarly,

$$|s(X_i, \hat{\alpha}^0) - s(X_i, \alpha)| \leq V_i \|\hat{\alpha}^0 - \alpha\|. \tag{2.56}$$

An application of Lemma 2.1(i) with $U_i = 1$ gives $\max_{1 \leq i \leq n} |V_i| = O_P(n^{1/\delta})$. Since $\|\hat{\alpha}^0 - \alpha\| = O_P(n^{-1/2})$ and $\delta > 2$, from (2.56), we have $\max_{1 \leq i \leq n} |s(X_i, \hat{\alpha}^0) - s(X_i, \alpha)| \leq \|\hat{\alpha}^0 - \alpha\| \max_{1 \leq i \leq n} |V_i| = o_P(1)$. Thus, by the condition $s(X_i, \alpha) > c > 0$ in Condition 2(ii), with probability tending to one, $s(X_i, \hat{\alpha}^0) > c/2$ for all i. Thus, in view of (2.55)–(2.56), we conclude that

$$\sup_{\|\vartheta\| \leq C} |\Delta_{2j}(\vartheta)| \leq \frac{O_P(1)}{n^{3/2}} \sum_{i=1}^{n} V_i^3 = o_P(1),$$

where the last convergence follows by applying Lemma (2.1)(iii) with $U_i = 1$.

Finally, let's consider $\Delta_{3j}(\vartheta)$. By expanding one more term in (2.56), we have $|s(X_i, \hat{\alpha}^0) - s(X_i, \alpha) - (\hat{\alpha}^0 - \alpha)^T \dot{s}(X_i, \alpha)| \leq V_i \|\hat{\alpha}^0 - \alpha\|^2$, which together with (2.55) gives

$$\sup_{\|\vartheta\| \leq C} |\Delta_{3j}(\vartheta)| \leq \frac{O_P(1)}{n^{3/2}} \sum_{i=1}^{n} V_i^2 = o_P(1).$$

This completes the proof. □

2.5.3 Proofs of Theorem 2.3, Theorem 2.4, Theorem 2.5, Theorem 2.6, Theorem 2.8

Recall the matrix Γ in (2.14). With $\tau_j = j/(k+1)$, direct calculation shows that

$$\Gamma^{-1} = (k+1) \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & -1 & 2 & -1 & \\ & & & -1 & 2 & \end{bmatrix}, \tag{2.57}$$

that is, Γ^{-1} is a tri-diagonal matrix with $2(k+1)$ on the diagonal, $-(k+1)$ on the super-/sub-diagonals, and 0 elsewhere.

Lemma 2.6. *Let $\ell(\tau) : (0,1) \to \mathbb{R}$ satisfy Condition 4, then for Γ defined in* (2.14)*, it holds that*

$$\lim_{k\to\infty} L^T\Gamma^{-1}L = \int_0^1 [\ell'(\tau)]^2 \,\mathrm{d}\tau, \text{ where } L = (\ell(\tau_1), \ldots, \ell(\tau_k))^T.$$

Proof. Let $\delta = 1/(k+1)$. By $\tau_j = j/(k+1)$ and (2.57), direct calculation shows that

$$\begin{aligned} L^T\Gamma^{-1}L &= (k+1)\left\{\ell^2(\tau_1) + \ell^2(\tau_k) + \sum_{j=2}^{k}[\ell(\tau_j) - \ell(\tau_{j-1})]^2\right\} \\ &= (k+1)[\ell^2(\tau_1) + \ell^2(\tau_k)] + R_k + \int_\delta^{1-\delta} [\ell'(t)]^2 \,\mathrm{d}t, \end{aligned} \tag{2.58}$$

where

$$R_k = -\frac{k+1}{2}\sum_{j=2}^{k}\int_{\tau_{j-1}}^{\tau_j}\int_{\tau_{j-1}}^{\tau_j} [\ell'(t) - \ell'(s)]^2 \,\mathrm{d}t\,\mathrm{d}s.$$

For $t, s \in [\tau_{j-1}, \tau_j]$, $|\ell'(t) - \ell'(s)| = |\int_s^t \ell''(v)\,\mathrm{d}v| \le \int_{\tau_{j-1}}^{\tau_j} |\ell''(v)|\,\mathrm{d}v$. By Schwarz inequality,

$$\max_{t,s\in[\tau_{j-1},\tau_j]} |\ell'(t) - \ell'(s)|^2 \le \left[\int_{\tau_{j-1}}^{\tau_j} |\ell''(v)|\,\mathrm{d}v\right]^2 \le \delta\int_{\tau_{j-1}}^{\tau_j} [\ell''(v)]^2\,\mathrm{d}v.$$

It then follows that

$$|R_k| \le \frac{k+1}{2}\sum_{j=2}^{k}(\tau_j - \tau_{j-1})^2 \max_{t,s\in[\tau_{j-1},\tau_j]} |\ell'(t) - \ell'(s)|^2 \le \frac{\delta^2}{2}\int_\delta^{1-\delta} |\ell''(t)|^2\,\mathrm{d}t.$$

As $k \to \infty$, $\int_\delta^{1-\delta}[\ell'(\tau)]^2\,\mathrm{d}\tau \to \int_0^1[\ell'(\tau)]^2\,\mathrm{d}\tau$, regardless of whether the latter integral is finite or infinite. The result then follows from the efficiency regularity condition of $\ell(\cdot)$, which is given as Condition 4. □

Proof of Theorem 2.3. We prove the result by constructing least squares estimates for a special linear regression model. Consider i.i.d. data $(X_1, Z_1), \ldots, (X_n, Z_n)$ from

$$Z = \mu + \frac{\dot{g}(X,\beta)^T}{s(X,\alpha)}\gamma + e, \quad e \sim \mathcal{N}(0,1) \text{ independent of } X, \tag{2.59}$$

where μ and γ are unknown parameters to be estimated from the data, whereas α and β are known. Consider the weighted least squares estimation of (μ, γ):

$$(\hat{\mu}, \hat{\gamma}_{\mathrm{LS}}(h)) = \arg\min_{u,r} \sum_{i=1}^{n} w_i \left[Z_i - u - \frac{\dot{g}(X_i, \beta)^T}{s(X_i, \alpha)} r \right]^2 \text{ with } w_i = \frac{s(X_i, \alpha)}{h(X_i)}.$$

By the classical least squares estimation theory, we have

$$\sqrt{n}[\hat{\gamma}_{\mathrm{LS}}(h) - \gamma] \Rightarrow \mathcal{N}(0, J_\beta(h)),$$

where $J_\beta(h)$ is defined in Theorem 2(i). On the other hand, by letting $w_i \equiv 1$ or equivalently $h(X_i) = s(X_i, \alpha)$, the resulting ordinary least squares estimator has asymptotic covariance matrix $J_\beta(s(\cdot, \alpha))$. Since e has a standard normal distribution, the ordinary least squares estimator is exactly the maximum likelihood estimator, which is most efficient and has the smallest variance. That is, $J_\beta(h) \geq J_\beta(s(\cdot, \alpha))$.

The case of $J_\alpha(h)$ can be treated by replacing $\dot{g}(X, \beta)$ in (2.59) with $\dot{s}(X, \alpha)$. □

Proof of Theorem 2.4. Using Γ and λ in (2.14), $I_\beta(w) = w^T \Gamma w / (w^T \lambda)^2$.

We first prove the assertion (i). Since $I_\beta(w) = I_\beta(cw)$ for $c \neq 0$, $\min_w I_\beta(w)$ is equivalent to $\min_w \{\tilde{I}_\beta(w) = w^T \Gamma w : w^T \lambda = 1\}$. For this constrained problem, the unique minimizer is $w = \Gamma^{-1}\lambda / (\lambda^T \Gamma^{-1} \lambda)$. Thus, the minimizer of $I_\beta(w)$ takes the form $w^\circ = c\Gamma^{-1}\lambda$ for some constant c.

To show the assertion (ii), plugging $w^\circ = c\Gamma^{-1}\lambda$ into $I_\beta(w) = w^T \Gamma w / (w^T \lambda)^2$, we get $I_\beta(w^\circ) = \dfrac{1}{\lambda^T \Gamma^{-1} \lambda}$. By Lemma 2.6, $\lambda^T \Gamma^{-1} \lambda \to \int_0^1 [\rho'(\tau)]^2 \, \mathrm{d}\tau$ with $\rho(\tau) = f_\varepsilon(Q_\varepsilon(\tau))$. Define the transformation $u = Q_\varepsilon(\tau)$. Then $\partial \tau / \partial u = f_\varepsilon(u)$. By the chain rule,

$$\rho'(\tau) = (\partial \rho / \partial u)(\partial u / \partial \tau) = f'_\varepsilon(u) / f_\varepsilon(u),$$
$$\int_0^1 [\rho'(\tau)]^2 \, \mathrm{d}\tau = \int_{\mathbb{R}} [f'_\varepsilon(u)]^2 / f_\varepsilon(u) \, \mathrm{d}u = \mathcal{F}(f_\varepsilon).$$

This completes the proof. □

Proof of Theorem 2.5. Write $\lambda_j = f_\varepsilon(Q_\varepsilon(\tau_j)), j = 1, \ldots, k$, and $\lambda_0 = \lambda_{k+1} = 0$. By (2.57), $\Gamma^{-1}\lambda = (k+1)(2\lambda_1 - \lambda_0 - \lambda_2, 2\lambda_2 - \lambda_1 - \lambda_3, \ldots, 2\lambda_k - \lambda_{k-1} - \lambda_{k+1})^T$. As in Theorem 2.4, write $\rho(\tau) = f(Q_\varepsilon(\tau))$. Define $u = Q_\varepsilon(\tau)$. By the proof of Theorem

2.4(ii), $\rho'(\tau) = f'_\varepsilon(u)/f_\varepsilon(u) = (\log f_\varepsilon(u))'$. Applying the chain rule once more gives that

$$\rho''(\tau) = \frac{\partial \rho'(\tau)}{\partial u}\frac{\partial u}{\partial \tau} = \frac{[\log f_\varepsilon(u)]''}{f_\varepsilon(u)}.$$

If $\log f_\varepsilon(u)$ is concave, then $[\log f_\varepsilon(u)]'' \leq 0$. By the above expression of $\rho''(\tau)$, $\rho''(\tau) \leq 0$ hence $\rho(\tau)$ is concave, which implies that $2\lambda_j - \lambda_{j-1} - \lambda_{j+1} \geq 0$ hence $\Gamma^{-1}\lambda$ satisfies the non-negative constraint.

Conversely, assume $\Gamma^{-1}\lambda \geq 0$ for every k, then for every $k \in \mathbb{N}$, it is easy to verify that if $i+j$ is even, then

$$\rho\left(\frac{i+j}{2(k+1)}\right) \geq \frac{1}{2}\rho\left(\frac{i}{k+1}\right) + \frac{1}{2}\rho\left(\frac{j}{k+1}\right).$$

Continuity argument then shows that for every $\tau_1, \tau_2 \in (0,1)$, it follows that $\rho((\tau_1+\tau_2)/2) \geq \rho(\tau_1)/2 + \rho(\tau_2)/2$, whence $\rho(\tau)$ is concave, consequently, $\log f_\varepsilon(u)$ must be concave. □

Proof of Theorem 2.6. Using Ξ and ζ in (2.17), we can write $I_\alpha(w) = \dfrac{w^T \Xi w}{(w^T\zeta)^2}$.

To prove (i), for even k, under Condition 3 and Condition 2(i), we have $Q_\varepsilon(\tau_j) \neq 0$ for all j. The result follows from the same argument in the proof for Theorem 2.4(i).

For odd k, $Q_\varepsilon(k^*) = 0$ and we can drop τ_{k^*} from Ξ and ζ to form Ξ^* and ζ^*. Let $\bar{w} = (w_1, \ldots, w_{k^*-1}, w_{k^*+1}, \ldots, w_k)^T$. Then $I_\alpha(w) = I_\alpha(\bar{w}) = \bar{w}^T\Xi^*\bar{w}/(\bar{w}^T\zeta^*)^2$ has minimizer $\bar{w} = c\Xi^{*-1}\zeta^*$.

The symmetry of the solution w°_α is clear from part (ii), by examining the forms of $\Xi, \Xi^*, \zeta, \zeta^*$ and using Condition 3.

We next prove (ii). For even k, from $w^\circ_\alpha = c\Xi^{-1}\zeta$ and $I_\alpha(w) = w^T\Xi w/(w^T\zeta)^2$, we have $I_\alpha(w^\circ_\alpha) = 1/(\zeta^T\Xi^{-1}\zeta)$. Let $P = diag(Q_\varepsilon(\tau_1), \ldots, Q_\varepsilon(\tau_k))$. With Γ in (2.14), we have $\Xi = P\Gamma P$ and $\zeta^T\Xi^{-1}\zeta = (P^{-1}\zeta)^T\Gamma^{-1}(P^{-1}\zeta)$. For $s(\tau) = Q_\varepsilon(\tau)f_\varepsilon(Q_\varepsilon(\tau))$, $P^{-1}\zeta = (s(\tau_1), \ldots, s(\tau_k))^T$. Thus, by Lemma 2.6, $\zeta^T\Xi^{-1}\zeta \to \int_0^1 [s'(\tau)]^2 \,\mathrm{d}\tau$.

For odd k, $I_\alpha(w^\circ_\alpha) = 1/(\zeta^{*T}\Xi^{*-1}\zeta^*)$. Recall P defined above and Γ in (2.14). Define P^* and Γ^* by removing the k^*-th row and k^*-th column from P and Γ. It can be verified that

$$\frac{\Gamma^{*-1}}{k+1} = \left(\begin{array}{cccccccccc}
2 & -1 & & & & & & & & \\
-1 & 2 & -1 & & & & & & & \\
 & \ddots & \ddots & \ddots & & & & & & \\
 & & -1 & 2 & -1 & & & & & \\
 & & & -1 & \frac{3}{2} & -\frac{1}{2} & & & & \\
 & & & & -\frac{1}{2} & \frac{3}{2} & -1 & & & \\
 & & & & & -1 & 2 & -1 & & \\
 & & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & & -1 & 2 & -1 \\
 & & & & & & & & -1 & 2
\end{array}\right)\begin{array}{c}
\\ \\ \\ \\ (k^*-1)\text{-th} \\ k^*\text{th} \\ \\ \\ \\ \\
\end{array}$$

Note that $\Xi^* = P^*\Gamma^*P^*$. Thus, using the above inversion matrix and by the same argument in Lemma 2.6, it can be shown that

$$\zeta^{*T}\Xi^{*-1}\zeta^* = (P^{*-1}\zeta^*)^T\Gamma^{*-1}(P^{*-1}\zeta^*) \to \int_0^1 [s'(\tau)]^2\,\mathrm{d}\tau.$$

The result then follows from $\int_0^1 [s'(\tau)]^2\,\mathrm{d}\tau = \mathcal{G}(f_\varepsilon)$ via the transformation $u = Q_\varepsilon(\tau)$. □

Proof of Theorem 2.8. By Theorem 2.1, $\|\hat{\alpha}^0 - \alpha\| = O_P(n^{-1/2})$ and $\|\hat{\beta}^0 - \beta\| = O_P(n^{-1/2})$. Since k is fixed, in order to prove $\hat{w}_j^* = w_j + o_P(1)$, it suffices to prove (i) $\tilde{Q}_\varepsilon(\tau) = Q_\varepsilon(\tau) + o_P(1)$ for each fixed $\tau \in (0,1)$; and (ii) $\tilde{f}_\varepsilon(u) = f_\varepsilon(u) + o_P(1)$ for each fixed u.

First, we prove $\tilde{Q}_\varepsilon(\tau) = Q_\varepsilon(\tau) + o_P(1)$. For $\hat{\varepsilon}_i$ in (2.20), we have

$$\hat{\varepsilon}_i = \varepsilon_i + \Delta_i, \Delta_i = \frac{g(X_i,\beta) - g(X_i,\hat{\beta}_0)}{s(X_i,\hat{\alpha}^0)} + \frac{s(X_i,\alpha) - s(X_i,\hat{\alpha}^0)}{s(X_i,\hat{\alpha}^0)}\varepsilon_i.$$

In the proof of Theorem 2.7, it has been shown that, with probability approaching to one, $s(X_i,\hat{\alpha}^0) > c/2 > 0$ for all i. Let V_i be defined as in (2.5). Similar to (2.56),

$$|\Delta_i| = \frac{O_P(1)}{\sqrt{n}}V_i(1 + |\varepsilon_i|). \tag{2.60}$$

Since $V_i \in \mathcal{L}^\delta$ (Condition 2(v)), $V_i \varepsilon_i \in \mathcal{L}^\delta$ (Condition 6(i)), and Lemma 2.1(i), $\max_{1\le i\le n} V_i(1+|\varepsilon_i|) = O_P(n^{1/\delta})$. Thus, $\hat{\varepsilon}_i = \varepsilon_i + o_P(1)$ uniformly for all i, and consequently $\tilde{Q}_\varepsilon(\tau) = \bar{Q}_\varepsilon(\tau) + o_P(1)$, where $\bar{Q}_\varepsilon(\tau)$ is the τ-th sample quantile of $\varepsilon_1, \ldots, \varepsilon_n$. By standard theory of sample quantiles, $\bar{Q}_\varepsilon(\tau) = Q_\varepsilon(\tau) + o_P(1)$. Therefore $\tilde{Q}_\varepsilon(\tau) = Q_\varepsilon(\tau) + o_P(1)$.

Next, we prove $\tilde{f}_\varepsilon(u) = f_\varepsilon(u) + o_P(1)$ for each fixed u. Since K is Lipschitz, it follows that

$$\tilde{f}_\varepsilon(u) = \frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{u-\varepsilon_i}{b_n}\right) + \frac{O(1)}{nb_n^2}\sum_{i=1}^{n} |\Delta_i|$$

By standard theory of nonparametric kernel density estimation,

$$\frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{u-\varepsilon_i}{b_n}\right) = f_\varepsilon(u) + o_P(1). \tag{2.61}$$

By (2.60),

$$\frac{1}{nb_n^2}\sum_{i=1}^{n} |\Delta_i| = O_P(1)\frac{1}{nb_n^2} n^{-1/2}\sum_{i=1}^{n} V_i(1+|\varepsilon_i|) = O_P((\sqrt{n}b_n^2)^{-1}) = o_P(1)$$

in view of $b_n \propto n^{-1/5}$ (Condition 6(ii)). This completes the proof. □

Chapter 3

Estimating Parameters in Nonlinear Heteroscedastic Models: Numerical Studies and Simulation

3.1 Introduction

To implement the adaptive estimation method proposed in Chapter 2, we are required to solve the following optimization problem:

$$\min_{\beta,\alpha,q_1,\ldots,q_k} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k h_i^{-1} \rho_{\tau_k}(y_i - g(x_i,\beta) - s(x_i,\alpha)q_k), \tag{3.1}$$

where $\beta \in \mathbb{R}^p$ and $\alpha \in \mathbb{R}^q$ are the regression parameter and the dispersion parameter respectively, $(q_1,\ldots,q_k)'$ is the nuisance parameter, representing the true quantiles of the error distribution. In (3.1), the weights $\{w_k\}$ and $\{h_i\}$ are known and satisfy $w_k = w_{K+1-k}, w_k > 0, k = 1,\ldots,K,\ h_i > 0, i = 1,\ldots,n$. Although our method can be generalized to asymmetric error case, throughout the chapter we will impose the symmetric error condition to ε, which implies an implicit constraint: $q_k = q_{K+1-k}, k = 1,\ldots,K$. This condition typically is fulfilled by manually symmetrizing an preliminary result or coding this constraint into the optimization routines explicitly.

Generally speaking, the optimization problem (3.1) is very challenging. The reason is threefold:

1. The non-linearity of the regression function $g(x, \cdot)$ and the dispersion function $s(x, \cdot)$ characterizes (3.1) a nonlinear optimization problem, which in general is considered to be very hard, in particular, no algorithm is guaranteed to converge theoretically for a general problem such as (3.1).

2. The non-differentiability of objective functions $\rho_{\tau_k}(\cdot)$ prohibits direct application of any efficient derivative-based optimization algorithm.

3. The multiplicative structure $s(x, \alpha)q$ typically makes α and $(q_1, \ldots, q_K)'$ non-identifiable, especially when they are minimized jointly (simultaneously). In addition, such structure makes the problem a non-convex optimization program.

To the best of our knowledge, Challenge 1 does not admit any definite answers — to discuss under which circumstances does (3.1) allow for a global (or local) minimizer and its convergence property is more a mission for operations researchers than statisticians. Therefore, although a general algorithm for solving (3.1) is proposed in Section 3.3, it is not warranted that this algorithm would work well for arbitrary specification of g and s. In this chapter, our main goal is to address Challenge 2 and Challenge 3 above and the details can be found in Section 3.3.

The remaining of this chapter is organized as follows: in Section 3.2 we review some representative algorithms used to solve regression quantiles. In literature, these algorithms are described under the linear single quantile regression case, in lieu of the nonlinear composite quantile regression case. Thus the leading purpose of this section is to discuss how these state-of-the-art algorithms may be generalized to the nonlinear composite quantile regression case. Section 3.3 gives a detailed description of our "MMLP" algorithm for solving the program (3.1). The name "MMLP" clearly indicates that our algorithm is a hybrid of the MM algorithm and the linear programming method. For completeness, a brief introduction to the MM algorithm is given at the beginning of this section. In Section 3.4 we will investigate the numerical performance of the MMLP algorithm proposed in Section 3.3 and recommend a modified version of MMLP for practical applications. In addition to this topic, a self-start procedure that computes reasonable initialized parameter values, including the nuisance parameters is also reported. In Section 3.5, we compare our method with several existing estimation methods by running Monte Carlo studies under two nonlinear heteroscedastic models.

3.2 Literature Review

To open the discussion of computational aspects of quantile regression, let's first focus on the classical linear model

$$y_i = x_i'\beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.2}$$

with $\{\varepsilon_i\}$ i.i.d. $\sim F$. By definition, for $\tau \in (0,1)$, the τ-th regression quantile $\hat{\beta}(\tau)$ is the solution to the following optimization problem:

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'b), \tag{3.3}$$

where $\rho_\tau(z) = (\tau - I(z < 0))z = \tau z^+ + (1-\tau)z^-$ is the well-known "check" function. When $\tau = 0.5$, (3.3) reduces to the famous *Least Absolute Deviation* problem: $\min_b \sum_{i=1}^{n} |y_i - x_i'b|$. Historically, the method of least absolute deviation (LAD) couldn't receive similar popularity as its closely related sibling: the method of least squares (LS), mainly due to by contrast to the LS method, it doesn't admit a closed form solution. For a historical account for this interesting topic, see Portnoy and Koenker (1997) and Chapter 1 of Koenker (2005).

It was not until the introduction of the simplex algorithm in the late 1940s that a practical, general method for computing (3.3) was made available. It is well-known that the problem (3.3) may be formulated as the linear program:

$$\min\{0'b + \tau e'u + (1-\tau)e'v : y = Xb + u - v, (u,v) \in \mathbb{R}_+^{2n}\} \tag{3.4}$$

and has dual formulation

$$\max\{y'a : X'a = (1-\tau)X'e, a \in [0,1]^n\}. \tag{3.5}$$

In the above formulations, $y = (y_1, \dots, y_n)' \in \mathbb{R}^{n\times 1}, X' = [x_1, \dots, x_n] \in \mathbb{R}^{n\times p}$ and e is an n-vector of all ones.

There are two state-of-the-art algorithms in literature to solve (3.4) and (3.5), namely, the *simplex methods* and the *interior point methods*. There are, however, other methods to solve (3.3) in addition to these two, for example, the finite smoothing algorithm proposed by Chen (2007). The computational aspects of quantile regression are also widely discussed under more complicated models or as a leading example of some innovative optimization algorithms, in this respect, see Fan et al. (2014), Fan et al. (2016) and Liu et al. (2016). In the following, we will briefly review the classical simplex methods and the interior point methods.

The simplex-type algorithms are guided by very strong geometric intuition — the basic idea is to look for vertex solution across the surface of the diamond-shaped constraint set. Simply speaking, the algorithm is divided into two phases: in phase I an initial feasible vertex of the problem is found, and then in phase II, we proceed from one such vertex to another until optimality is achieved. In quantile regression problems, the implementation of phase I is typically easy, and the phase II is implemented in the path-breaking algorithm of Barrodale and Roberts (1974). The innovative part of the Barrodale and Roberts (1974) algorithm is that rather than simply adopting the conventional simplex strategy of traveling only as far as the next vertex, they proposed to continue in the original direction as long as doing so continued to reduce the value of the objective function. In this way they dramatically reduce the number of simplex pivot operations required when the basis changes. For details of this algorithm and its application in quantile regression problems, we refer readers to Barrodale and Roberts (1974) and Section 6.2 of Koenker (2005).

The simplex algorithm is one strategy of the *exterior point methods*, in the sense that it relies on an iterative path along the exterior of the constraint set. By contrast, the interior point methods “work systematically from the interior of the admissible region and employ some barrier function as a guide to avoid crossing the boundary” (Koenker (2005), p.191). In quantile regression problems, the algorithm is the Frisch-Newton method proposed by Portnoy and Koenker (1997). Specifically, the formulation of Portnoy and Koenker (1997)’s algorithm corresponding to the dual formulation (3.5) employs the logarithm barrier function

$$B(a, s, \mu) = y'a + \mu \sum_{i=1}^{n} (\log a_i + \log s_i),$$

which should be maximized subject to the constraints $X'a = (1 - \tau)X'e$ and $a + s = e$. A Newton step is then followed to bring down the objective value of $B(a, s, \mu)$. Throughout, the *barrier parameter* μ is gradually decreased to 0 so that a *central path* is traced and along which the iterations converge to the optimal solution. There are many technical issues need to be addressed in implementing the interior point method to solve (3.4) and (3.5). Among others, Portnoy and Koenker (1997) gives the primal-dual formulation of the interior point algorithm, Mehrotra (1992) provides the general guidance of how to decrease μ and handle the nonlinearity during the Newton step. Further details about the history, convergence property and advantages of the interior point algorithm may be found in a series of excellent expository papers of Wright (1992) and Wright (2005).

Let's conclude this section by making some comparison between the simplex algorithm and the interior point algorithm. As pointed out by Portnoy and Koenker (1997), for problems of modest size, both algorithms are competitive with least squares in terms of computational speed. For small problems, the simplex implementation is the clear winner, but the interior point algorithm does considerably better than simplex at larger sample sizes. More importantly, as noted by Wright (1992), unlike the simplex method, the interior point techniques can obviously be applied to nonlinear optimization problems. In particular, while it seems difficult to directly apply the simplex method to nonlinear quantile regression problems, the interior point method has been successfully generalized to these problems, see Koenker and Park (1996) and Section 6.6 of Koenker (2005).

3.3 An "MMLP" Algorithm for Nonlinear Composite Quantile Regression

This nonlinear quantile regression problem:

$$\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - g_0(x_i, \theta)), \tag{3.6}$$

where g_0 is some known function that is nonlinear in the parameter θ, is clearly harder to solve than its linear counterpart (3.3). As pointed out at the end of Section 3.2, (3.6) may be solved by the interior point algorithm proposed by Koenker and Park (1996). On the other hand, Hunter and Lange (2000) introduce the MM algorithm to the nonlinear quantile regression community and soon becomes a strong rival to the interior points methods (Kai et al. (2010) have successfully applied this MM algorithm to a nonparametric composite quantile regression problem, though the optimization problem studied there is linear in parameters.). In Section 5 of Hunter and Lange (2000), the authors made a comprehensive numerical comparison between the MM algorithm and the interior point algorithm, concluding that "although neither of the two algorithms outpaces the other or produces more accurate solutions in Table 1,* the MM algorithm is the more stable of the two". Due to its better numerical stability, as well as its conceptual simplicity and ease of implementation, we choose to extend the MM algorithm, instead of the interior point algorithm to the current setting (3.1) that we are interested in.

The nature of the MM algorithm can be briefly described as follows: suppose we want to minimize the objective function $L(\theta) : \mathbb{R}^p \to \mathbb{R}$. If $\theta^{(k)}$ denotes the current iterate in finding the minimum point, then the MM algorithm proceeds in two steps. First, we create a surrogate function $Q(\theta|\theta^{(k)}) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ satisfying

$$\begin{aligned} &Q(\theta^{(k)}|\theta^{(k)}) = L(\theta^{(k)}) \\ &Q(\theta|\theta^{(k)}) \geq L(\theta) \text{ for all } \theta. \end{aligned}$$

The function $Q(\theta|\theta^{(k)})$ is said to *majorize* $L(\theta)$ at $\theta^{(k)}$ (aka, the *surrogate function*).

*In this table, 14 different nonlinear models are investigated and the thousands of FLOPs required until convergence for two algorithms are reported.

In the second step, we choose $\theta^{(k+1)}$ to minimize $Q(\theta|\theta^{(k)})$ with respect to θ. In general, it is a challenge to construct a good surrogate function which simultaneously majorizes $L(\theta)$ at $\theta^{(k)}$ and is itself easy to minimize.

Since the check function $\rho_\tau(z)$ is not differentiable at $z = 0$, so instead of seeking for the surrogate function for $\rho_\tau(z)$, Hunter and Lange (2000) suggested seeking for the surrogate function for the perturbation of $\rho_\tau(z)$, which is $\rho_\tau^\varepsilon(z) := \rho_\tau(z) - \frac{\varepsilon}{2}\log(\varepsilon + |z|)$. It is remarkable that for any $\varepsilon > 0$, $\rho_\tau^\varepsilon(z)$ is smooth everywhere in $\mathbb{R}^1$. In Hunter and Lange (2000), they showed that the $\rho_\tau^\varepsilon(z)$ is majorized at $z^{(k)}$ by the quadratic function

$$\zeta_\tau^\varepsilon(z|z^{(k)}) = \frac{1}{4}\left[\frac{z^2}{\varepsilon + |z^{(k)}|} + (4\tau - 2)z + c\right],$$

where c is a constant chosen so that $\zeta_\tau^\varepsilon(z^{(k)}|z^{(k)}) = \rho_\tau^\varepsilon(z^{(k)})$. The relations between $\rho_\tau(z)$, $\rho_\tau^\varepsilon(z)$ and $\zeta_\tau^\varepsilon(z|z^{(k)})$ are shown in Figure 3.1.

As emphasized before, any joint minimization procedure to the problem (3.1) may bring numerical stability issues. Therefore, in order to deploy the MM algorithm, let's assume the values of nuisance parameters $q_1, \ldots, q_K$ are known temporarily, and the goal now is to solve the following nested problem of (3.1):

$$\min_{\beta,\alpha} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k h_i^{-1} \rho_{\tau_k}(y_i - g(x_i, \beta) - s(x_i, \alpha) q_k), \tag{3.7}$$

To derive the MM algorithm for (3.7) using the surrogate functions $\zeta_{\tau_k}^\varepsilon, k = 1, \ldots, K$, let $\theta \in \mathbb{R}^{p+q}$ be the vector of parameters $(\beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_q)'$ and assume we are currently at the $(m+1)$-st iterate. The MM algorithm operates by minimizing the majorizer

$$Q_\varepsilon(\theta|\theta^{(m)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k h_i^{-1} \zeta_{\tau_k}^\varepsilon(r_{ik}|r_{ik}^{(m)})$$

with respect to θ. In the above expression, $r_{ik} = r_{ik}(\theta) = y_i - g(x_i, \beta) - s(x_i, \alpha) q_k$,

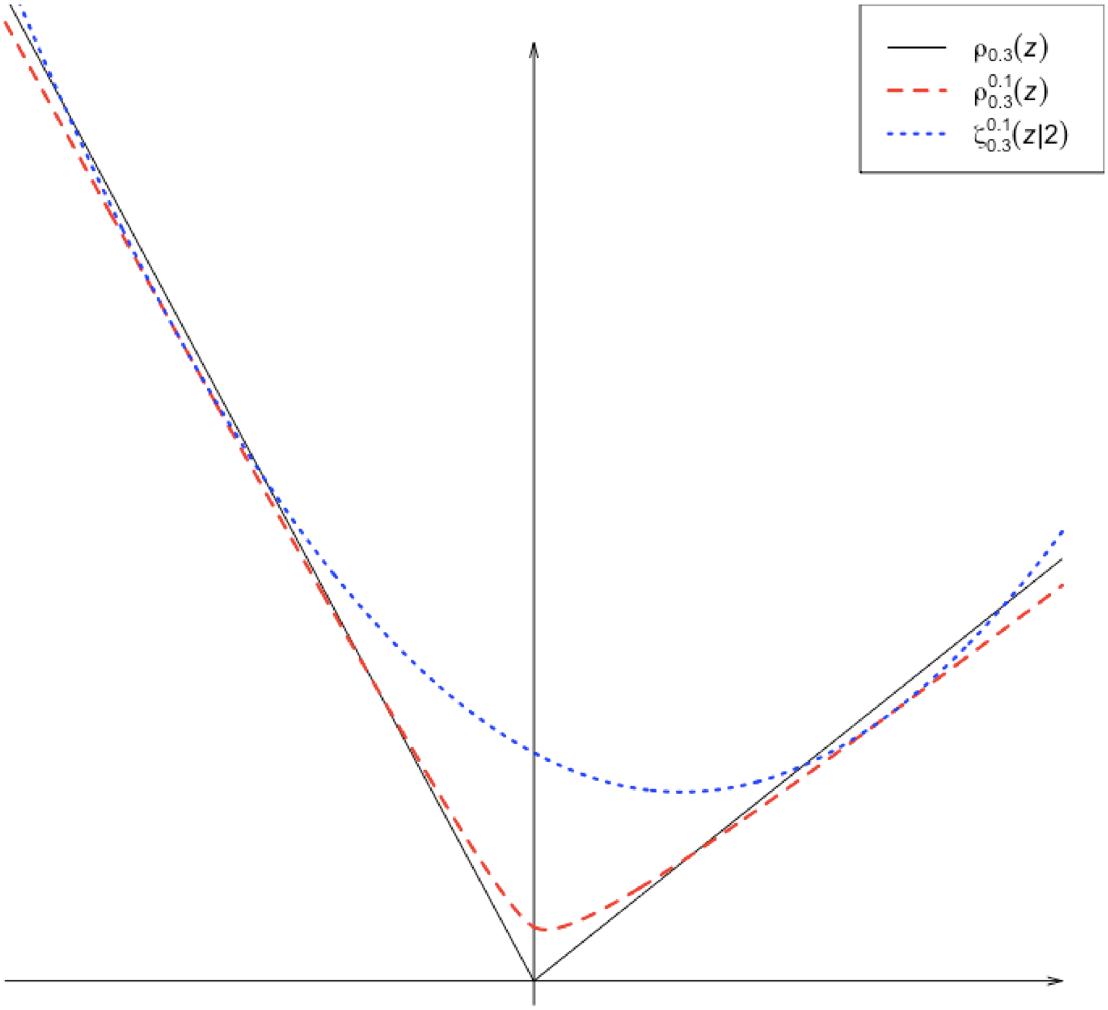

Figure 3.1. *The original check function $\rho_\tau(z)$ (black solid), its perturbation function $\rho_\tau^\varepsilon(z)$ (red dashed), and the surrogate function $\zeta_\tau^\varepsilon(z|z^{(k)})$ (blue dotted) of $\rho_\tau^\varepsilon(z)$ at $z^{(k)}$. In this picture, $\tau = 0.3, \varepsilon = 0.1$ and $z^{(k)} = 2$.*

$i = 1, \ldots, n, k = 1, \ldots, K$. Direct calculation shows that

$$\frac{\partial Q_\varepsilon(\theta|\theta^{(m)})}{\partial \beta} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left[\frac{r_{ik}}{\varepsilon + |r_{ik}^{(m)}|} + 2\tau_k - 1\right]\dot{g}(x_i, \beta),$$

$$\frac{\partial Q_\varepsilon(\theta|\theta^{(m)})}{\partial \alpha} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left[\frac{r_{ik}}{\varepsilon + |r_{ik}^{(m)}|} + 2\tau_k - 1\right] q_k \dot{s}(x_i, \alpha),$$

and

$$
\begin{aligned}
\frac{\partial Q_{\varepsilon}^{2}(\theta|\theta^{(m)})}{\partial\beta^{2}} =& \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left\{\frac{1}{\varepsilon+|r_{ik}^{(m)}|}\dot{g}(x_i,\beta)\dot{g}(x_i,\beta)'\right.\\
&\left.+\left[\frac{r_{ik}}{\varepsilon+|r_{ik}^{(m)}|}+2\tau_k-1\right]\ddot{g}(x_i,\beta)\right\},\\
\frac{\partial Q_{\varepsilon}^{2}(\theta|\theta^{(m)})}{\partial\alpha^{2}} =& \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left\{\frac{1}{\varepsilon+|r_{ik}^{(m)}|}q_k^2\dot{s}(x_i,\alpha)\dot{s}(x_i,\alpha)'\right.\\
&\left.+\left[\frac{r_{ik}}{\varepsilon+|r_{ik}^{(m)}|}+2\tau_k-1\right]q_k\ddot{s}(x_i,\alpha)\right\},\\
\frac{\partial Q_{\varepsilon}^{2}(\theta|\theta^{(m)})}{\partial\alpha\partial\beta} =& \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\frac{1}{\varepsilon+|r_{ik}^{(m)}|}q_k\dot{g}(x_i,\beta)\dot{s}(x_i,\alpha)'.
\end{aligned}
$$

In line with Hunter and Lange (2000), after omitting the second-order partial derivatives $\ddot{g}$ and $\ddot{s}$ from the above results, the Newton direction $[\delta_{\beta}',\delta_{\alpha}']'$ can be found by solving the following linear system:

$$
\begin{aligned}
&\begin{bmatrix}
\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\dot{g}(x_i,\beta^{(m)})\dot{g}(x_i,\beta^{(m)})'}{\varepsilon+|r_{ik}^{(m)}|} & \displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{q_k\dot{g}(x_i,\beta^{(m)})\dot{s}(x_i,\alpha^{(m)})'}{\varepsilon+|r_{ik}^{(m)}|}\\
\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{q_k\dot{s}(x_i,\alpha^{(m)})\dot{g}(x_i,\beta^{(m)})'}{\varepsilon+|r_{ik}^{(m)}|} & \displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{q_k^2\dot{s}(x_i,\alpha^{(m)})\dot{g}(x_i,\alpha^{(m)})'}{\varepsilon+|r_{ik}^{(m)}|}
\end{bmatrix}
\begin{bmatrix}\delta_{\beta}\\ \delta_{\alpha}\end{bmatrix}\\
=&\begin{bmatrix}
\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left[\frac{r_{ik}}{\varepsilon+|r_{ik}^{(m)}|}+2\tau_k-1\right]\dot{g}(x_i,\beta)\\
\displaystyle\sum_{i=1}^{n}\sum_{k=1}^{K} w_k h_i^{-1}\left[\frac{r_{ik}}{\varepsilon+|r_{ik}^{(m)}|}+2\tau_k-1\right]q_k\dot{s}(x_i,\alpha)
\end{bmatrix}. \qquad (3.8)
\end{aligned}
$$

Denote the solution to the above system by $\Delta_{\varepsilon}^{(m)}$. To guarantee the value of the surrogate function is decreased, following Hunter and Lange (2000), we take an appropriate fractional step size $\alpha^{(m)}\in(0,1]$:

$$
\theta^{(m+1)}=\theta^{(m)}+\alpha^{(m)}\Delta_{\varepsilon}^{(m)}, \qquad (3.9)
$$

where

$$\alpha^{(m)} = \max\{2^{-\nu} : Q_\varepsilon(\theta^{(m)} + 2^{-\nu}\Delta_\varepsilon^{(m)}|\theta^{(m)}) < Q_\varepsilon(\theta^{(m)}|\theta^{(m)}), \nu \in \mathbb{N}\}.$$

Suppose we have obtained a convergent solution $(\beta_*', \alpha_*')'$ to the program (3.7), the next step is to update the parameters $q_1, \ldots, q_K$. Denote $y_i - g(x_i, \beta_*)$ by $\tilde{y}_i$, $s(x_i, \alpha_*)$ by $\tilde{x}_i$, $i = 1, \ldots, n$, and the goal is to solve:

$$\min_{q_1,\ldots,q_K} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k h_i^{-1} \rho_{\tau_k}(\tilde{y}_i - \tilde{x}_i q_k). \tag{3.10}$$

It can be easily identified that the program (3.10) has the following linear programming formulation:

$$\begin{aligned} &\min_{\xi,u,v} 0'\xi + c_1'u + c_2'v, \text{ subject to} \\ &\begin{bmatrix} \tilde{X}_1 & I & 0 & \cdots & 0 & -I & 0 & \cdots & 0 \\ \tilde{X}_2 & 0 & I & \cdots & 0 & 0 & -I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{X}_K & 0 & 0 & \cdots & I & 0 & 0 & \cdots & -I \end{bmatrix} \begin{bmatrix} \xi \\ u \\ v \end{bmatrix} = \begin{bmatrix} \tilde{y} \\ \tilde{y} \\ \vdots \\ \tilde{y} \end{bmatrix} \\ &u \geq 0, \quad v \geq 0. \end{aligned} \tag{3.11}$$

In the above formulation, $\xi = (q_1, \ldots, q_K)' \in \mathbb{R}^{K\times 1}$, $c_1 = (w_1\tau_1, \ldots, w_K\tau_K)' \otimes (h_1^{-1}, \ldots, h_n^{-1})' \in \mathbb{R}^{(Kn)\times 1}$, $c_2 = (w_1(1-\tau_1), \ldots, w_K(1-\tau_K))' \otimes (h_1^{-1}, \ldots, h_n^{-1})' \in \mathbb{R}^{(Kn)\times 1}$, $u = (u_1', \ldots, u_K')' \in \mathbb{R}_+^{(Kn)\times 1}$, $v = (v_1', \ldots, v_K')' \in \mathbb{R}_+^{(Kn)\times 1}$. For each $k \in \{1, \ldots, K\}$, $u_k = (u_{1k}, \ldots, u_{nk})' \in \mathbb{R}_+^{n\times 1}$, $v_k = (v_{1k}, \ldots, v_{nk})' \in \mathbb{R}_+^{n\times 1}$. I is an n-by-n identity matrix, $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)' \in \mathbb{R}^{n\times 1}$, and $\tilde{X}_k$ is an n-by-K matrix with its k-th column $(\tilde{x}_1, \ldots, \tilde{x}_n)'$ and all the other entries zero. As a linear program, (3.11) can be handled by directly calling the optimization routine, say, the MATLAB function `linprog`. If it is presumed that the error term is symmetric, then we may add further constraints $\xi_j = \xi_{K+1-j}, j = 1, \ldots, \lfloor K/2 \rfloor$ into (3.11) to address the symmetry condition.

Now we have seen why our algorithm is termed as "MMLP": it consists of alternating steps of updating the parameter of interest $(\beta', \alpha')'$, given the nuisance parameter $(q_1, \ldots, q_K)'$, using the MM algorithm, and of updating the nuisance parameter $(q_1, \ldots, q_K)'$, given $(\beta', \alpha')'$, using the LP method. The algorithm is terminated when we repeat such outer iterates for $\mathbb{C}$ times. Intuitively, it is expected that through the outer iterates between MM steps and LP steps, the parameter of interest would converge to the optimal solutions (at least to a local optimum). Although it is our belief that a better estimates of $(q_1, \ldots, q_K)'$ would result in more accurate estimate of $(\beta', \alpha')'$, there is no guarantee for such convergence in theory, partially may be due to the fact that updating $(q_1, \ldots, q_K)'$ doesn't necessarily decrease the objective value of the surrogate function, compared to that at the completion of the last MM step. We will come back to discuss this issue later in Section 3.4.

The MMLP algorithm for solving the problem (3.1) derived in this section can be summarized as the pseudo code in Table 1.

Algorithm 1 General MMLP algorithm for nonlinear heteroscedastic composite quantile regression, with the weights $\{w_k\}$ and $\{h_i\}$ as fixed inputs.

```
Initialize $\theta^{(0)} = (\beta^{(0)\prime}, \alpha^{(0)\prime})'$, $\xi^{(0)} = (q_1^{(0)}, \ldots, q_K^{(0)})'$. Set the outer iterate index $\ell = 1$.
repeat
    MM Step: Set the starting value for the MM step as $\theta^{(\ell-1)}$, assume $\xi^{(\ell-1)}$ to be known constants. Update $\theta^{(\ell-1)}$ to $\theta^{(\ell)}$ by completing the MM steps (3.8) and (3.9) multiple times, until the MM algorithm's convergence condition is met.
    LP Step: Evaluate $\tilde{y}$, $\tilde{X}_1, \ldots, \tilde{X}_K$ in (3.11) using $\theta^{(\ell)}$. Solve the linear program (3.11) to update $\xi^{(\ell-1)}$ to $\xi^{(\ell)}$.
    Increase $\ell$ to $\ell + 1$.
until $\ell = \mathbb{C}$
```

3.4 Further Discussions on the Algorithm

In this section, two important aspects related to the MMLP algorithm proposed in Section 3.3 will be discussed. The method to pick up a reasonable starting values, under the assumption of symmetric errors, is studied in Section 3.4.1. In Section 3.4.2, we investigate two ramifications of the Algorithm 1, and provide some guidance on which algorithm should be chosen for different purposes.

3.4.1 Parameters Initialization

Since the core of our Algorithm 1 is MM algorithm, which heavily relies on the choice of initial values, it is natural to expect that good initial values would speed up the convergence, and increase the chance of achieving the global minimizer. Therefore it is critical for us to develop some automatic procedure that provides users with reasonably good initial values of β, α, as well as $(q_1, \ldots, q_K)'$.

Recall the model under consideration takes the form:

$$y_i = g(x_i, \beta) + s(x_i, \alpha)\varepsilon_i, \quad i = 1, \ldots, n,$$

where $s(\cdot, \alpha) > 0$, $\{\varepsilon_i\}$ i.i.d. $\sim \varepsilon$, whose distribution function is F. Usually we are willing to further assume that F is symmetric about 0, i.e., $F(x) = 1 - F(-x)$ for any $x \in \mathbb{R}^1$. For the ease of reference, we will call this condition as *symmetric error* condition. Throughout this chapter, without explicit specification, we tacitly assume the symmetric error condition holds.

Since we don't impose conditions on the moments of ε, but do have $Q_\varepsilon(0.5) = 0$ due to the symmetric error condition, the parameter β can be initialized by solving the nonlinear least absolute deviation problem:

$$\hat{\beta}^{(0)} = \arg\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} |y_i - g(x_i, b)|$$

Other estimation method, such as least squares, may also be used to initialize β, but for robustness consideration, we recommend using $\hat{\beta}^{(0)}$. After $\hat{\beta}^{(0)}$ is obtained, we are able to calculate the residuals $r_i = y_i - g(x_i, \hat{\beta}^{(0)}), i = 1, \ldots, n$. It is then conventional to postulate

$$|r_i| \approx s(x_i, \alpha)\varepsilon_i,$$

or equivalently,

$$\log |r_i| \approx \log s(x_i, \alpha) + \log |\varepsilon_i| \tag{3.12}$$

Keep in mind that q_k has the direct interpretation that it is the τ_kth quantile of ε, thus if we are able to uncover the relationship between the quantiles of ε and $\log |\varepsilon|$, then (3.12) suggests a plausible quantile regression framework to estimate q_k, $k = 1, \ldots, K$. Under the symmetric error condition, we have the following proposition:

Proposition. *For any absolute continuous random variable ε whose distribution is symmetric about* 0 *and $\tau \in (0,1)$, we have*

$$Q_\varepsilon\left(\frac{1+\tau}{2}\right) = \exp[Q_{\log|\varepsilon|}(\tau)]. \tag{3.13}$$

Proof. Denote $Q_{\log|\varepsilon|}(\tau)$ by ξ. By definition,

$$\tau = P[\log|\varepsilon| \leq \xi] = P[|\varepsilon| \leq \exp(\xi)] = 2P[\varepsilon \leq \exp(\xi)] - 1, \tag{3.14}$$

which implies $P[\varepsilon \leq \exp(\xi)] = (1+\tau)/2$, i.e., $Q_\varepsilon((1+\tau)/2) = e^\xi = \exp[Q_{\log|\varepsilon|}(\tau)]$. The symmetric error condition is used in the last equality of (3.14). □

In light of (3.12) and the proposition above, for k such that $\tau_k > 1/2$, we can perform a nonlinear $(2\tau_k - 1)$-quantile regression:

$$Q_{\log|r_i|}(2\tau_k - 1|x_i) = \log s(x_i, \alpha) + \gamma_k,$$

in which we treat $\log|r_i|, i = 1, \ldots, n$ as responses and γ_k as the intercept. When α and γ_k is identifiable with each other (i.e., both of them are estimable without confusion), q_k can be initialized as

$$q_k^{(0)} = \exp(\hat{\gamma}_k), \text{ if } k \text{ satisfies } \tau_k > 1/2.$$

For those k such that $\tau_k < 1/2$, q_k is initialized by symmetry. Additionally, if the function $\log s(x_i, \alpha)$ is linear in α, then α can be initialized simultaneously with q_k. Otherwise, the special structure of the dispersion function $s(\cdot, \alpha)$ needs to be exploited and some ad-hoc self starting procedure needs to be designed, see Venables and Ripley (2003), pp.216–217.

To solve the nonlinear quantile regression problem (3.12), we can call the R function `nlrq` in the `quantreg` package directly. The `nlrq` function is based on the interior point ideas described in Koenker and Park (1996).

It is notable that the general initialization method proposed above may not be the best way for a specific problem. Whenever possible, we should take advantage of the special structure of the model under consideration. In particular, we need to take advantage of linear parameters, as pointed out by Venables and Ripley (2003, pp.218–220) and is termed as the *partially linear* algorithm. It can be much more stable than methods that do not take advantage of linear parameters, it requires fewer initial values and it can often converge from poor starting positions where other procedures fail. To illustrate this point under our setting, we describe the initialization procedure here for the Model 2 in Section 3.5. Note that before reparametrization, both the regression function g and the dispersion function s are linear in parameters, hence it is the classic location-scale model (see Jurečková and Procházka (1994), Koenker and Zhao (1994)). Because of this nice structure, none initial values need to be provided to complete the initialization stage. Specifically, we can do:

Step 1: Perform a linear LAD regression to obtain the initial values $\beta^{(0)}$, as well as the residuals $r_t = y_t - x_t'\beta^{(0)}, t = 1, \ldots, n$.

Step 2: In light of $r_t \approx (1 + \alpha_1|y_{t-1}| + \alpha_2|y_{t-2}|)\varepsilon_t$, q_k may be estimated by solving

$$\min_{q_k, a_{1k}, a_{2k}} \sum_{t=3}^{n} \rho_{\tau_k}(r_t - q_k - a_{1k}|y_{t-1}| - a_{2k}|y_{t-2}|).$$

In addition, for each k, we estimate α_j by $\hat{\alpha}_{jk} = \hat{a}_{jk}/\hat{q}_k, j = 1, 2$.

Step 3: Finalize $q^{(0)}$ by taking the symmetric error condition into account, and initialize $\alpha_j^{(0)}$ by $K^{-1}\sum_{k=1}^{K}\hat{\alpha}_{jk}$, $j = 1, 2$.

3.4.2 Ramifications of the MMLP Algorithm

Now that we have some proposal on the initialization stage of the Algorithm 1, a natural follow-up question is how we should determine the number of outer iterates $\mathbb{C}$ (aka, “cycles”). Similar question has been discussed in great detail in pp.14–18 of Carroll and Ruppert (1988), under the generalized least squares (GLS) framework. It is worth pointing out that the algorithm for GLS described in Carroll and Ruppert (1988, pp.69–70) and our MMLP algorithm Algorithm 1 for CQR bear in essential resemblance, at least from the iterative structure point of view. Specifically, both algorithms have to alternate between steps of estimating regression parameters and dispersion parameters. In their monograph, Carroll and Ruppert (1988) acknowledge that "there has been no clear consensus about the best of the number of cycles, which reflects the indeterminate nature of the theoretical calculations." Through a series of simulation study, they recommend "at least two cycles of generalized least squares, largely to eliminate the effect of the inefficient unweighted least-squares estimate."

Before giving any recommendation on the best choice of $\mathbb{C}$ in Algorithm 1, let's point out that in Algorithm 1, the weights $\{w_k\}$ and $\{h_i\}$ are treated as fixed during the whole implementation process. In other words, both $\{w_k\}$ and $\{h_i\}$ are inputs of the function that implements Algorithm 1 and are never updated in the execution period of the function. Such specification is indeed useful in some situations, in particular when we want to evaluate the performance of the algorithm under the true weights or when we have reliable weights a priori. In practice, however, it seems heuristic that the optimal weights should be calculated adaptively using the rules derived in Chapter 2, thus depend on the latest values of β, α[†]. Therefore, if $\mathbb{C} \geq 2$, the Algorithm 1 can be modified accordingly to the Algorithm 2 below, which takes the varying weights into account.

Algorithm 2 General MMLP algorithm for nonlinear heteroscedastic composite quantile regression, with weights updated during each outer iterate

```
Initialize $\theta^{(0)} = (\beta^{(0)\prime}, \alpha^{(0)\prime})'$, $\xi^{(0)} = (q_1^{(0)}, \dots, q_K^{(0)})'$. Set the outer iterate index $\ell = 1$, weights $w_k = 1, k = 1, \dots, K$, $h_i = 1, i = 1, \dots, n$.
repeat
    MM Step: Set the starting value for the MM step as $\theta^{(\ell-1)}$, assume $\xi^{(\ell-1)}$ to be known constants. Update $\theta^{(\ell-1)}$ to $\theta^{(\ell)}$ by completing the MM steps (3.8) and (3.9) multiple times, until the MM algorithm's convergence condition is met.
    Compute the optimal weights $\{w_k\}$ and $\{h_i\}$ using $\theta^{(\ell)}$. These weights will be used in the LP step in this outer iterate and the MM step in the next outer iterate.
    LP Step: Evaluate $\tilde{y}$, $\tilde{X}_1, \dots, \tilde{X}_K$ in (3.11) using $\theta^{(\ell)}$. Solve the linear program (3.11) to update $\xi^{(\ell-1)}$ to $\xi^{(\ell)}$.
    Increase $\ell$ to $\ell + 1$.
until $\ell = \mathbb{C}$
```

[†] Interestingly, these weights do not depend on $(q_1, \dots, q_K)'$, on the contrary, they influence the LP step that updates $(q_1, \dots, q_K)'$, see Algorithm 2 for details.

As in the GLS scenario, there seems no theoretical guidance to help us determine the best choice of $\mathbb{C}$. In addition to that, it is also of our interest that when $\mathbb{C} \geq 2$, which of Algorithm 1 and Algorithm 2 gives better estimates. Clearly, as in Carroll and Ruppert (1988), these two questions need to be answered by simulation study, as well as from the practical perspective. We will relegate this simulation study to Section 3.5.

Note that in Algorithm 2, due to the newly added statement 4, in principle we are facing a distinctive optimization problem (3.1) for each outer iterate. This feature prevents the declaration of convergence of the algorithm based on the change of successive objective function values. On the other hand, although Algorithm 2 need not converge theoretically, by simulation we found that the convergence is usual in the sense that the change of parameters estimates $(\beta', \alpha')'$ is tiny when $\mathbb{C}$ is large (in experience, when $\mathbb{C}$ is around 20). This observation suggests the following modification of Algorithm 2.

Algorithm 3 General MMLP algorithm for nonlinear heteroscedastic composite quantile regression, with prespecified convergence threshold ϵ_0

```
1: Initialize $\theta^{(0)} = (\beta^{(0)\prime}, \alpha^{(0)\prime})'$, $\xi^{(0)} = (q_1^{(0)}, \ldots, q_K^{(0)})'$. Set the outer iterate index $\ell = 1$, weights $w_k = 1, k = 1, \ldots, K$, $h_i = 1, i = 1, \ldots, n$ and the threshold $\epsilon_0$. Set $\delta = 1.7977 \times 10^{308}$.
2: **while** $\|\delta\| > \epsilon_0$ **do**
3:     **MM Step:** Set the starting value for the MM step as $\theta^{(\ell-1)}$, assume $\xi^{(\ell-1)}$ to be known constants. Update $\theta^{(\ell-1)}$ to $\theta^{(\ell)}$ by completing the MM steps (3.8) and (3.9) multiple times, until the MM algorithm's convergence condition is met.
4:     Compute $\delta = \|\theta^{(\ell)} - \theta^{(\ell-1)}\|$.
5:     Compute the optimal weights $\{w_k\}$ and $\{h_i\}$ using $\theta^{(\ell)}$. These weights will be used in the **LP step** in this outer iterate and the **MM step** in the next outer iterate.
6:     **LP Step:** Evaluate $\tilde{y}$, $\tilde{X}_1, \ldots, \tilde{X}_K$ in (3.11) using $\theta^{(\ell)}$. Solve the linear program (3.11) to update $\xi^{(\ell-1)}$ to $\xi^{(\ell)}$.
7:     Increase $\ell$ to $\ell + 1$.
8: **end while**
```

Let me conclude this section by briefly commenting Algorithm 1–Algorithm 3. Algorithm 1 is most useful when we have prior knowledge about the weights $\{w_k\}$ and $\{h_i\}$, for which case we use Algorithm 1 by substituting $\{w_k\}$ and $\{h_i\}$ into (3.1) and typically setting $\mathbb{C} = 2$. This is particularly useful in simulation studies where the true underlying error distributions are known to us, hence both $\{w_k\}$ and $\{h_i\}$ can be calculated theoretically. Algorithm 2 and Algorithm 3 are essentially the same, except for different termination rules. The simulation result in Section 3.5 reveals that the Algorithm 3 is in general not preferable to the Algorithm 2, even at the cost of much more computational time. Thus in practice we recommend using Algorithm 2 with $\mathbb{C} = 2$ or $\mathbb{C} = 3$.

It is worth pointing out that the initialization steps in both algorithms can be completed by using the method proposed in Section 3.4.1. In experience, we found that all the MMLP algorithms are fairly sensitive to the initial values of ξ (by comparison, the initial values of θ is less sensitive), and when applicable, the method proposed in Section 3.4.1 does provide a warm start of ξ to ensure the good behavior of the MMLP algorithm. It is also straightforward to note that the CQR procedure with uniform weights (i.e., non-weighting) and the adaptive procedure mentioned in Section 2.4.2 can be effectively carried out by running Algorithm 2 with $\mathbb{C} = 1$ and $\mathbb{C} = 2$, respectively. More cycles may be exerted if higher precision of estimates is expected.

3.5 Monte Carlo Studies

In this section, we conduct extensive Monte Carlo experiments to demonstrate the numerical performance of the proposed MMLP algorithm. The purpose of this section is threefold, namely:

Purpose 1: To corroborate the recommendation made at the end of Section 3.4.2.

Purpose 2: To illustrate how different choices of the weights $\{w_k\}$ and $\{h_i\}$ affect the estimation accuracy.

Purpose 3: To demonstrate the advantage of the adaptive CQR method using the optimal weights proposed in Chapter 2 over other conventional statistical methods, under a wide scope of error distributions.

Throughout this section, Monte Carlo samples will be simulated from the following two models:

$$\text{Model 1: } y_t = \beta_0 + \exp(\beta_1 x_t) + \exp(0.3 + \alpha x_t^2)\varepsilon_t, \quad t = 1, \ldots, n.$$

$$\text{Model 2: } y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + (1 + \alpha_1 |y_{t-1}| + \alpha_2 |y_{t-2}|)\varepsilon_t,$$
$$\alpha_1 > 0, \alpha_2 > 0, \quad t = 1, \dots, n.$$

In Model 1, the true parameters are set to be $(\beta_0, \beta_1)' = (1, 2)'$ and $\alpha = 0.3$. The regressors are independent uniform random variables: $\{x_t\}$ i.i.d. $\sim \mathcal{U}(0, 1)$. Similar models to Model 1 are termed as "model the logarithm of the variances as linear in predictors" in p.65 of Carroll and Ruppert (1988).

In Model 2, the true parameters are set to be $(\beta_0, \beta_1, \beta_2)' = (1, -0.4, 0.2)'$ and $(\alpha_1, \alpha_2)' = (0.3, 0.1)'$. The regressors are independent uniform random variables: $\{x_{jt}\}$ i.i.d. $\sim \mathcal{U}(0, 1), j = 1, 2$ and they are independent to each other. Model 2 differs from the celebrated ARCH-type model (Engle (1982)) by replacing the "square-root" volatility function $\sqrt{1 + \alpha_1 y_{t-1}^2 + \alpha_2 y_{t-2}^2}$ with the "absolute-value" volatility function $1 + \alpha_1 |y_{t-1}| + \alpha_2 |y_{t-2}|$, which is widely adopted in the quantile regression literature, for instance, see Koenker and Zhao (1996), Xiao and Koenker (2009). Model 2 is essentially a time series model, but the exogenous factors $\{x_{jt}\}$ are also included to show the flexibility of our method. At the first glance, both the regression function g and the dispersion function s are linear in parameters, but the constraints $\alpha_1 > 0$ and $\alpha_2 > 0$ effectively disqualifies s to be treated as linear function of α. In literature, this issue is usually handled by estimating α as if it is unconstrained, and only the positive estimates would result in a meaningful model. Here, we take a simple reparametrization approach: let $\alpha_1 = e^{\gamma_1}, \alpha_2 = e^{\gamma_2}$ so that γ_1 and γ_2 are unconstrained, and Model 2 can be rewritten as

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + (1 + e^{\gamma_1} |y_{t-1}| + e^{\gamma_2} |y_{t-2}|)\varepsilon_t, \quad t = 1, \dots, n,$$

which matches our nonlinear heteroscedastic model specification, therefore the previous algorithms are readily applicable to estimate $(\beta_0, \beta_1, \beta_2, \gamma_1, \gamma_2)'$. After the estimates $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are obtained, we estimate α_j by $\hat{\alpha}_j = \exp(\hat{\gamma}_j)$, $j = 1, 2$.

Before reporting the simulation results below, I must acknowledge that though the initialization methods proposed in Section 3.4.1 could be used, it is found through simulation that the quality of estimates are greatly influenced by the initial values of $(q_1, \dots, q_K)'$. In other words, the precision of the estimates of $(\alpha, \beta)'$ are very sensitive to the correct specification of $(q_1^{(0)}, \dots, q_K^{(0)})'$. Considering the general difficulty of solving (3.1) and our main purpose is to demonstrate the advantage of using composite quantile regression with optimal weights, we will initialize $(q_1, \dots, q_K)'$ with their true values in all the subsequent experiments. Of course, we view this as a potential drawback of our algorithm as well as a tantalizing future research topic.

To address Purpose 1, we focus our attention on Model 1 only. Three types of the error distributions $\mathcal{N}(0,1)$ (Normal), Student's t_3 and $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(1,1)$ (MG2) are considered. The sample size we took was $n = 600$, and there were 400 simulations in the experiment. For each case, we initialize β by $\beta^{(0)} = (0,3)'$, $\alpha^{(0)} = 1$. In line with the Table 2.2 in Carroll and Ruppert (1988), the results of our simulation are summarized in the Table 3.1 below.

Table 3.1. *Mean squared errors* $(\times 10^{-2})$ *of estimator* $(\hat{\beta}_0(\mathbb{C}), \hat{\beta}_1(\mathbb{C}), \hat{\alpha}(\mathbb{C}))'$ *under Model 1, for different number of cycles* $\mathbb{C}$*. The* ϵ_0 *in the* $\mathbb{C}$ *column corresponds to Algorithm 3. For columns initiated with* β_0 *and* β_1*, the weights* $\{w_k\}$ *are computed using* (2.13)*, while for columns initiated with* α*, the weights* $\{w_k\}$ *are computed using* (2.16)*.*

	Normal			t_3			MG2		
$\mathbb{C}$	β_0	β_1	α	β_0	β_1	α	β_0	β_1	α
1	0.740	0.121	0.577	1.104	0.182	0.883	1.723	0.234	0.550
2	0.704	0.122	0.693	1.112	0.184	1.152	1.489	0.209	0.633
3	0.710	0.123	0.835	1.117	0.185	1.425	1.504	0.213	0.750
5	0.711	0.124	1.018	1.118	0.185	1.717	1.503	0.215	0.895
10	0.715	0.125	1.144	1.117	0.186	1.901	1.508	0.215	0.988
20	0.714	0.125	1.190	1.114	0.185	1.990	1.513	0.216	1.029
ϵ_0	0.712	0.125	1.198	1.111	0.185	2.017	1.508	0.214	1.033

It can be seen from the Table 3.1 that in general the increasing of iteration times doesn't improve the estimation accuracy, but for some cases increasing $\mathbb{C}$ from 1 to 2 does enhance the efficiency a little. Considering the initialization issue mentioned above probably makes the scenario $\mathbb{C} = 1$ overly optimistic, in practice we recommend using Algorithm 2 with $\mathbb{C} = 2$, by noting that more cycles is immaterial and even harmful.

To investigate Purpose 2, for each of Model 1 and Model 2, the estimators of β and α are reported for six symmetric error distributions, using adaptive weights with varying objective functions, adaptive weights with unchanged objective function, theoretical weights ((2.13), (2.16)) and the uniform weights. These methods are labelled as “Adaptive 1”, “Adaptive 2”, “Theoretical” and “Uniform” respectively in Table 3.2 and Table 3.3. For each of specifications, the Monte Carlo replications is $N = 400$, and for each run, the sample size is $n = 600$. All the runs using uniform weights and theoretical weights are implemented using Algorithm 1 with $\mathbb{C} = 2$, and all the runs using adaptive weights are implemented using Algorithm 2 with $\mathbb{C} = 2$. The initialization method for Model 1 is the same as that for Table 3.1, whereas for Model 2, we set $(\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)})' = (0.6, -0.2, 0.1)'$, $(\alpha_1^{(0)}, \alpha_2^{(0)})' = (0.2, 0.2)'$.

The simulation results are summarized in Table 3.2 and Table 3.3 below. From these two tables, we can make the following observations:

1. Among the six errors, using the theoretical weights w_β^* always results in the best performance in estimating β. By comparison, using the theoretical weights w_α^* always results in the best performance in estimating α for Model 2. This empirical evidence corroborates our theoretical arguments in Chapter 2.

2. Except for a few cases in Table 3.2, both adaptive methods outperform the uniform method by some notable margin, which highlights the benefit (efficiency gain) of weighting whenever it is applicable. On the other hand, the “Adaptive 1” method and the “Adaptive 2” method are generally comparable.

3. For some unclear reason, the “Uniform” method gives the best result for Laplace, logistic, MG1 and t_3 errors when estimating α under Model 1, though the leading advantage may typically be considered as negligible. This unexpected output deserves further study in our future study.

Finally, to address Purpose 3, we will compare our adaptive method WCQR (Algorithm 2 with $\mathbb{C} = 2$ and the initialization method as before) with the following conventional estimation methods:

1. Generalized least squares (GLS): This is generalized least squares estimation produced by iteratively reweighted least squares algorithm introduced in Carroll and Ruppert (1988), p.69. To obtain the weights estimation in the intermediate steps, we regress the squared residuals on the dispersion function

Table 3.2. *Mean squared errors* $(\times 10^{-3})$ *of different weighting schemes under Model 1, for a fixed error distribution. In the heading, "Laplace", "logistic" and "Normal" all refer to their standard distributions, "MG1" refers to* $0.9\mathcal{N}(0,1)+0.1\mathcal{N}(0,10)$ *and "MG2" refers to* $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(1,1)$.

	Laplace			logistic			MG1		
	α	β_0	β_1	α	β_0	β_1	α	β_0	β_1
Adaptive 1	12.674	8.7206	1.1853	9.4725	21.508	3.0689	9.1527	10.553	1.3979
Adaptive 2	12.689	8.6483	1.1779	9.3	21.628	3.0739	9.2	10.601	1.3978
Theoretical	12.815	8.3654	1.1405	9.2485	21.003	3.0001	9.0719	10.368	1.3931
Uniform	12.279	9.4436	1.2958	9.0968	21.327	3.0449	9.0483	10.384	1.3957
	MG2			Normal			t_3		
	α	β_0	β_1	α	β_0	β_1	α	β_0	β_1
Adaptive 1	4.7738	14.861	2.1779	7.4534	7.5984	1.1366	10.87	11.503	1.7774
Adaptive 2	4.6549	14.86	2.1945	7.1712	7.6028	1.1387	10.9	11.439	1.7786
Theoretical	4.8254	14.529	2.1364	7.2374	7.4402	1.1053	10.931	11.288	1.7264
Uniform	5.2565	17.864	2.6398	7.7377	7.8152	1.1563	10.819	11.871	1.7953

$s(\cdot,\alpha)$. As noted before, the GLS method they employed is similar to our MMLP algorithm. In particular, I will also set $\mathbb{C}=2$ in the GLS algorithm.

2. Unweighted LAD (UWLAD): The unweighted Least Absolute Deviation estimation.
3. Weighted LAD (WLAD): Weighted LAD estimation, for which the weights are estimated by performing a median regression of absolute value of residuals on the dispersion function $s(\cdot,\alpha)$.
4. Weighted LAD using true weights (TWLAD): Weighted LAD estimation, for which the weights are computed using the true parameter α.

Since all the above methods are proposed to estimate the regression parameter β only, the results listed in Table 3.4 and Table 3.5 only include the mean standard errors of different $\hat{\beta}$s, though our method may well be applied to estimate the dispersion parameter, as illustrated in earlier Monte Carlo studies. Same remarks that explain the entries in Table 3.2 and Table 3.3 invariably work for Table 3.4 and Table 3.5, with necessary changes self-explanatory.

It can be seen from Table 3.4 and Table 3.5 that the proposed WCQR method dominates other methods for every combination (except for Model 1, $\mathcal{N}(0,1)$

Table 3.3. *Mean squared errors* $(\times 10^{-3})$ *of different weighting schemes under Model 2, for a fixed error distribution. In the heading, "Laplace", "logistic" and "Normal" all refer to their standard distributions, "MG1" refers to* $0.9\mathcal{N}(0,1)+0.1\mathcal{N}(0,10)$ *and "MG2" refers to* $0.5\mathcal{N}(-1,1)+0.5\mathcal{N}(1,1)$.

	Laplace					logistic				
	α_1	α_2	β_0	β_1	β_2	α_1	α_2	β_0	β_1	β_2
Adaptive 1	2.5982	1.4343	41.274	64.687	63.534	1.3644	0.89809	128.5	215.34	212.62
Adaptive 2	2.4867	1.394	41.831	66.601	66.538	1.265	0.87434	129.23	216.35	212.88
Theoretical	2.3665	1.3825	39.823	62.076	63.925	1.248	0.86326	126.37	209.32	207.37
Uniform	2.8144	1.4858	48.526	74.133	79.418	1.6605	1.1112	135.04	228.39	233.6
	MG1					MG2				
	α_1	α_2	β_0	β_1	β_2	α_1	α_2	β_0	β_1	β_2
Adaptive 1	2.169	1.0943	36.005	70.679	67.793	1.1045	0.69326	69.099	115.13	116.11
Adaptive 2	1.9842	1.0559	35.889	70.376	67.725	1.0436	0.66715	70.461	116.35	116.17
Theoretical	1.9334	1.0278	35.162	69.77	66.443	1.0223	0.67305	68.251	114.23	113.58
Uniform	2.5649	1.2632	39.016	73.004	74.936	1.2862	0.84492	87.732	150.35	137.62
	Normal					t_3				
	α_1	α_2	β_0	β_1	β_2	α_1	α_2	β_0	β_1	β_2
Adaptive 1	2.205	1.3454	26.067	46.673	45.009	2.7316	1.8005	47.804	82.137	83.012
Adaptive 2	2.0518	1.2977	25.732	46.44	45.315	2.6277	1.8321	48.825	82.013	83.426
Theoretical	2.0654	1.2626	25.296	46.119	43.352	2.293	1.2	47.293	81.273	82.021
Uniform	2.3803	1.4128	28.296	49.398	48.633	3.1199	1.6016	50.596	83.962	94.766

error, the GLS method works slightly better, which is easy to understand). In particular, for Model 2, except for the logistic error, our WCQR method on average outperforms the other methods by at least one magnitude. Although the MMLP algorithm has the aforementioned initialization restriction, we may still consider, as an innovative estimation method, the adaptive optimal double-weighted composite quantile regression method proposed in Chapter 2 can substantially escalate estimation efficiency.

3.6 Conclusion

In this chapter, we generalize the MM algorithm proposed by Hunter and Lange (2000), which works only for the single-quantile case, to the composite quantile regression situation. In addition to that, our algorithm also takes the heterogeneity part of the model into account by setting the nuisance parameter $(q_1, \ldots, q_K)'$ to be design variables, and update them by solving a linear programming problem. This

Table 3.4. *Mean squared errors* ($\times 10^{-3}$) *of different estimation methods under Model 1, for a fixed error distribution. In the heading, "Laplace", "logistic" and "Normal" refer to their standard distributions, "MG1" refers to* $0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(0,10)$ *and "MG2" refers to* $0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(1,1)$.

	Laplace		logistic		MG1		MG2		Normal		t_3	
	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
GLS	14.72	2.46	23.29	3.94	12.99	2.29	13.96	2.29	6.54	1.09	23.39	3.79
UWLAD	8.99	1.32	25.05	3.64	11.98	1.92	30.34	4.48	10.63	1.56	13.15	2.14
WLAD	9.04	1.33	25.05	3.67	12.04	1.93	29.59	4.36	10.56	1.55	13.12	2.14
TWLAD	8.91	1.28	25.51	3.70	12.60	1.99	30.10	4.36	10.66	1.58	13.27	2.13
WCQR	8.65	1.18	21.51	3.07	10.55	1.40	14.86	2.18	7.60	1.14	11.50	1.78

Table 3.5. *Mean squared errors* ($\times 10^{-2}$) *of different estimation methods under Model 2, for a fixed error distribution. In the heading, "Laplace", "logistic" and "Normal" refer to their standard distributions, "MG1" refers to* $0.9\mathcal{N}(0,1) + 0.1\mathcal{N}(0,10)$ *and "MG2" refers to* $0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(1,1)$.

	Laplace			logistic			MG1		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
GLS	25.851	19.730	15.308	44.493	52.324	44.930	23.301	16.389	14.991
UWLAD	19.914	10.977	6.460	32.403	32.531	29.254	21.615	14.368	9.923
WLAD	19.824	10.497	6.690	32.482	30.299	29.435	21.770	13.881	10.372
TWLAD	19.936	10.432	6.644	32.239	30.494	29.390	21.990	13.679	10.001
WCQR	4.127	6.469	6.353	12.850	21.534	21.262	3.589	7.038	6.773
	MG2			Normal			t_3		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
GLS	26.036	20.551	18.344	18.095	8.768	6.490	28.434	28.910	29.101
UWLAD	30.018	28.564	28.444	18.937	11.167	7.894	21.810	14.498	12.327
WLAD	28.884	27.382	30.342	18.719	11.145	8.343	21.758	15.199	11.675
TWLAD	29.557	27.853	29.462	18.521	10.967	8.375	21.438	14.822	11.555
WCQR	6.910	11.513	11.611	2.573	4.644	4.532	4.780	8.214	8.301

explains the name of our algorithm: the "MMLP" algorithm. The Monte Carlo studies performed in Section 3.5 confirm the theoretical assertions in Chapter 2, also show that the proposed DWCQR method overall outperforms the conventional statistical methods such as the generalized least squares, provided a reasonably good initial values of $(q_1, \ldots, q_K)'$ are supplied to the MMLP algorithm.

In developing this chapter, some technical issues, such as the number of cycles of the algorithm, the initialization method are also discussed. By simulation, we recommend using $\mathbb{C} = 2$. On the other hand, it seems difficult to come up with decent initial values of $(q_1, \ldots, q_K)'$, which should be deemed as a major weakness of our algorithm. We consider this unsolved problem as an interesting future research topic, either from the statistical or the computational perspective.

Chapter 4
Efficient Quantile Regression for Linear Heteroscedastic Models: an Alternative Approach

4.1 Introduction

Inhomogeneity of the variances of the errors is a common phenomenon in practice. As a major tool of modeling such inhomogeneity, heteroscedastic regression models have received many important applications (among others, the celebrated ARCH model (Engle (1982)) is a leading example) and have been extensively studied in literature. Anscombe (1961) conducted the pioneering work of pointing out the fitting of an ideal linear model is often only a first tentative step in the analysis of data and has proposed some simple test statistics for heteroscedasticity. This research is followed up by Bickel (1978), which investigated the power of Anscombe's procedures when the error distributions are not normal and compared these procedures with some natural alternative tests which are robust against gross errors. To the estimation end, Welsh et al. (1994) have proposed two estimation methods for a broad class of heteroscedastic regression models. For a comprehensive treatment to heteroscedastic models by the classical least squares methods, we refer readers to Carroll and Ruppert (1988).

After the introduction of the seminal concept *regression quantiles* by Koenker and Bassett Jr. (1978), it immediately becomes a powerful tool of studying heteroscedastic models. Very much in the spirit of Bickel (1978)'s work, Koenker and Bassett Jr. (1982) propose an alternative approach to robustify tests for heteroscedasticity based on regression quantiles. For the location-scale heteroscedastic models

$$y_t = x_t'\beta + (x_t'\alpha)u_t, \qquad t = 1, \ldots, n, \tag{4.1}$$

under the *local heteroscedasticity* assumption, Gutenbrunner and Jurečková (1992) derive the asymptotic distributions for $\{n^{1/2}(\hat{\beta}_n(\tau) - \beta(\tau)), 0 < \tau < 1\}$, namely, the *regression quantile processes*, where

$$\beta(\tau) = \beta + F^{-1}(\tau)\alpha, \tag{4.2}$$

$\hat{\beta}_n(\tau) = \arg\min\limits_{b \in \mathbb{R}^p} \sum\limits_{t=1}^{n} \rho_\tau(y_t - x_t'b)$. For the same location-scale model (4.1), if a $\sqrt{n}$-consistent estimator $\hat{\gamma}$ of γ is available, Koenker and Zhao (1994) investigate the uniform Bahadur representation of the *weighted regression quantiles* $n^{1/2}(\hat{\beta}(\tau, \hat{\alpha}) - \beta(\tau))$, where $\hat{\beta}(\tau, \hat{\alpha}) = \arg\min\limits_{b \in \mathbb{R}^p} \sum\limits_{t=1}^{n} (x_t'\hat{\gamma})^{-1} \rho_\tau(y_t - x_t'b)$, and then propose the way of constructing efficient L-estimators of β. The similar proof techniques has also been successfully applied to ARCH models (Koenker and Zhao (1996)) to establish the Bahadur representation of the ARCH parameter's estimator. Zhao (2001) considers the linear model with unknown heteroscedasticity form, and estimates β by using weighted least absolute deviation regression, in which the weights are estimated by k-nearest neighbors approach. For other works on heteroscedastic models using quantile regression methods, let's mention Welsh (1996), Zhou and Portnoy (1998) and Xiao and Koenker (2009).

In this chapter, we consider a class of heteroscedastic models which is more general than (4.1) as follows:

$$y_t = x_t'\beta + s(x_t, \alpha)u_t, \qquad t = 1, \ldots, n, \tag{4.3}$$

where $\beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^k$, $s(x_t, \alpha)$ is some smooth known nonlinear function of α such that $s(x_t, \alpha) > 0$ for all t, and $u_t, t = 1, \ldots, n$ are i.i.d. errors. Our goal is seeking for efficient estimators of the regression parameter β, provided the existence of $\sqrt{n}$-consistent estimator $\hat{\alpha}_n$ of the dispersion parameter α. It is worth pointing out that due to the nonlinearity of $s(x_t, \cdot)$, it's infeasible to form a transformed parameter $\beta(\tau)$ such as (4.2) which simultaneously contains regression parameter β and dispersion parameter α, hence we need to look for another appropriate normalizing constant when discussing the asymptotic distribution of any sensible estimator of β. Under model (4.3), for any $\tau \in (0,1)$, our main contribution is to propose a weighted estimation method which aims to estimate β directly (which is in marked contrast to estimate the transformed version of β), and then to establish the Bahadur representation of this estimator, which will still be denoted by $\hat{\beta}(\tau)$ with the abuse of notation. Interestingly, we find that the effect of the preliminary estimator $\hat{\alpha}$ propagates in the Bahadur representation of $n^{1/2}(\hat{\beta}(\tau) - \beta)$. This undesirable asymptotic bias can be eliminated by wisely incorporating the information across different quantile positions, as suggested by Zhao and Xiao (2014). The resulting *weighted quantile average estimator* (WQAE) may then be further polished to achieve optimal asymptotic efficient performance, leading to the *optimal weighted quantile average estimator* (OWQAE).

The asymptotic theory on an estimator which is constructed based on a preliminary estimator has been studied to some extent in literature. Among others, Bickel (1975) and Ruppert and Carroll (1980) have become classics. However, both of the models in these two papers are restricted to homoscedastic linear models. Koenker and Zhao (1994) generalized the case to linear location-scale models, but their result failed to reveal an explicit asymptotic relation between the estimator of interest and the preliminary estimator. In this paper, under the more general nonlinear heteroscedastic models, we succeeded in showing how the preliminary estimator effects the asymptotic behavior of the estimator of central interest (similar research to ours can be found in Koenker and Zhao (1996), while firstly, their model is essentially location-scale, and secondly, some proofs are lacking of important intermediate steps.).

The rest of this paper is organized as follows: in Section 4.2 we describe the weighted quantile regression procedure of estimating β, for any fixed quantile position $\tau \in (0,1)$, and then give the Bahadur representation of $n^{1/2}(\hat{\beta}(\tau) - \beta)$. In Section 4.3 we report an adaptive procedure on how to construct OWQAE of β based on the results obtained in Section 4.2, and then demonstrate its good efficiency property. We discuss computational aspects and compare our method with other estimation methods by simulation in Section 4.4. Technical proofs and auxiliary results are collected in Section 4.5.

Before presenting the main result of this project, we briefly commenting the connections between the method to be introduced in this chapter and the CQR method proposed in Chapter 2. In short, the method considered in this chapter is residual-based, i.e., depends on a preliminary estimator of the dispersion parameter. By contrast, the CQR method estimates all the parameters jointly, hence does not require any preliminary estimators. This seemingly advantage, however, is at the cost of solving a very challenging optimization problem, as fully explained in Chapter 4. In fact, it is the difficulty of coming up with an effective algorithm to solve the CQR optimization problem that motivates me to initiate this project from scratch. Consequently, residual-based method, which results in an estimator that is similar to the L-estimator is computationally convenient. In spite of the very disparate schemes in obtaining the estimates, these two methods share one vital feature — the information across several different quantile positions is synthesized by weighting (see Koenker (1984)), and the weights are determined based on the result of asymptotic analysis to achieve the optimal efficiency. As can be seen from Section 2.3 and Section 4.3, despite of the contrastive asymptotics, the efficiency analysis for the CQR and for the L-estimator is almostly the same.

4.2 Bahadur Representation of $\hat{\beta}(\tau)$

In this paper, we restrict the independent variables $\{x_t\}$ to be nonrandom, though the results obtained throughout will continue to hold if $\{x_t\}$ is independent of $\{u_t\}$, or $\{(x_t, u_t)\}$ is strictly stationary and ergodic, without increasing essential difficulty in the proofs.

For fixed $\tau \in (0, 1)$, as mentioned in Section 4.1, due to the infeasibility of incorporating the dispersion parameter α and the τ-quantile of error $F^{-1}(\tau)$ into the regression parameter β, we should view $F^{-1}(\tau)$ as a nuisance parameter and then estimate $F^{-1}(\tau)$ and β jointly by solving the optimization problem as follows

$$\min_{q\in\mathbb{R}^1, b\in\mathbb{R}^p} \sum_{t=1}^{n} \sigma_t^{-1}\rho_\tau(y_t - \sigma_t q - x_t' b), \tag{4.4}$$

where $\sigma_t \equiv s(x_t, \alpha)$, $\rho_\tau(z) = z(\tau - I[z < 0])$. Here and in the sequel, for simplicity, we write $I[z \in A]$ to stand for the indicator function $I_A(z)$, for any subset $A \subset \mathbb{R}^1$. The rationale of coming up with problem (4.4) can be explained as follows: by denoting y_t/σ_t by y_t^*, x_t/σ_t by x_t^*, model (4.3) can be rewritten as

$$y_t^* = x_t^{*\prime}\beta + u_t, \qquad t = 1, \dots, n.$$

Hence the τ-quantile of y_t^* is $Q_{y_t^*}(\tau) = x_t^{*\prime}\beta + F^{-1}(\tau) = (1, x_t^{*\prime})(F^{-1}(\tau), \beta')'$, suggesting that $(F^{-1}(\tau), \beta')'$ can be estimated by minimizing $\sum_{t=1}^{n} \rho_\tau(y_t^* - q - x_t^{*\prime}b) = \sum_{t=1}^{n} \sigma_t^{-1}\rho_\tau(y_t - \sigma_t q - x_t' b)$ with respect to $(q, b')'$. Reasoning in this way, we automatically assigns (4.4) with a *weighted quantile regression* form, which is expected to result in an estimator that has higher efficiency than its unweighted counterpart. In practice, σ_t is unavailable and will be estimated by $\hat{\sigma}_t \equiv s(x_t, \hat{\alpha})$, where $\hat{\alpha} \equiv \hat{\alpha}_n$ is any $\sqrt{n}$-consistent estimator of α. Substituting σ_t in (4.4) by $\hat{\sigma}_t$ yields the *estimated weighted regression quantile estimator*:

$$(\widehat{F^{-1}(\tau)}, \hat{\beta}(\tau)) = \underset{q \in \mathbb{R}^1, b \in \mathbb{R}^p}{\arg\min} \sum_{t=1}^{n} \hat{\sigma}_t^{-1}\rho_\tau(y_t - \hat{\sigma}_t q - x_t' b). \tag{4.5}$$

To derive the Bahadur representation of $\hat{\beta}(\tau)$ (as well as of $\widehat{F^{-1}(\tau)}$), we will employ the following conditions:

F. F has positive density f such that $f \in \mathscr{C}'(\mathbb{R}^1)$. Moreover, both f and f' are bounded functions on $\mathbb{R}^1$.

SX1. $\sigma_t \equiv s(x_t, \alpha) \geq c_0 > 0, t = 1, \dots, n$, for some constant c_0.

SX2. For every $\tau \in (0, 1)$,

$$n^{-1} \sum_{t=1}^{n} \sigma_t^{-r} \begin{bmatrix} \xi_t \xi_t' & \xi_t z_t' \\ z_t \xi_t' & z_t z_t' \end{bmatrix} \to Q_r \equiv \begin{bmatrix} Q_r^{(11)} & Q_r^{(12)} \\ Q_r^{(21)} & Q_r^{(22)} \end{bmatrix}$$

as $n \to \infty$, where $\xi_t \equiv \dot{s}(x_t, \alpha), z_t \equiv (\sigma_t, x_t')'$, $Q_r \in \mathbb{R}^{(k+p+1)\times(k+p+1)}$: $r = 0, 1, 2$ are positive definite.

SX3. $\sum_{t=1}^{n} \|(\xi_t', z_t')'/\sigma_t\|^3 = O(n)$, $\sum_{t=1}^{n} \|(\xi_t', z_t')'/\sigma_t\| = O(n^{1/2})$.

SX4. $\max_{t \le n} \|(\xi_t', z_t')'/\sigma_t\| = O(n^{1/4})$.

SX5. There exist some $\varepsilon_0 > 0$ and $K > 0$ such that all the second partial derivatives of s are continuous on $\{\xi : \|\xi - \alpha\| < \varepsilon_0\}$, and for every $t \in \{1, \ldots, n\}$

$$\sup_{\|\xi\|<\varepsilon_0} \|\dot{s}(x_t, \alpha + \xi)\| \le K, \qquad \sup_{\|\xi\|<\varepsilon_0} \|\ddot{s}(x_t, \alpha + \xi)\| \le K^*.$$

Moreover,

$$\sup_{\|\xi\|\le\varepsilon_0} \frac{s(x_t, \alpha + \xi)}{s(x_t, \alpha)} \le K$$

uniformly in t.

We briefly comment the conditions above. Condition F is a classical condition imposed on error distributions under the quantile regression setting. Conditions SX1–SX4 naturally adapt conditions imposed on the location-scale model for the nonlinearity of the dispersion function $s(x_t, \cdot)$: these conditions reduce to conditions C1–C4 in Koenker and Zhao (1994) if $s(x_t, \alpha) = x_t'\alpha$. Condition SX5 imposes some uniform (in t) smoothness restrictions on the function $s(x_t, \cdot)$, which will be used to control the bounds of remainders in asymptotic expansions.

Under these conditions, we have the following main theorem:

Theorem 4.1. *Consider the heteroscedastic model*

$$y_t = x_t'\beta + \sigma_t u_t, \qquad t = 1, \ldots, n,$$

where $\sigma_t \equiv s(x_t, \alpha)$. *Given a preliminary* $\sqrt{n}$*-consistent estimator* $\hat{\alpha}_n$ *of* α, *define* $\hat{\delta}_n = n^{1/2}(\hat{\alpha}_n - \alpha)$. *Let* $(\hat{q}(\tau), \hat{\beta}(\tau)')'$ *be the solution to the optimization problem*

[*] Here we use 2-norm for a matrix A (Golub and Van Loan (1996), p.55):

$$\|A\| = \sup_{x \ne 0} \frac{\|Ax\|}{\|x\|}.$$

Throughout this article, by convention, for a vector $x \in \mathbb{R}^q$, $\|x\|$ means its Euclidean norm $\sum_{j=1}^{p} |x_j|^2$.

(4.5). *Then under conditions F, SX1–SX5, $(\hat{q}(\tau), \hat{\beta}(\tau)')'$ has the following Bahadur representation:*

$$n^{1/2}\begin{bmatrix}\hat{q}(\tau) - F^{-1}(\tau)\\ \hat{\beta}(\tau) - \beta\end{bmatrix} = \frac{Q_2^{(22)^{-1}}}{f(F^{-1}(\tau))} n^{-1/2}\sum_{t=1}^{n}\begin{bmatrix}\sigma_t\\ x_t\end{bmatrix}\psi_\tau(u_t - F^{-1}(\tau)) - F^{-1}(\tau)Q_2^{(22)^{-1}}Q_2^{(21)}\hat{\delta}_n + o_P(1), \tag{4.6}$$

where $\psi_\tau(z) = \tau - I[z < 0], z \in \mathbb{R}^1$.

By performing some matrix computations, we can extract the Bahadur representation of β from (4.6) as follows:

Corollary. *Partition $Q_2^{(22)}$ and $Q_2^{(21)}$ according to $(F^{-1}(\tau), \beta')'$ as*

$$Q_2^{(22)} = \begin{bmatrix}a & v'\\ v & A\end{bmatrix}, \qquad Q_2^{(21)} = \begin{bmatrix}b'\\ B\end{bmatrix},$$

where $a \in \mathbb{R}^1$, $v \in \mathbb{R}^p$, $A \in \mathbb{R}^{p\times p}$, $b \in \mathbb{R}^k$, $B \in \mathbb{R}^{p\times k}$. Then under the same conditions of Theorem 4.1, $\hat{\beta}(\tau)$ has the following Bahadur representation:

$$n^{1/2}(\hat{\beta}(\tau)-\beta) = \frac{1}{n^{1/2}f(F^{-1}(\tau))}\sum_{t=1}^{n} C_t\psi_\tau(u_t-F^{-1}(\tau))-F^{-1}(\tau)D\hat{\delta}_n+o_P(1), \tag{4.7}$$

where $C_t = (A - a^{-1}vv')^{-1}(x_t - a^{-1}v\sigma_t), t = 1, \ldots, n$, $D = (A - a^{-1}vv')^{-1}(B - a^{-1}vb')$.

In view of (4.7), due to the existence of $\hat{\delta}_n$, $n^{1/2}(\hat{\beta}(\tau) - \beta)$ does not have zero mean asymptotically, which is evidently unsatisfactory. Therefore, it is undesirable to estimate β by only using one single weighted regression quantile $\hat{\beta}(\tau)$. To annihilate such bias caused by $\hat{\delta}_n$, it is natural to consider the weighted quantile average estimator (WQAE) proposed by Zhao and Xiao (2014):

$$\hat{\beta}_{\mathrm{WQAE}}(w) = \sum_{j=1}^{K} w_j\hat{\beta}(\tau_j). \tag{4.8}$$

(4.8) is also known as the Mosteller's estimator (Koenker (1984)) or (a discrete version of) the L-estimatior (Gutenbrunner and Jurečková (1992)). It is easy to see that under the constraints

$$\sum_{j=1}^{K} w_j = 1, \qquad \sum_{j=1}^{K} w_j F^{-1}(\tau_j) = 0, \tag{4.9}$$

$\hat{\beta}_{\text{WQAE}}(w)$ has the Bahadur representation

$$n^{1/2}(\hat{\beta}_{\text{WQAE}}(w) - \beta) = \frac{1}{n^{1/2}} \sum_{j=1}^{K} \frac{w_j}{f(F^{-1}(\tau_j))} \sum_{t=1}^{n} C_t \psi_{\tau_j}(u_t - F^{-1}(\tau_j)) + o_P(1), \tag{4.10}$$

which implies the asymptotic normality of $\hat{\beta}_{\text{WQAE}}(w)$. We summarize this observation to a theorem as follows:

Theorem 4.2. *Given any vector of weights $w = (w_1, \ldots, w_K)'$ that satisfies* (4.9)*, define the weighted quantile average estimator $\hat{\beta}_{WQAE}(w)$ as in* (4.8)*. Then under conditions F, SX1–SX5, $n^{1/2}(\hat{\beta}_{WQAE}(w) - \beta)$ converges weakly to a normal distribution with mean* 0 *and covariance matrix $\Sigma_\alpha^{-1} S(w)$, where*

$$\Sigma_\alpha^{-1} = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} C_t C_t' = (A - a^{-1} v v')^{-1},$$

$$S(w) = w' H w \ \textit{with}\ H = \left[\frac{\tau_j \wedge \tau_{j'} - \tau_j \tau_{j'}}{f(F^{-1}(\tau_j)) f(F^{-1}(\tau_{j'}))} \right] \in \mathbb{R}^{K\times K}.$$

In practice, without loss of generality, we can always fix K to be an odd number, and $\tau_j = j/(K+1), j = 1, \ldots, K$ are placed evenly on $(0,1)$. It's remarkable that if f is symmetric about 0, which is a common condition imposed on error distributions, then the second constraint in (4.9) simplifies to $w_j = w_{K+1-j}, j = 1, \ldots, (K-1)/2$, which is independent of the quantiles $F^{-1}(\tau_j), j = 1, \ldots, K$. For easier reference, in the sequel we will refer the constraints $\sum\limits_{j=1}^{K} w_j = 1$ and $w_j = w_{K+1-j}, j = 1, \ldots, (K-1)/2$ as the *unity constraint* and the *symmetry constraint*, respectively.

4.3 Adaptive Weighted Quantile Average Estimator

By Theorem 4.2, the asymptotic variance of $n^{1/2}(\hat{\beta}_{\text{WQAE}}(w) - \beta)$ consists of two parts: the first part Σ_{α}^{-1} depends only on the design X and the true dispersion function $s(x_t, \cdot)$, while the second part $S(w)$ depends on the weights w and sparsity function of error u. As proposed in Zhao and Xiao (2014), this structure offers us a straightforward way to select the "optimal" weights w so that the asymptotic variance is "minimized". The following theorem shows that under the symmetry distribution assumption, the optimal weights have a closed form solution:

Theorem 4.3. *Under the same conditions of Theorem 4.2, and further assume that the density function of error is symmetric about* 0*, then*

$$w^* = \frac{H^{-1}\mathbf{1}}{\mathbf{1}'H^{-1}\mathbf{1}}, \qquad \textit{where } \mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^K$$

minimizes $S(w)$ and satisfies the unity constraint and the symmetry constraint simultaneously. We call $\hat{\beta}_{WQAE}(w^*)$ the Optimal Weighted Quantile Average Estimator (OWQAE) *of β, and denote it by $\hat{\beta}_{OWQAE}$. It has the following limiting distribution:*

$$n^{1/2}[\hat{\beta}_{OWQAE} - \beta] \Rightarrow \mathcal{N}(0, \Sigma_{\alpha}^{-1}\Omega_K^{-1}), \ \textit{where } \Omega_K = \mathbf{1}'H^{-1}\mathbf{1}.$$

In practice, $\hat{\beta}_{\text{OWQAE}}$ is not directly applicable for it contains the unknown function $\ell(\tau) \equiv f(F^{-1}(\tau))$. To further study the asymptotic property of its plug-in version, in line with Zhao and Xiao (2014), we make the following assumption on its estimator $\hat{\ell}(\tau)$.

E. $\sup\limits_{1 \le j \le K} |\hat{\ell}(\tau_j) - \ell(\tau_j)| = o_P(1)$.

We then denote corresponding plug-in estimators of H and w^* by $\hat{H}$ and $\hat{w}^*$ respectively. Furthermore, denote $\hat{\beta}_{\text{WQAE}}(\hat{w}^*)$ by $\hat{\beta}_{\text{EOWQAE}}$. It then follows by Slutsky's theorem that under condition E, $\hat{\beta}_{\text{EOWQAE}}$ and $\hat{\beta}_{\text{OWQAE}}$ have the same limiting distribution. In other words, the estimator $\hat{\beta}_{\text{EOWQAE}}$ is adaptive.

Here we propose an adaptive procedure to estimate $\ell(\tau)$ so that condition E is satisfied, provided a $\sqrt{n}$-consistent estimator $\hat{\alpha}$ of α is available (the existence of $\hat{\alpha}$

is assumed throughout the chapter). A general procedure is as follows:

Step 1: Obtain a $\sqrt{n}$-consistent estimator $\hat{\beta}^0$ of β.

Step 2: Compute the estimated noises:

$$\hat{u}_t = \frac{y_t - x_t'\hat{\beta}^0}{s(x_t, \hat{\alpha})}, \quad t = 1, \ldots, n. \tag{4.11}$$

Step 3: Estimate $f(u)$ through the nonparametric kernel density estimator:

$$\tilde{f}(z) = \frac{1}{nb_n} \sum_{t=1}^{n} K\left(\frac{z - \hat{u}_t}{b_n}\right) \tag{4.12}$$

for a bandwidth $b_n > 0$ and kernel function $K(\cdot)$.

Step 4: Estimate $F^{-1}(\tau)$ by the sample τ-th quantile of $\hat{u}_1, \ldots, \hat{u}_n$, denoted by $\tilde{F}^{-1}(\tau)$.

Step 5: To ensure the symmetric density condition in Theorem 4.3, we symmetrize $\tilde{f}$ and $\tilde{F}^{-1}(\tau)$ through

$$\hat{f}(z) = \frac{1}{2}(\tilde{f}(z) + \tilde{f}(-z)) \quad \text{and} \quad \hat{F}^{-1}(\tau) = \frac{1}{2}(\tilde{F}^{-1}(\tau) - \tilde{F}^{-1}(1 - \tau))$$

Step 6: Plug $\hat{f}(\hat{F}^{-1}(\tau))$ into H in Theorem 4.2 to obtain $\hat{H}$, then plug $\hat{H}$ into w^* in Theorem 4.3 to obtain the estimated optimal weights $\hat{w}^*$.

Under the assumption that the error u is symmetric, the $\hat{\beta}^0$ in Step 1 can be obtained in various ways, among which the least squares estimator $\hat{\beta}^0_{\text{LS}}$ and the least absolute deviation estimator $\hat{\beta}^0_{\text{LAD}}$ are two obvious candidates. Alternatively, practitioners also use the *composite quantile regression* (Zou and Yuan (2008)) estimator $\hat{\beta}_{\text{CQR}}$ as $\hat{\beta}^0$, regardless the heteroscedasticity of error terms. For simplicity and robustness consideration, we recommend using $\hat{\beta}_{\text{LAD}}$ and $\hat{\beta}^0$.

In Step 3, we follow Silverman (1986) to use the rule-of-thumb bandwidth b_n:

$$b_n = 0.9n^{-1/5} \min\left\{\text{sd}(\hat{u}_1, \ldots, \hat{u}_n), \frac{\text{IQR}(\hat{u}_1, \ldots, \hat{u}_n)}{1.34}\right\}$$

where, "sd" and "IQR" are the sample standard deviation and the sample interquartile range.

To show that the resulting estimators $\hat{\ell}(\tau_j), j = 1, \ldots, K$ fulfill condition E, we need to impose the following condition on bandwidth b_n and kernel function $K(\cdot)$:

N. (i) The bandwidth $b_n \propto n^{-1/5}$, (ii) The kernel $K(\cdot)$ is Lipschitz continuous and $\displaystyle\int_{\mathbb{R}^1} K(x)\,\mathrm{d}x = 1$.

Theorem 4.4. *Under conditions F, SX1, SX3–SX5, N, then* $\hat{\ell}(\tau_j) \equiv \hat{f}(\hat{F}^{-1}(\tau_j))$, $j = 1, \ldots, K$ *obtained from Step 6 above satisfy condition E. Consequently, the above procedure is adaptive.*

The good efficiency property of $\hat{\beta}_{\mathrm{EOWQAE}}$ is justified by the following theorem:

Theorem 4.5. *Suppose the function* $\ell(\tau) : (0, 1) \to \mathbb{R}^1$ *satisfies the* efficiency regularity condition*:*

$$\frac{1}{\tau}(\ell^2(\tau) + \ell^2(1-\tau)) + \tau^2 \int_{\tau}^{1-\tau} |\ell''(t)|^2\,\mathrm{d}t \to 0$$

as $\tau \to 0$*, then*

$$\lim_{K\to\infty} \Omega_K = \mathcal{F}(f),$$

where

$$\mathcal{F}(f) = \int_{\mathbb{R}^1} \frac{[f'(u)]^2}{f(u)}\,\mathrm{d}u$$

is the Fisher information of the error distribution f*.*

4.4 Monte Carlo Studies

In this section, we conduct Monte Carlo studies to investigate the sampling performance of the proposed procedures in two different models. In all settings below, we use 600 realizations to evaluate the performance of various methods.

Each model under consideration is of the form

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + s(x_{1t}, x_{2t}, \alpha) u_t, \qquad t = 1, \ldots, n.$$

For our first model, we use a modified version of the scale function from Carroll and Ruppert (1982) as the following expression:

$$\text{Model 1: } s(x_1, x_2, \alpha) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2). \tag{4.13}$$

So in this setting, the dispersion parameter α coincides with the regression parameter β, in which case the existence of $\sqrt{n}$-consistent estimator $\hat{\alpha}$ is automatically guaranteed if a $\sqrt{n}$-consistent preliminary estimator of β is available. Although we require the design $\{x_t\}$ is nonrandom during the theoretical development, here for convenience we assume the regressors x_1 and x_2 are generated by (see Zhao (2001)):

$$x_1 = U + 0.2V, \quad x_2 = 0.2U + V, \quad U \sim \mathcal{N}(5, 9), \quad V \sim U(0, 4).$$

The true parameter is set to be $\beta = (\beta_0, \beta_1, \beta_2)' = (1, -0.4, 0.2)'$. The distributions of errors will be specified shortly together with Model 2.

We specify the dispersion function of our second model as follows:

$$\text{Model 2: } s(x_1, x_2, \alpha) = \sqrt{\alpha_0 + \alpha_1 x_1^2 + \alpha_2 x_2^2}. \tag{4.14}$$

Thus the form of Model 2 resembles the celebrated ARCH model. The true parameters are set to be $\beta = (\beta_0, \beta_1, \beta_2)' = (2, 1, 2)', (\alpha_0, \alpha_1, \alpha_2) = (1, 2, 1)'$. The regressors are drawn from $\mathcal{N}_2((0, 0)', I_2)$.

For both Model 1 and Model 2, the distributions of error u are taken to be *normalized* normal, Student's t_3, Cauchy, mixture normal $(0.5\mathcal{N}(-2, 1) + 0.5\mathcal{N}(2, 1))$ and Laplace so that $F^{-1}(0.5) = 0$ and $G^{-1}(0.5) = 1$, where G is the distribution function of $|u|$. Note that all the distributions under consideration are symmetric about zero.

For Model 1, since $\alpha \equiv \beta$ and $\hat{\beta}^0$ is $\sqrt{n}$-consistent for β, $\hat{\beta}^0$ can be naturally taken as a $\sqrt{n}$-consistent estimator of α. For Model 2 (or more generally, models that α and β are different), the presumed $\sqrt{n}$-consistent estimator α is obtained by performing a least absolute deviation (LAD) regression of absolute values of residuals $y_t - x_t'\hat{\beta}^0$, $t = 1, \ldots, n$ on the nonlinear function $s(x_t, \cdot)$. We wish to give theoretical justifications for $\hat{\alpha}$ obtained in this way is indeed $\sqrt{n}$-consistent in our future work.

We will compare the following six estimation methods:

1. Generalized least squares (GLS): This is generalized least squares estimation produced by iteratively reweighted least squares algorithm introduced in Carroll and Ruppert (1988, p.69). For Model 2, to obtain the weights estimation in the intermediate steps, we regress the square of residuals on x_1^2 and x_2^2.

2. Unweighted LAD (UWLAD): The unweighted LAD estimation.

3. Weighted LAD (WLAD): Weighted LAD estimation, for which the weights are obtained as described in the last paragraph.

4. Weighted LAD using true weights (TWLAD): Weighted LAD estimation, for which the weights are computed using the true α values.

5. WQAE using theoretical optimal weights (TWQAE): It is $\hat{\beta}(w^*)$ in Theorem 4.3.

6. WQAE using estimated optimal weights (EWQAE): It is the adaptive estimator $\hat{\beta}(\hat{w}^*)$, where $\hat{w}^*$ are estimated from data following the procedure proposed in Section 4.3.

For each of the method mentioned above, the Monte Carlo replications is $N = 1000$, and for each run, the number of data points generated from Model 1 and Model 2 is $n = 600$. With N replications, we use EWQAE as the benchmark to which the other five methods are compared based on the empirical relative efficiency:

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(\hat{\beta}_\ell(j) - \beta_\ell)^2 \text{ and } \text{RE(Method)} = \frac{MSE(\text{Method})}{MSE(\text{EWQAE})},$$

where “Method” stands for one of UWLAD, WLAD, TWLAD and TWQAE, and $\hat{\beta}_\ell(j)$ is the estimator of β_ℓ in the j-th run, $\ell = 1, 2$, $j = 1, \ldots, N$. A value of RE that is greater than 1 indicates better performance of EWQAE.

The simulation results are summarized in Table 4.1 and Table 4.2. From Table 4.1 and Table 4.2, we can make the following observations:

Table 4.1. *Empirical relative efficiency of EWQAE compared to other five methods under Model 1. The last row gives the empirical MSE of EWQAE, the remaining entries are the empirical relative efficiencies of the method in that row when estimating β, for a fixed error distribution.*

	Normal		Student's t_3		Cauchy		Mixture Normal		Laplace	
Methods	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
GLS	399.7	8.440	38081	29711	3.23×10^8	9.68×10^8	44.30	4.212	9859	787.2
UWLAD	4.694	2.564	3.279	1.867	0.398	0.444	27.14	10.78	2.000	1.708
WLAD	1.332	1.144	1.075	1.034	0.545	0.636	4.262	2.595	0.974	0.988
TWLAD	1.328	1.129	1.079	1.027	0.544	0.652	4.149	2.537	0.967	0.988
TWQAE	1.005	0.982	0.993	0.946	0.812	11.27	1.067	0.935	0.974	0.988
EWQAE.MSE ($\times 10^{-5}$)	6.17	24.3	7.69	29.2	51.2	98.8	2.94	12.8	9.24	28.3

Table 4.2. *Empirical relative efficiency of EWQAE compared to other five methods under Model 2. The last row gives the empirical MSE of EWQAE, the remaining entries are the empirical relative efficiencies of the method in that row when estimating β, for a fixed error distribution.*

	Normal		Student's t_3		Cauchy		Mixture Normal		Laplace	
Methods	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
GLS	1.359	1.781	3.118	4.934	5265	2402	1.621	2.086	6.402	5.845
UWLAD	1.533	1.434	1.201	1.049	0.265	0.113	17.82	21.40	0.940	0.963
WLAD	1.503	1.535	1.276	1.235	0.430	0.211	17.46	21.86	0.948	0.957
TWLAD	1.439	1.501	1.231	1.107	0.259	0.109	17.39	21.34	0.917	0.933
TWQAE	0.993	0.990	1.019	0.985	0.926	0.817	1.251	1.282	0.947	0.957
EWQAE.MSE	0.0221	0.0177	0.0262	0.0227	0.094	0.169	0.0071	0.0045	0.0214	0.0163

1. In general, the quantile-based methods outperform the mean-based method significantly, even for the Gaussian-error case. This suggests that the existence of heterogeneity greatly impairs the applicability of GLS method, especially when the heterogeneity is strong. Consequently, the quantile-based methods are better choices for the parameter estimation problem under heteroscedastic models.

2. Among all the quantile-based methods under consideration, by comparing results between UWLAD and WLAD/TWLAD/TWQAE, we see that the efficiency gain through weighting the heteroscedasticity is also substantial. Hence weighted scheme is typically more preferable than its unweighted counterpart.

3. Except for the Cauchy-error case and Laplace-error case, our method overall outperforms methods that only use the 0.5 quantile position information, in particular when the error is mixture normal, suggesting that our method may have great potential when error distribution is not uni-modal. In addition, it is remarkable that the EWQAE with estimated optimal weights has comparable (many times slightly better) performance than the TWQAE with theoretical optimal weights.

4. The inferiority of our method compared to other median-regression methods under Laplace-error is explicable since LAD estimation is the MLE if errors are Laplacian, for which no wonder the TWLAD method gives the optimal estimation. Nevertheless, we see that our method is still comparable to the TWLAD for this case.

5. For the Cauchy-error case, the single-quantile-based methods outperform our proposed method by somewhat noticeable margin. This may be explained as for either TWQAE or EWQAE, we need to estimate the dispersion parameter α based on residuals. The accuracy of this estimation can be seriously deteriorated because of the high kurtosis of Cauchy error. It can be expected our method would again be competitive if we were able to get more reliable estimate of α.

6. By comparing EWQAE.EMSE rows of Table 4.1 and Table 4.2, it can be seen that the estimation accuracy of β under Model 1 is much higher than that under Model 2, which again demonstrates the importance of obtaining a reliable estimate of α: in Model 1, the $\sqrt{n}$-consistency of $\hat{\alpha} \equiv \hat{\beta}^0$ is guaranteed by theory while the $\sqrt{n}$-consistency of the residual-based $\hat{\alpha}$ in Model 2 hasn't been justified in theory.

In conclusion, substantial efficiency gain can be achieved by our proposed method under various error distributions, and how to generate a theoretically-insured $\sqrt{n}$-consistent estimator of α before applying our proposed method will be an interesting research topic in the future.

4.5 Technical Proofs

The proof of Theorem 4.1 is quite complicated, so we will first prepare a series of lemmas.

Lemma 4.1. *Define*

$$V_1(\delta, \Delta) \equiv \frac{1}{n^{1/2}} \sum_{t=1}^{n} \frac{1}{\sigma_t} \begin{bmatrix} \xi_t \\ z_t \end{bmatrix} \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}z_t'\Delta).$$

Then, under assumptions F, SX1–SX3, we have

$$\sup_{\|(\delta',\Delta')'\| \le M} \left\| V_1(\delta, \Delta) - V_1(0,0) + f(F^{-1}(\tau))Q_2 \begin{bmatrix} \delta F^{-1}(\tau) \\ \Delta \end{bmatrix} \right\| = o_P(1).$$

for fixed M, $0 < M < \infty$.

Proof. For simplicity, denote $(\delta' F^{-1}(\tau), \Delta')'$ by $d \in \mathbb{R}^{k+p+1}$, $(\xi_t', z_t')'$ by $Z_t \in \mathbb{R}^{k+p+1}$, then

$$V_1(\delta, \Delta) \equiv V_1(d) = n^{-1/2} \sum_{t=1}^{n} \sigma_t^{-1} Z_t \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1} Z_t' d).$$

Denote $\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}Z_t'd)$ by $\eta_t(d)$. It then can be seen that:

$$\begin{aligned} & |\eta_t(d) - \eta_t(0)| \\ =& |I(u_t - F^{-1}(\tau) < 0) - I(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}Z_t'd < 0)| \\ \le& I(-n^{-1/2}\|Z_t/\sigma_t\|M \le u_t - F^{-1}(\tau) \le n^{-1/2}\|Z_t/\sigma_t\|M). \end{aligned}$$

Therefore, by Lagrange mean value theorem and assumption F:

$$\begin{aligned} & E[|\eta_t(\Delta) - \eta_t(0)|^2] \\ \le& F(F^{-1}(\tau) + n^{-1/2}\|Z_t/\sigma_t\|M) - F(F^{-1}(\tau) - n^{-1/2}\|Z_t/\sigma_t\|M) \\ \le& c_1 n^{-1/2}\|Z_t/\sigma_t\| \end{aligned}$$

for some constant c_1 that is independent of t and n.

Denote the jth coordinate of $V_1(\delta, \Delta)$ by $V_1^{(j)}(\delta, \Delta)$, $j = 1, \ldots, k, k+1, \ldots, k+p+1$. Given $\varepsilon > 0$, for $j \in \{1, \ldots, k\}$, it follows by Chebyshev's inequality that:

$$\begin{aligned}
&P[|V_1^{(j)}(d) - V_1^{(j)}(0) - E[V_1^{(j)}(d) - V_1^{(j)}(0)]| \geq \varepsilon] \\
\leq&\frac{1}{\varepsilon^2}\operatorname{Var}(V_1^{(j)}(d) - V_1^{(j)}(0)) = \frac{1}{\varepsilon^2}\operatorname{Var}\left(n^{-1/2}\sum_{t=1}^n \sigma_t^{-1}\xi_{tj}[\eta_t(d) - \eta_t(0)]\right) \\
=&\frac{1}{n\varepsilon^2}\sum_{t=1}^n \operatorname{Var}(\sigma_t^{-1}\xi_{tj}[\eta_t(d) - \eta_t(0)]) \leq \frac{1}{n\varepsilon^2}\sum_{t=1}^n E[\sigma_t^{-2}\xi_{tj}^2[\eta_t(d) - \eta_t(0)]^2] \\
\leq&\frac{c_1}{n^{3/2}\varepsilon^2}\sum_{t=1}^n \|Z_t/\sigma_t\|^3 = O(n^{-1/2}),
\end{aligned}$$

where the last equality follows from condition SX3. In the same manner, it can be shown that for $j \in \{k+1, \ldots, k+p+1\}$, we have

$$V_1^{(j)}(d) - V_1^{(j)}(0) - E[V_1^{(j)}(d) - V_1^{(j)}(0)] = o_P(1).$$

Hence $V_1(d) - V(0) - E[V_1(d) - V_1(0)] = o_P(1)$.

Next, using Bickel's chaining approach (Bickel (1975)), it can be shown that

$$\sup_{\|d\| \leq M} \|V_1(d) - V_1(0) - E[V_1(d) - V_1(0)]\| = o_P(1). \tag{4.15}$$

Finally, we need to show

$$\sup_{\|d\| \leq M} \|E[V_1(d) - V_1(d)] + f(F^{-1}(\tau))Q_2 d\| = o_P(1),$$

which, by condition SX2, is implied by

$$\begin{aligned}
\sup_{\|d\| \leq M} \Bigg\| n^{-1/2}\sum_{t=1}^n \sigma_t^{-1} Z_t \left[F(F^{-1}(\tau)) - F(F^{-1}(\tau) + n^{-1/2}\sigma_t^{-1} Z_t' \Delta)\right] \\
+ n^{-1}\sum_{t=1}^n \sigma_t^{-2} Z_t Z_t' f(F^{-1}(\tau)) d \Bigg\| = o_P(1).
\end{aligned} \tag{4.16}$$

By Taylor's theorem with the integral remainder, we have

$$
\begin{aligned}
&F(F^{-1}(\tau)) - F(F^{-1}(\tau) + n^{-1/2}\sigma_t^{-1}Z_t'd) = -f(F^{-1}(\tau))n^{-1/2}\sigma_t^{-1}Z_t'd \\
&- n^{-1}(\sigma_t^{-1}Z_t'd)^2 \int_0^1 (1-s)f'(F^{-1}(\tau) + sn^{-1/2}\sigma_t^{-1}Z_t'd)\,\mathrm{d}s.
\end{aligned}
$$

Hence the left hand side of (4.16) is bounded by

$$
\begin{aligned}
&\sup_{\|d\|\le M} \left\| n^{-3/2}\sum_{t=1}^n \sigma_t^{-1}Z_t(\sigma_t^{-1}Z_t'\Delta)^2 \int_0^1 (1-s)f'(F^{-1}(\tau) + sn^{-1/2}Z_t'\Delta)\,\mathrm{d}s \right\| \\
\le& c_2 n^{-3/2}\sum_{t=1}^n \|Z_t/\sigma_t\|^3 = O(n^{-1/2}). \qquad (4.17)
\end{aligned}
$$

Combining (4.15) and (4.17) completes the proof. □

When the estimated weights $\hat{\sigma}_t$ is used, the following lemma asserts that the resulting criterion function is close to the theoretical criterion function.

Lemma 4.2. *Define*

$$
\hat{V}_1(\delta, \Delta) \equiv n^{-1/2}\sum_{t=1}^n \hat{\sigma}_t^{-1} \begin{bmatrix} \xi_t \\ z_t \end{bmatrix} \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}z_t'\Delta).
$$

Then under conditions F, SX1–SX5,

$$
\sup_{\|(\delta', \Delta')'\|\le M} \|\hat{V}_1(\delta, \Delta) - V_1(\delta, \Delta)\| = o_P(1).
$$

for fixed M, $0 < M < \infty$.

Proof. Adopting the notations used in Lemma 4.1, write $\hat{V}_1(\delta, \Delta)$ as

$$
\hat{V}_1(d) = n^{-1/2}\sum_{t=1}^n \hat{\sigma}_t^{-1}Z_t\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}Z_t'd).
$$

By Taylor's theorem (Ferguson (1996), Section 4):

$$
\hat{\sigma}_t^{-1} - \sigma_t^{-1} = -\frac{\dot{s}(x_t, \alpha)'}{s^2(x_t, \alpha)}(\hat{\alpha}_n - \alpha) + r_{t,n},
$$

where

$$r_{t,n} = (\hat{\alpha}_n - \alpha)' \left[\int_0^1 \int_0^1 vM(x_t, \alpha + uv(\hat{\alpha}_n - \alpha)) \, \mathrm{d}u \, \mathrm{d}v \right] (\hat{\alpha}_n - \alpha),$$

with, for $\xi \in \mathbb{R}^k$,

$$M(x_t, \xi) = \frac{2\dot{s}(x_t, \xi)\dot{s}(x_t, \xi)'}{s^3(x_t, \xi)} - \frac{\ddot{s}(x_t, \xi)}{s^2(x_t, \xi)} \in \mathbb{R}^{k \times k}. \tag{4.18}$$

Thus $\hat{V}_1(d) - V_1(d)$ can be decomposed as $T_1 + T_2$, where

$$\begin{aligned} T_1 &= -n^{-1/2} \sum_{t=1}^{n} \frac{\xi_t' \hat{\delta}_n}{\sigma_t^2} Z_t \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1} Z_t' \Delta), \\ T_2 &= n^{-1/2} \sum_{t=1}^{n} r_{t,n} Z_t \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1} Z_t' \Delta), \end{aligned}$$

where $\hat{\delta}_n = \hat{\alpha}_n - \alpha = O_P(n^{-1/2})$. By conditions SX3, SX4, it follows that

$$|T_1| \leq \|\hat{\delta}_n\| n^{-1/2} \sum_{t=1}^{n} \|\xi_t/\sigma_t\| \|Z_t/\sigma_t\| = O_P(n^{-1/4}).$$

Since $\hat{\delta}_n = o_P(1)$, with probability tending to 1, $\|\hat{\delta}_n\| < \varepsilon_0$ for all sufficiently large n. It then follows by(4.18) and condition SX5 that

$$\begin{aligned} |T_2| \leq & \, n^{-1/2} \sum_{t=1}^{n} \left\| \frac{Z_t}{\sigma_t} \right\| \int_0^1 \int_0^1 \sigma_t |\hat{\delta}_n' M(x_t, \alpha + uv\hat{\delta}_n)\hat{\delta}_n| \, \mathrm{d}u \, \mathrm{d}v \\ \leq & \frac{1}{n^{1/2}} \sum_{t=1}^{n} \left\| \frac{Z_t}{\sigma_t} \right\| \int_0^1 \int_0^1 \frac{2s(x_t, \alpha) \|\dot{s}(x_t, \alpha + uv\hat{\delta}_n)\dot{s}(x_t, \alpha + uv\hat{\delta}_n)'\|}{s^3(x_t, \alpha + uv\hat{\delta}_n)} \|\hat{\delta}_n\|^2 \, \mathrm{d}u \, \mathrm{d}v \\ & + \frac{1}{n^{1/2}} \sum_{t=1}^{n} \left\| \frac{Z_t}{\sigma_t} \right\| \int_0^1 \int_0^1 \frac{s(x_t, \alpha)\|\ddot{s}(x_t, \alpha + uv\hat{\delta}_n)\|}{s^2(x_t, \alpha + uv\hat{\delta}_n)} \|\hat{\delta}_n\|^2 \, \mathrm{d}u \, \mathrm{d}v \\ = & O_P(n^{-1}). \end{aligned}$$

This completes the proof. □

Considering the last $p + 1$ components of $V_1(\delta, \Delta)$ and $\hat{V}_1(\delta, \Delta)$, from Lemma 4.1 and Lemma 4.2, we have the following corollary:

Corollary. *Put*

$$\begin{aligned} V_1^{(2)}(\delta,\Delta) &= n^{-1/2}\sum_{t=1}^{n}\sigma_t^{-1}z_t\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}z_t'\Delta),\\ \hat{V}_1^{(2)}(\delta,\Delta) &= n^{-1/2}\sum_{t=1}^{n}\hat{\sigma}_t^{-1}z_t\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}z_t'\Delta). \end{aligned} \tag{4.19}$$

Then, under the conditions of Lemma 4.2,

$$\begin{aligned} \sup_{\|(\delta',\Delta')'\|\le M} \Big\| V_1^{(2)}(\delta,\Delta) - V_1^{(2)}(0,0) + &\\ f(F^{-1}(\tau))(Q_2^{(21)}\delta F^{-1}(\tau) + Q_2^{(22)}\Delta)\Big\| &= o_P(1),\\ \sup_{\|(\delta',\Delta')'\|\le M} \Big\| \hat{V}_1^{(2)}(\delta,\Delta) - V_1^{(2)}(\delta,\Delta)\Big\| &= o_P(1) \end{aligned}$$

for fixed M, $0 < M < \infty$.

Lemma 4.3. *Let $(\hat{q},\hat{\beta})$ be the minimizer of the function*

$$\sum_{t=1}^{n}\hat{\sigma}_t^{-1}\rho_\tau(y_t - \hat{\sigma}_t q - x_t'b)$$

where $\hat{\sigma}_t \ge c_0 > 0$. Then under assumptions F, SX4, with probability 1,

$$\begin{aligned} &\left| n^{-1/2}\sum_{t=1}^{n}\psi_\tau(u_t - \sigma_t^{-1}\hat{\sigma}_t F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\hat{\sigma}_t\hat{\Delta}_1 - n^{-1/2}\sigma_t^{-1}x_t'\hat{\Delta}_2)\right|\\ \le & n^{-1/2}p = o_P(1);\\ &\left| n^{-1/2}\sum_{t=1}^{n}\hat{\sigma}_t^{-1}x_{tj}\psi_\tau(u_t - \sigma_t^{-1}\hat{\sigma}_t F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\hat{\sigma}_t\hat{\Delta}_1 - n^{-1/2}\sigma_t^{-1}x_t'\hat{\Delta}_2)\right|\\ \le & pn^{-1/2}\max_{t\le n}\|x_t/\hat{\sigma}_t\| = o_P(1), \qquad j = 1,\ldots,p, \end{aligned} \tag{4.20}$$

where $\hat{\Delta}_1 = n^{1/2}(\hat{q} - F^{-1}(\tau))$, $\hat{\Delta}_2 = n^{1/2}(\hat{\beta} - \beta)$.

Remark 1. *For the inequality part, many existing papers (for instance, Koenker and Zhao (1994), Koenker and Zhao (1996)) cited Lemma A.2 of Ruppert and*

Carroll (1980). But a careful examination of the proof of this lemma shows that it is at least unclear, if not incorrect. Here I will first restate the result as a proposition then give a rigorous proof to it using Knight's identity (Knight (1998)).

Proposition. *Suppose $\hat{B} \in \mathbb{R}^P$ solves the following optimization problem*

$$\min_{B\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta(Y_i - X_i'B).$$

Then,

$$\left|\sum_{i=1}^{n} X_{ij}\psi_\theta(Y_i - X_i'\hat{B})\right| \le \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}]|X_{ij}|, \qquad j = 1, \dots, P. \tag{4.21}$$

Proof of Proposition. By Knight's identity, for every $t > 0$ and $w \in \mathbb{R}^P$ such that $\|w\| = 1$, we have

$$\begin{aligned} G_i(t) \equiv& \rho_\theta(Y_i - X_i'(\hat{B} + tw)) - \rho_\theta(Y_i - X_i'\hat{B}) \\ =& - tX_i'w\psi_\theta(Y_i - X_i'\hat{B}) + \int_0^{tX_i'w} \Big\{ I[Y_i - X_i'\hat{B} < s] - I[Y_i - X_i'\hat{B} < 0] \Big\}\, \mathrm{d}s. \end{aligned}$$

If $Y_i = X_i'\hat{B}$, then $G_i(t) = -tX_i'w\theta + \int_0^{tX_i'w} I[s > 0]\,\mathrm{d}s$. For this case, if $X_i'w = 0$, then $G_i(t) = 0$; if $X_i'w > 0$, then $G_i(t) = -tX_i'w\theta + tX_i'w = tX_i'w(1-\theta)$; if $X_i'w < 0$, then $G_i(t) = -tX_i'w\theta$. These three cases can be written compactly as $G_i(t) = tX_i'w(I[X_i'w > 0] - \theta)$. Thus

$$\lim_{t\to 0^+} \frac{G_i(t)}{t} = X_i'w(I[X_i'w > 0] - \theta). \tag{4.22}$$

If $Y_i \ne X_i'\hat{\beta}$, a careful evaluation shows that the integral term, when divided by t always tends to 0 as $t \to 0^+$, hence

$$\lim_{t\to 0^+} \frac{G_i(t)}{t} = -X_i'w\psi_\theta(Y_i - X_i'\hat{B}). \tag{4.23}$$

By (4.22) and (4.23), it follows that

$$
\begin{aligned}
0 \leq & \lim_{t\to 0^+} \frac{1}{t} \sum_{i=1}^{n} \left[\rho_\theta(Y_i - X_i'(\hat{B} + tw)) - \rho_\theta(Y_i - X_i'\hat{B}) \right] \\
= & \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] X_i'w(I[X_i'w > 0] - \theta) - \sum_{i=1}^{n} I[Y_i \neq X_i'\hat{B}] X_i'w\psi_\theta(Y_i - X_i'\hat{B}).
\end{aligned}
$$

Or equivalently,

$$
\sum_{i=1}^{n} I[Y_i \neq X_i'\hat{B}] X_i'w\psi_\theta(Y_i - X_i'\hat{B}) \leq \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] X_i'w(I[X_i'w > 0] - \theta).
$$

Adding $\sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] X_i'w\psi_\theta(Y_i - X_i'\hat{B})$ to both sides of the above inequality then yields

$$
\begin{aligned}
\sum_{i=1}^{n} X_i'w\psi_\theta(Y_i - X_i'\hat{B}) \leq & \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] X_i'w(I[X_i'w > 0] - I[Y_i < X_i'\hat{B}]) \\
= & \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] |X_i'w| I[X_i'w > 0]
\end{aligned}
$$

Similar analysis for $\lim_{t\to 0^-} \sum_{i=1}^{n} G_i(t)/t \leq 0$ yields

$$
\sum_{i=1}^{n} X_i'w\psi_\theta(Y_i - X_i'\hat{B}) \geq -\sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] |X_i'w| I[X_i'w \leq 0].
$$

These two inequalities together imply

$$
\begin{aligned}
& \left| \sum_{i=1}^{n} X_i'w\psi_\theta(Y_i - X_i'\hat{B}) \right| \\
\leq & \max\left\{ \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] |X_i'w| I[X_i'w > 0], \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] |X_i'w| I[X_i'w \leq 0] \right\} \\
\leq & \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}] |X_i'w|.
\end{aligned}
$$

Taking w to be $e_1, \ldots, e_P$ respectively in above inequality then gives

$$\left|\sum_{i=1}^{n} X_{ij}\psi_\theta(Y_i - X_i'\hat{B})\right| \leq \sum_{i=1}^{n} I[Y_i = X_i'\hat{B}]|X_{ij}|, \qquad j = 1, \ldots, P.$$

This completes the proof of proposition. □

Proof of Lemma 4.3. Substitute $Y_i = y_t/\hat{\sigma}_t$, $X_i = (1, x_t'/\hat{\sigma}_t)'$, $B = (q, b')'$, $\hat{B} = (F^{-1}(\tau) + n^{-1/2}\hat{\Delta}_1, \beta' + n^{-1/2}\hat{\Delta}_2')'$ in (4.21), and then divide cases yields the inequality parts of (4.20), in which we also used condition F (the distribution of error is continuous).

To show $n^{-1/2}\max_{t\leq n}\|x_t/\hat{\sigma}_t\| = o_P(1)$, first apply (4.24) below, then use assumption SX4. □

The next lemma builds a bound of the difference between the weighted criterion function of (4.5) and its first-order approximation (4.19).

Lemma 4.4. *For $\delta \in \mathbb{R}^k, \Delta = (\Delta_1, \Delta_2')' \in \mathbb{R}^{p+1}$, define*

$$\begin{aligned}\hat{V}_2(\delta, \Delta) =& n^{-1/2}\sum_{t=1}^{n}\hat{\sigma}_t^{-1}\begin{bmatrix}\hat{\sigma}_t \\ x_t\end{bmatrix}\psi_\tau(u_t - \sigma_t^{-1}s(x_t, \alpha + n^{-1/2}\delta)F^{-1}(\tau) - \\ & n^{-1/2}\sigma_t^{-1}s(x_t, \alpha + n^{-1/2}\delta)\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2),\end{aligned}$$

Then under conditions F, SX1–SX5, it holds that

$$\sup_{\substack{\|\delta\|\leq M \\ \|\Delta\|\leq M}}\left\|\hat{V}_2(\delta, \Delta) - \hat{V}_1^{(2)}(\delta, \Delta)\right\| = o_P(1)$$

for fixed M, $0 < M < \infty$.

Proof. In Lemma 4.2 we have shown that

$$\hat{\sigma}_t^{-1} - \sigma_t^{-1} = -\frac{\xi_t'(\hat{\alpha}_n - \alpha)}{\sigma_t^2} + \sigma_t^{-2}\hat{r}_{t,n}, \tag{4.24}$$

where the remainder term $\hat{r}_{t,n}$ satisfies (with $\hat{d}_n$ denoting $\hat{\alpha}_n - \alpha$)

$$|\hat{r}_{t,n}| = \left|\hat{d}_n'\left[\int_0^1\int_0^1 \sigma_t^2 vM(x_t, \alpha + uv\hat{d}_n)\,\mathrm{d}u\,\mathrm{d}v\right]\hat{d}_n\right|$$

$$
\begin{aligned}
\leq &\int_0^1\int_0^1 \frac{\sigma_t^2}{s^3(x_t, \alpha + uv\hat{d}_n)} \|\dot{s}(x_t, \alpha + uv\hat{d}_n)\dot{s}'(x_t, \alpha + uv\hat{d}_n)\| \|\hat{d}_n\|^2 \,\mathrm{d}u\,\mathrm{d}v \\
&+ \int_0^1\int_0^1 \frac{\sigma_t^2}{s^2(x_t, \alpha + uv\hat{d}_n)} \|\ddot{s}(x_t, \alpha + uv\hat{d}_n)\| \|\hat{d}_n\|^2 \,\mathrm{d}u\,\mathrm{d}v \\
\leq &c_1 \|\hat{d}_n\|^2
\end{aligned}
$$

for some constant c_1 independent of t and n. Hence (4.24) implies

$$\hat{\sigma}_t^{-1} \leq \sigma_t^{-1} + \|\xi_t/\sigma_t^2\| \|\hat{\alpha}_n - \alpha\| + c_1\sigma_t^{-2}\|\hat{\alpha}_n - \alpha\|^2. \tag{4.25}$$

To quantify the difference between $\hat{V}_2$ and $\hat{V}_1^{(2)}$, we also need expand $s(x_t, \alpha + n^{-1/2}\delta)$ explicitly as follows:

$$s(x_t, \alpha + n^{-1/2}\delta) = \sigma_t + n^{-1/2}\xi_t'\delta + \rho_{t,n} \tag{4.26}$$

where $\rho_{t,n} = n^{-1}\delta'\left[\int_0^1\int_0^1 v\ddot{s}(x_t, \alpha + n^{-1/2}uv\delta)\,\mathrm{d}u\,\mathrm{d}v\right]\delta$ satisfies

$$|\rho_{t,n}| \leq c_2 n^{-1}\|\delta\|^2 \tag{4.27}$$

for some constant c_2 independent of t and n.

Denote the argument of $\psi_\tau(\cdot)$ in the definition of $\hat{V}_2(\delta, \Delta)$ by $\eta_t(\delta, \Delta)$. By (4.26), it follows that

$$\eta_t(\delta, \Delta) = u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2 - R_{t,n}$$

where $R_{t,n} = n^{-1}\sigma_t^{-1}\xi_t'\delta\Delta_1 + \sigma_t^{-1}\rho_{t,n}F^{-1}(\tau) + n^{-1/2}\sigma_t^{-1}\rho_{t,n}\Delta_1$. By (4.27) and SX4, we have

$$
\begin{aligned}
|R_{t,n}| &\leq n^{-1}c_3\|\xi_t/\sigma_t\|\|\delta\|^3\|\Delta_1\| + n^{-1}c_3\sigma_t^{-1}\|\delta\|^2 + n^{-3/2}c_3\sigma_t^{-1}\|\delta\|^2\|\Delta_1\| \\
&\leq c_4 n^{-3/4}
\end{aligned} \tag{4.28}
$$

for some constant c_4 independent of t and n.

After above preparation, now we shall prove the statement coordinate-wisely. We first compare the jth coordinate of $\hat{V}_2(\delta, \Delta)$ and that of $\hat{V}_1^{(2)}(\delta, \Delta)$ for $j = 2, \ldots, p+1$.

By definition, it is easily seen that

$$
\begin{aligned}
&|\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2 - R_{t,n}) \\
&\quad - \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2)| \\
=&|I(u_t - F^{-1}(\tau) < n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) + n^{-1/2}\Delta_1 + n^{-1/2}\sigma_t^{-1}x_t'\Delta_2) - \\
&\quad I(u_t - F^{-1}(\tau) < n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) + n^{-1/2}\Delta_1 + n^{-1/2}\sigma_t^{-1}x_t'\Delta_2 + R_{t,n})| \\
\le& I(F^{-1}(\tau) - B_n < u_t < F^{-1}(\tau) + B_n), \qquad (4.29)
\end{aligned}
$$

where, since $\|\delta\| \le M$ and $\|\Delta\| \le M$,

$$
\begin{aligned}
B_n &= n^{-1/2}\max_{t\le n}\left\|\frac{\xi_t}{\sigma_t}\right\| MF^{-1}(\tau) + n^{-1/2}M + n^{-1/2}\max_{t\le n}\left\|\frac{x_t}{\sigma_t}\right\| M + \max_{t\le n}|R_{t,n}| \\
&\le c_5 n^{-1/4}
\end{aligned}
$$

for some constant independent of n.

By (4.29) and (4.24), the difference between the $(j+1)$st coordinate of $\hat{V}_2(\delta,\Delta)$ and that of $\hat{V}_1^{(2)}(\delta,\Delta), j = 1,\ldots,p$ is bounded by

$$
\begin{aligned}
&n^{-1/2}\sum_{t=1}^n \hat{\sigma}_t^{-1}|x_{tj}|I(F^{-1}(\tau) - B_n < u_t < F^{-1}(\tau) + B_n) \\
\le& n^{-1/2}\sum_{t=1}^n \sigma_t^{-1}|x_{tj}|I(F^{-1}(\tau) - B_n < u_t < F^{-1}(\tau) + B_n) \\
&\quad + \|\hat{\alpha}_n - \alpha\| \cdot \sum_{t=1}^n n^{-1/2}\left\|\frac{\xi_t}{\sigma_t^2}\right\| |x_{tj}|I(F^{-1}(\tau) - B_n < u_t < F^{-1}(\tau) + B_n) \\
&\quad + c_1\|\hat{\alpha}_n - \alpha\|^2 \cdot n^{-1/2}\sum_{t=1}^n \sigma_t^{-2}|x_{tj}|I(F^{-1}(\tau) - B_n < u_t < F^{-1}(\tau) + B_n).
\end{aligned}
$$

By Markov's inequality and conditions F, SX1–SX4, it can be shown that each term of the right hand side of the above inequality is $o_P(1)$.

Next we compare the first component of $\hat{V}_2(\delta,\Delta)$ with that of $\hat{V}_1^{(2)}(\delta,\Delta)$, which may be decomposed as $T_1(\delta,\Delta) + T_2(\delta,\Delta)$ with

$$T_1(\delta,\Delta)$$

$$
\begin{aligned}
&=\frac{1}{n^{1/2}}\sum_{t=1}^{n}[\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2\\
&\quad - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1 - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2)]\\
&\quad T_2(\delta,\Delta)\\
&=\frac{1}{n^{1/2}}\sum_{t=1}^{n}\frac{\hat\sigma_t-\sigma_t}{\hat\sigma_t}\psi_\tau(u_t - F^{-1}(\tau) - n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) - n^{-1/2}\Delta_1\\
&\quad - n^{-1/2}\sigma_t^{-1}x_t'\Delta_2)
\end{aligned}
$$

By (4.24), (4.26) and (4.27) (with $\hat\delta_n$ denoting $n^{1/2}(\hat\alpha_n - \alpha)$),

$$
\begin{aligned}
&\left|\frac{\hat\sigma_t-\sigma_t}{\hat\sigma_t}\right|\\
\le&\left(\sigma_t^{-1} + n^{-1/2}\left\|\xi_t/\sigma_t^2\right\|\|\hat\Delta_n\| + c_1n^{-1}\sigma_t^{-2}\|\hat\delta_n\|^2\right)\left(n^{-1/2}\|\xi_t\|\|\hat\delta_n\| + c_2n^{-1}\|\hat\delta_n\|^2\right)\\
\le& n^{-1/2}\|\xi_t/\sigma_t\|\|\hat\delta_n\| + c_0^{-1}c_2n^{-1}\|\hat\delta_n\|^2 + n^{-1}\|\xi_t/\sigma_t\|^2\|\hat\delta_n\|^2 + c_2n^{-3/2}\|\xi_t/\sigma_t^2\|\|\hat\delta_n\|^3\\
&+ c_0^{-1}c_1n^{-3/2}\|\xi_t/\sigma_t\|\|\hat\delta_n\|^3 + c_0^{-2}c_1c_2n^{-2}\|\hat\delta_n\|^4.
\end{aligned}
$$

By this inequality, conditions SX3, SX4 and $|\psi_\tau(\cdot)|\le 1$, it follows that

$$
\begin{aligned}
&|T_2(\delta,\Delta)|\\
\le& n^{-1}\|\hat\delta_n\|\sum_{t=1}^{n}\|\xi_t/\sigma_t\| + n^{-1/2}c_0^{-1}c_2\|\hat\delta_n\|^2 + n^{-5/4}\|\hat\delta_n\|^2\sum_{t=1}^{n}\|\xi_t/\sigma_t\|\\
&+ n^{-2}c_0^{-1}c_2\|\hat\delta_n\|^3\sum_{t=1}^{n}\|\xi_t/\sigma_t\| + n^{-2}c_0^{-1}c_1\|\hat\delta_n\|^3\sum_{t=1}^{n}\|\xi_t/\sigma_t\| + n^{-3/2}c_0^{-2}c_1c_2\|\hat\delta_n\|^4\\
=&O_P(n^{-1/2}) = o_P(1).
\end{aligned}
$$

To bound $T_1(\delta,\Delta)$, the inequality (4.29) is not tight enough, which needs to be tightened a little more. For simplicity, denote $n^{-1/2}\sigma_t^{-1}\xi_t'\delta F^{-1}(\tau) + n^{-1/2}\Delta_1 + n^{-1/2}\sigma_t^{-1}x_t'\Delta_2$ by $\zeta_t(\delta,\Delta)$. To bound $E[|\psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta) - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta))|]$, we consider the following three cases:

Case 1: $R_{t,n} = 0$. Trivial.

Case 2: $R_{t,n} > 0$. For this case, we have

$$\begin{aligned}
&E[|\psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta) - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta))|] \\
=&P\left[F^{-1}(\tau) + \zeta_t(\delta,\Delta) \le u_t < F^{-1}(\tau) + \zeta_t(\delta,\Delta) + R_{t,n}\right] \\
=&F(F^{-1}(\tau) + \zeta_t(\delta,\Delta) + R_{t,n}) - F(F^{-1}(\tau) + \zeta_t(\delta,\Delta)) \\
\le&c_6 R_{t,n} \le c_6 c_4 n^{-3/4}.
\end{aligned}$$

The last inequality follows from (4.28).

Case 3: $R_{t,n} < 0$. Similarly to case 2, we have

$$\begin{aligned}
&E[|\psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta) - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta))|] \\
=&P\left[F^{-1}(\tau) + \zeta_t(\delta,\Delta + R_{t,n}) \le u_t < F^{-1}(\tau) + \zeta_t(\delta,\Delta)\right] \\
=&F(F^{-1}(\tau) + \zeta_t(\delta,\Delta)) - F(F^{-1}(\tau) + \zeta_t(\delta,\Delta + R_{t,n})) \\
\le&-c_6 R_{t,n} \le c_6 c_4 n^{-3/4}.
\end{aligned}$$

So for each case we have the bound

$$E[|\psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta) - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta))|] \le c_4 c_6 n^{-3/4}.$$

Therefore

$$\begin{aligned}
&E[|T_1(\delta,\Delta)|] \\
\le&n^{-1/2}\sum_{t=1}^{n} E[|\psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta) - R_{t,n}) - \psi_\tau(u_t - F^{-1}(\tau) - \zeta_t(\delta,\Delta))|] \\
\le&c_4 c_6 n^{-1/4} = O(n^{-1/4}).
\end{aligned}$$

By Markov's inequality, $T_1(\delta,\Delta) = o_P(1)$.

In summary, we have shown that for fixed δ and Δ such that $\|\delta\| \le M$ and $\|\Delta\| \le M$, it holds that $\|\hat{V}_2(\delta,\Delta) - \hat{V}_1^{(2)}(\delta,\Delta)\| = o_P(1)$. The proof will then be completed by routinely invoking the chaining arguments. □

Combining the results of Corollary 4.5 and Lemma 4.4, we obtain

Corollary. *Under conditions F, SX1–SX5,*

$$\sup_{\substack{\|\delta\|\le M\\ \|\Delta\|\le M}} \left\| \hat{V}_2(\delta,\Delta) - V_1^{(2)}(0,0) + f(F^{-1}(\tau))(Q_2^{(21)}\delta F^{-1}(\tau) + Q_2^{(22)}\Delta) \right\| = o_P(1)$$

for fixed M, $0 < M < \infty$.

In view of Lemma 4.3 and Corollary 4.5, in order to establish the Bahadur representation of $\hat{\Delta}$, it remains to show $\hat{\Delta} = O_P(1)$, which can be done nicely by following the proof of Lemma 5.2 in Jurečková (1977). Once this last step is completed, we can prove our main theorem as below.

Proof of Theorem 4.1. For $\Delta = (\Delta_1, \Delta_2')' \in \mathbb{R}^{p+1}$, define $\hat{V}(\Delta) \coloneqq \hat{V}_2(\hat{\delta}_n, \Delta)$. It then can be easily verified that

$$-\Delta'\hat{V}(\lambda\Delta) \ge -\Delta'\hat{V}(\Delta), \quad \lambda \ge 1.$$

On the other hand, $\hat{\delta}_n = O_P(1)$ and Corollary 4.5 together imply that

$$\sup_{\|\Delta\|\le M} \left\| \hat{V}(\Delta) - V_1^{(2)}(0,0) + f(F^{-1}(\tau))(Q_2^{(21)}\hat{\delta}_n F^{-1}(\tau) + Q_2^{(22)}\Delta) \right\| = o_P(1).$$

Finally, for $\hat{\Delta} = n^{1/2}(\hat{q}(\tau) - F^{-1}(\tau), \hat{\beta}(\tau)' - \beta')'$, Lemma 4.3 asserts that

$$\|\hat{V}(\hat{\Delta})\| = o_P(1).$$

The result of the theorem then follows by applying Lemma 3.4 in Koenker and Zhao (1996). □

Proof of Theorem 4.2. Let's denote $\sum_{j=1}^{K} f^{-1}(F^{-1}(\tau_j)) w_j \psi_{\tau_j}(u_t - F^{-1}(\tau_j))$ by ζ_t, $t = 1, \ldots, n$. By (4.10) and the Cramér-Wold device, it suffices to show that for any $c \in \mathbb{R}^p$,

$$n^{-1/2} \sum_{t=1}^{n} c' C_t \zeta_t \Rightarrow \mathcal{N}\left(0, c'\Sigma_\alpha^{-1} S(w) c\right).$$

Note that ζ_t can be further written as $\zeta_t = w'\Lambda\phi_t$, $t = 1, \ldots, n$, with

$$\Lambda = diag(f^{-1}(F^{-1}(\tau_1)), \ldots, f^{-1}(F^{-1}(\tau_K))),$$

$$\phi_t = (\psi_{\tau_1}(u_t - F^{-1}(\tau_1)), \ldots, \psi_{\tau_K}(u_t - F^{-1}(\tau_K)))'.$$

Using this form, it is then straightforward to show that

$$E[c'C_t\zeta_t] = 0, \ \mathrm{Var}(c'C_t\zeta_t) = (c'C_t)^2 S(w), \quad t = 1, \ldots, n.$$

The result then follows from Lindeberg's central limit theorem. □

Proof of Theorem 4.3. It is easily seen that w^* proposed in Theorem 4.3 minimizes $S(w)$ under the unity constraint. Therefore it remains to show this w^* is also symmetric.

Let Λ be defined as in the proof of Theorem 4.2, also define $\Gamma = [\tau_j \wedge \tau_{j'} - \tau_j\tau_{j'}] \in \mathbb{R}^{K\times K}$. It can be seen that $H = \Lambda\Gamma\Lambda$ and both Γ and Λ are invertible[†]. Let e_j be the K-vector with its jth entry 1 and all other entries 0, thus to show $w_j = w_{K+1-j}$ for $j \in \{1, 2, \ldots, (K-1)/2\}$ (recall that K is odd) is equivalent to show that

$$e_j^T H^{-1}\mathbf{1} = e_{K+1-j}^T H^{-1}\mathbf{1},$$

which is further equivalent to

$$e_j^T \Lambda^{-1}\Gamma^{-1}\Lambda^{-1}\mathbf{1} = e_{K+1-j}^T \Lambda^{-1}\Gamma^{-1}\Lambda^{-1}\mathbf{1}.$$

Therefore it is sufficient to show that $e_j^T\Lambda^{-1} = e_{K+1-j}^T\Lambda^{-1}$. Since $e_j^T\Lambda^{-1}$ gives the jth row of Λ^{-1} and $e_{K+1-j}^T\Lambda^{-1}$ gives the $(K+1-j)$th row of Λ^{-1}, these two quantities are identical since $f(F^{-1}(\tau_j)) = f(F^{-1}(\tau_{K+1-j}))$, by symmetric error assumption. The proof is completed by substituting $S(w)$ in Theorem 4.2 by $S(w^*)$. □

Proof of Theorem 4.4. By the umbrella assumption and Step 1, we have $\|\hat{\alpha} - \alpha\| = O_P(n^{-1/2})$ and $\|\hat{\beta}^0 - \beta\| = O_P(n^{-1/2})$. Since K is fixed, condition E ensues if we can show that (i) $\tilde{F}^{-1}(\tau) = F^{-1}(\tau) + o_P(1)$ for each fixed $\tau \in (0, 1)$; and (ii) $\tilde{f}(z) = f(z) + o_P(1)$ for each fixed z.

[†] We will give the explicit form of Γ^{-1} in the proof of Theorem 4.5.

First, we prove $\tilde{F}^{-1}(\tau) = F^{-1}(\tau) + o_P(1)$. For $\hat{u}_t$ in (4.11), we have

$$\hat{u}_t = \varepsilon_t + \Delta_t,\ \Delta_t = \frac{x_t'\beta - x_t'\hat{\beta}^0}{s(x_t, \hat{\alpha})} + \frac{s(x_t, \alpha) - s(x_t, \hat{\alpha})}{s(x_t, \hat{\alpha})} u_t.$$

Denote $\hat{\alpha} - \alpha$ by $\hat{\delta}$. By (4.25),

$$\left| \frac{x_t'\beta - x_t'\hat{\beta}^0}{s(x_t, \hat{\alpha})} \right| \leq (\|x_t/\sigma_t\| + \|\xi_t/\sigma_t\| \|x_t/\sigma_t\| \|\hat{\delta}\| + c\|x_t/\sigma_t\| \|\hat{\delta}\|)\|\beta - \hat{\beta}^0\|. \quad (4.30)$$

By (4.25), (4.26) and (4.27),

$$\begin{aligned} &\left| \frac{s(x_t, \alpha) - s(x_t, \hat{\alpha})}{s(x_t, \hat{\alpha})} u_t \right| \\ \leq &(\sigma_t^{-1} + \sigma_t^{-1}\|\xi_t/\sigma_t\| \|\hat{\delta}\| + c\|x_t/\sigma_t\| \|\hat{\delta}\|)(\|\xi_t\| \|\hat{\delta}\| + c\|\hat{\delta}\|^2)|u_t| \end{aligned} \quad (4.31)$$

In view of (4.30), (4.31) and conditions SX3, SX4, we have

$$\left| \frac{x_t'\beta - x_t'\hat{\beta}^0}{s(x_t, \hat{\alpha})} \right| = O_P(n^{-1/4}), \quad \left| \frac{s(x_t, \alpha) - s(x_t, \hat{\alpha})}{s(x_t, \hat{\alpha})} u_t \right| = O_P(n^{-1/4}) \quad (4.32)$$

uniformly in t, which implies $\Delta_t = o_P(1)$ uniformly in t. Consequently, $\tilde{F}^{-1}(\tau) = \overline{F}^{-1}(\tau) + o_P(1)$, where $\overline{F}^{-1}(\tau)$ is the τth sample quantile of $u_1, \ldots, u_n$. By standard theory of sample quantiles, $\overline{F}^{-1}(\tau) = F^{-1}(\tau) + o_P(1)$. Therefore $\tilde{F}^{-1}(\tau) = F^{-1}(\tau) + o_P(1)$.

Next, we prove $\tilde{f}(z) = f(z) + o_P(1)$ for each fixed z. Since $K(\cdot)$ is Lipschitz, it follows that

$$\tilde{f}(z) = \frac{1}{nb_n} \sum_{t=1}^{n} K\left(\frac{z - u_t}{b_n}\right) + \frac{O(1)}{nb_n^2} \sum_{t=1}^{n} |\Delta_t|$$

By standard theory of nonparametric kernel density estimation,

$$\frac{1}{nb_n} \sum_{t=1}^{n} K\left(\frac{z - u_t}{b_n}\right) = f(u) + o_P(1). \quad (4.33)$$

Finally, by (4.30), (4.31) and conditions SX3, SX4, it follows that $\sum_{t=1}^{n} |\Delta_t| =$

$O_P(n^{-1/2})$. Hence

$$\frac{1}{nb_n^2}\sum_{t=1}^{n}|\Delta_t| = O_P((\sqrt{n}b_n^2)^{-1}) = o_P(1)$$

in view of $b_n \propto n^{-1/5}$. This completes the proof. □

Proof of Theorem 4.5. The proof can be found in Zhao and Xiao (2014). For completeness, we include their proof here.

Let Γ be defined in the proof of Theorem 4.3, we first show that under the efficiency regularity condition,

$$\lim_{K\to\infty} L'\Gamma^{-1}L = \int_0^1 [\ell'(\tau)]^2 \,\mathrm{d}\tau, \text{ where } L = (\ell(\tau_1),\ldots,\ell(\tau_K))'. \tag{4.34}$$

By the definition of Γ, direct calculation shows that

$$\Gamma^{-1} = (K+1)\begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix}, \tag{4.35}$$

that is, Γ^{-1} is a tri-diagonal matrix with $2(K+1)$ on the diagonal, $-(K+1)$ on the super-/sub-diagonals, and 0 elsewhere.

Let $\delta = 1/(K+1)$. By $\tau_j = j/(K+1)$ and (4.35), direct calculation shows that

$$\begin{aligned} L'\Gamma^{-1}L &= (K+1)\left\{\ell^2(\tau_1) + \ell^2(\tau_K) + \sum_{j=2}^{K}[\ell(\tau_j) - \ell(\tau_{j-1})]^2\right\} \\ &= (K+1)[\ell^2(\tau_1) + \ell^2(\tau_k)] + R_K + \int_\delta^{1-\delta}[\ell'(t)]^2\,\mathrm{d}t, \end{aligned} \tag{4.36}$$

where

$$R_K = -\frac{K+1}{2}\sum_{j=2}^{K}\int_{\tau_{j-1}}^{\tau_j}\int_{\tau_{j-1}}^{\tau_j}[\ell'(t) - \ell'(s)]^2\,\mathrm{d}t\,\mathrm{d}s.$$

For $t, s \in [\tau_{j-1}, \tau_j]$, $|\ell'(t) - \ell'(s)| = |\int_s^t \ell''(v)\,\mathrm{d}v| \le \int_{\tau_{j-1}}^{\tau_j}|\ell''(v)|\,\mathrm{d}v$. By Schwarz inequality,

$$\max_{t,s\in[\tau_{j-1},\tau_j]} |\ell'(t)-\ell'(s)|^2 \le \left[\int_{\tau_{j-1}}^{\tau_j} |\ell''(v)|\,\mathrm{d}v\right]^2 \le \delta \int_{\tau_{j-1}}^{\tau_j} [\ell''(v)]^2\,\mathrm{d}v.$$

It then follows that

$$|R_K| \le \frac{K+1}{2}\sum_{j=2}^{K}(\tau_j-\tau_{j-1})^2 \max_{t,s\in[\tau_{j-1},\tau_j]} |\ell'(t)-\ell'(s)|^2 \le \frac{\delta^2}{2}\int_{\delta}^{1-\delta} |\ell''(t)|^2\,\mathrm{d}t.$$

As $K\to\infty$, $\int_{\delta}^{1-\delta}[\ell'(\tau)]^2\,\mathrm{d}\tau \to \int_0^1[\ell'(\tau)]^2\,\mathrm{d}\tau$, regardless of whether the latter integral is finite or infinite. (4.34) then follows from the efficiency regularity condition of $\ell(\cdot)$.

Recall the definition of L in (4.34), then by the last result of Theorem 4.3 and the intermediate steps in the proof of Theorem 4.3, it follows that

$$\Omega_K = \mathbf{1}'H^{-1}\mathbf{1} = \mathbf{1}'\Lambda^{-1}\Gamma^{-1}\Lambda^{-1}\mathbf{1} = L'\Gamma^{-1}L.$$

It then follows by (4.34) and the chain rule of differentiation that as $K\to\infty$,

$$\Omega_K \to \int_0^1 [\ell'(\tau)]^2\,\mathrm{d}\tau = \int_{\mathbb{R}^1}\frac{[f'(u)]^2}{f(u)}\,\mathrm{d}u = \mathcal{F}(f).$$

This completes the proof. □

Appendix

Applications of Chaining Arguments in Probability and Mathematical Statistics

1 Introduction

We have seen in the proofs of Theorem 2.9 in Chapter 2 and Lemma 4.1 in Chapter 4 that how the point-wise convergence result may be generalized to its strengthened uniform convergence result, thus induces some deeper consequences. In the quantile regression literature, such technique is termed as the "*chaining technique*" or "*chaining arguments*". Nevertheless, this useful technique is by no means the patent of quantile regression researchers — it has roots in the elegant proofs of many classical probability and mathematical statistics theorems. In this appendix, it is our hope to present a comprehensive account for the classical and modern applications of the chaining technique, as well as the general steps to call for it. In Section 2, we review some celebrated historical examples which could be viewed as the prelude of the chaining technique. After unveiling the common feature appeared in the proofs in Section 2, with the possible risk of oversimplification, in Section 3 we streamline the invocation of the chaining technique into three key steps. We conclude this chapter by revisiting several remarkable applications of the chaining technique in the quantile regression literature.

The purpose of this chapter is to give theoretical statisticians an overview of the powerful, yet easy-to-use chaining technique.

2 Historical Examples

In this section, we review four classical probability theorems whose proofs clearly utilized the chaining argument (explicit definition will be given in Section 3) in different ways. These theorems are also from different areas: one-sample uniform weak convergence, uniform strong convergence of empirical distribution functions, uniform strong convergence of general functions, and weak convergence of random functions. The structure of this section's exposition is standard: the theorems are given first and are followed by detailed proofs. The proofs of Theorem A.2 and Theorem A.4 are taken from the references without much change, while the proofs of Theorem A.1 and Theorem A.3 are adapted by the author.

Our first historical example is the well-known *Polyá Theorem* (Polyá (1920)):

Theorem A.1. *Let $X_1, X_2, \ldots$ be a sequence of random variables that converges weakly to a random variable X with continuous distribution function. Then*

$$\sup_{x\in\mathbb{R}} |P[X_n \leq x] - P[X \leq x]| \to 0.$$

Proof. Let F_n and F denote the distribution functions of X_n and X, respectively.

Given $\varepsilon > 0$, choose $k \in \mathbb{N}$ sufficiently large such that $k^{-1} < \varepsilon/2$. Since F is continuous on $\mathbb{R}$, by the intermediate value theorem, there exist $-\infty = x_0 < x_1 < \cdots < x_{k-1} < x_k = +\infty$ such that $F(x_i) = i/k, i = 0, 1, \ldots, k$. For each $x \in \mathbb{R}$, there exists $i \in \{0, 1, \ldots, k-1\}$ such that $x \in (x_i, x_{i+1}]$ (of course, $x \neq x_k$). By monotonicity of F and F_n, we have

$$\begin{aligned}
F_n(x) - F(x) \leq & F_n(x_{i+1}) - F(x_i) = F_n(x_{i+1}) - F(x_{i+1}) + F(x_{i+1}) - F(x_i) \\
\leq & \sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| + \frac{1}{k}; \\
F_n(x) - F(x) \geq & F_n(x_i) - F(x_{i+1}) = F_n(x_i) - F(x_i) + F(x_i) - F(x_{i+1}) \\
\geq & - \sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| - \frac{1}{k},
\end{aligned}$$

which is equivalent to $|F_n(x) - F(x)| \leq \sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| + k^{-1}$. Since this inequality holds for each $x \in \mathbb{R}$, and the upper bound is independent of x, it follows that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| + \frac{1}{k}. \tag{A.1}$$

Since $F_n \Rightarrow F$, and F is continuous everywhere, there exists $N \in \mathbb{N}$ such that $\sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| < \varepsilon/2$ for all $n > N$. Hence for each $n > N$,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \leq \sup_{0\leq j\leq k-1} |F_n(x_j) - F(x_j)| + \frac{1}{k} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This completes the proof. □

The second example is the celebrated *Glivenko-Cantelli Theorem*, which generalizes the point-wise strong law of large numbers to the uniform case. The statement and the proof of the theorem are taken from Billingsley (1995, pp.268–269).

Theorem A.2. *Suppose that $X_1, X_2, \ldots$ are independent and have a common distribution function F, put*

$$D_n(\omega) = \sup_x |F_n(x,\omega) - F(x)|.$$

Then $D_n \to 0$ with probability 1.

Proof. By the strong law of large numbers, for each x there is a set A_x and a set B_x, both are of probability 0 such that $\lim_n F_n(x,\omega) = F(x)$ except on A_x and $\lim_n F_n(x-,\omega) = F(x-)$ except on B_x. Let $\varphi(u) = \inf[x : F(x) \geq u]$ for $0 < u < 1$, and put $x_{m,k} = \varphi(k/m), m \geq 1, 1 \leq k \leq m$. It is not hard to see that $F(\varphi(u)-) \leq u \leq F(\varphi(u))$; hence $F(x_{m,k}) - F(x_{m,k-1}) \leq m^{-1}$, $F(x_{m,1}-) \leq m^{-1}$, and $F(x_{m,m-1}) \geq 1 - m^{-1}$. Let $D_{m,n}(\omega)$ be the maximum of the quantities $|F_n(x_{m,k}) - F(x_{m,k})|$ and $|F_n(x_{m,k}-) - F(x_{m,k}-)|$ for $k = 1, \ldots, m$.

If $x_{m,k-1} \leq x < x_{m,k}$, then $F_n(x,\omega) \leq F_n(x_{m,k}-,\omega) \leq F(x_{m,k}-) + D_{m,n}(\omega) \leq F(x) + m^{-1} + D_{m,n}(\omega)$ and $F_n(x,\omega) \geq F_n(x_{m,k-1},\omega) \geq F(x_{m,k-1}) - D_{m,n}(\omega) \geq F(x) - m^{-1} - D_{m,n}(\omega)$. Together with similar arguments for the cases $x < x_{m,1}$ and $x \geq x_{m,m-1}$, this shows that

$$D_n(\omega) \leq D_{m,n}(\omega) + m^{-1}. \tag{A.2}$$

If ω lies outside the union A of all the $A_{x_{mk}}$ and $B_{x_{mk}}$, then $\lim_n D_{m,n}(\omega) = 0$ and hence $\lim_n D_n(\omega) = 0$ by (A.2). But A has probability 0. This completes the proof. □

Unlike the preceding two pure probabilistic results, the third example below has some clear statistical inference implications — it can be used to show the strong consistency of the sample average of some function $U(x, \hat{\theta}_n)$, provided that $\hat{\theta}_n$ itself is strong consistent. This example is excerpted from Section 16, Ferguson (1996).

Theorem A.3 (A Uniform Strong Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common distribution function $F(x)$, and let $U(x, \theta)$ be a measurable function of x for all θ in some parameter space Θ. If*

1. *Θ is compact,*
2. *$U(x, \theta)$ is continuous in θ for all x,*
3. *There exists a function $K(x)$ such that $E[K(X)] < \infty$ and $|U(x, \theta)| \leq K(x)$ for all x and θ.*

Then

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} U(X_i, \theta) - \mu(\theta) \right| \to 0$$

with probability 1.

To prove this theorem, we need the lemma below, and it is this lemma that illustrates the chaining technique:

Lemma A.1. *If*

1. *Θ is compact,*
2. *$U(x, \theta)$ is upper semicontinuous[*] in θ for all x,*
3. *There exists a function $K(x)$ such that $E[K(X)] < \infty$ and $|U(x, \theta)| < K(x)$ for all x and θ,*

[*] A function $f(\theta)$ is *upper semicontinuous* at θ_0 if for any $\varepsilon > 0$, there is a positive δ such that $|\theta - \theta_0| < \delta$ implies $f(\theta) < f(\theta_0) + \varepsilon$.

4. *For all θ and for all sufficiently small $\rho > 0$, $\sup_{|\theta'-\theta|<\rho} U(x, \theta')$ is measurable in x.*

Then

$$\limsup_n \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} U(X_i, \theta) \leq \sup_{\theta \in \Theta} \mu(\theta)$$

with probability 1.

Proof of Lemma A.1. Define $\varphi(x, \theta, \rho) := \sup_{|\theta'-\theta|<\rho} U(x, \theta')$. By Condition 4, $\varphi(x, \theta, \rho)$ is measurable in x for all θ and all sufficiently small $\rho > 0$. By Condition 2, for all x and fixed $\theta \in \Theta$, $\varphi(x, \theta, \rho) \downarrow U(x, \theta)$ as $\rho \downarrow 0$. Therefore by Condition 3 and Lebesgue's dominated convergence theorem,

$$\int \varphi(x, \theta, \rho) dF(x) \downarrow \int U(x, \theta) dF(x) = \mu(\theta) \tag{A.3}$$

as $\rho \downarrow 0$, for every $\theta \in \Theta$. Given $\varepsilon > 0$, (A.3) then implies that for each $\theta \in \Theta$, there exists a sufficiently small $\rho_\theta > 0$ such that $\int \varphi(x, \theta, \rho_\theta) dF(x) < \mu(\theta) + \varepsilon$. Clearly, the family $\{B(\theta, \rho_\theta) : \theta \in \Theta\}$ forms an open cover of Θ, and since Θ is compact, we can choose a finite sub-cover $\{B(\theta_j, \rho_{\theta_j}) : j = 1, \ldots, m\}$ such that $\Theta = \bigcup_{j=1}^{m} B(\theta_j, \rho_{\theta_j})$. Now for each $\theta \in \Theta$, there exists an index j, $1 \leq j \leq m$ such that $\theta \in B(\theta_j, \rho_{\theta_j})$. Hence by the definition of $\varphi(x, \theta, \rho)$, it follows that $U(X_i, \theta) \leq \varphi(X_i, \theta_j, \rho_{\theta_j})$, $i = 1, \ldots, n$. Thus

$$\frac{1}{n} \sum_{i=1}^{n} U(X_i, \theta) \leq \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i, \theta_j, \rho_{\theta_j}) \leq \sup_{1 \leq j \leq m} \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i, \theta_j, \rho_{\theta_j}).$$

Consequently,

$$\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} U(X_i, \theta) \leq \sup_{1 \leq j \leq m} \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i, \theta_j, \rho_{\theta_j}). \tag{A.4}$$

For each $j \in \{1, \ldots, m\}$, by the strong law of large numbers,

$$\lim_n \frac{1}{n} \sum_{i=1}^{n} \varphi(X_i(\omega), \theta_j, \rho_{\theta_j}) = \int \varphi(x, \theta_j, \rho_{\theta_j}) dF(x) < \mu(\theta_j) + \varepsilon$$

except on a set A_{θ_j} with probability 0. If ω lies outside the union A of all the A_{θ_j}, then $\lim_n \sup_{1\le j\le m} n^{-1}\sum_{i=1}^n \varphi(X_i(\omega),\theta_j,\rho_{\theta_j}) < \sup_{1\le j\le m}\mu(\theta_j)+\varepsilon \le \sup_{\theta\in\Theta}\mu(\theta)+\varepsilon$. Together with (A.4), it can be seen that

$$\limsup_n \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) < \sup_{\theta\in\Theta}\mu(\theta)+\varepsilon$$

with probability 1. Since ε is arbitrary, the result follows. □

Proof of Theorem A.3. Because of Condition 2, Condition 4 in Lemma A.1 is satisfied. By Condition 2 and Lebesgue's dominated convergence theorem, $\mu(\theta)$ is continuous on Θ. Consequently, both $U(x,\theta)-\mu(\theta)$ and $-U(x,\theta)+\mu(\theta)$ are upper semicontinuous in θ for all x. Therefore, we may apply Lemma A.1 to $U(x,\theta)-\mu(\theta)$ and $-U(x,\theta)+\mu(\theta)$ respectively and claim that with probability 1,

$$\limsup_n \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n [U(X_i,\theta)-\mu(\theta)] \le 0,$$
$$\limsup_n \sup_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^n [-U(X_i,\theta)+\mu(\theta)] \le 0,$$

which is equivalent to

$$\lim_n \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n U(X_i,\theta)-\mu(\theta)\right| = 0$$

with probability 1. □

As the final historical example, let's recall the notable *Arzelà-Ascoli theorem*. This theorem has many variants across different analysis branches, among which the following version regarding the characterization of compact subsets of the space $C[0,1]$, is cited from Billingsley (1999, p.81).

Theorem A.4. *Let $C \equiv C[0,1]$ be the space of continuous functions on the unit interval. The set $A \subset C$ is relatively compact*[†] *if and only if*

$$\sup_{x\in A}|x(0)| < \infty$$

[†] A set A is *relatively compact* if its closure $\overline{A}$ is compact.

and the family A of continuous functions is uniformly equicontinuous*, i.e.,*

$$\lim_{\delta\to 0}\sup_{x\in A} w_x(\delta) = 0. \tag{A.5}$$

Here the modulus of continuity *of an arbitrary function $x(\cdot)$ on $[0,1]$ is defined by*

$$w_x(\delta) = w(x,\delta) = \sup_{|s-t|\le\delta} |x(s) - x(t)|, \quad 0 < \delta \le 1. \tag{A.6}$$

Proof. If $\overline{A}$ is compact, then $\sup_{x\in A}|x(0)| \le \sup_{x\in A}\|x\| < \infty$, (A.5) follows. Since $w(x, n^{-1})$ is continuous in x[‡] and nonincreasing in n, $\lim_{\delta\to 0} w_x(\delta) = 0$ holds uniformly on A if $\overline{A}$ is compact (Billingsley (1999, p.242, M8)), and (A.6) holds.

Suppose now (A.5) and (A.6) hold. Choose k large enough that $\sup_{x\in A} w_x(k^{-1})$ is finite. Since

$$|x(t)| \le |x(0)| + \sum_{i=1}^{k} |x(it/k) - x((i-1)t/k)|,$$

it follows that

$$\sup_t \sup_{x\in A} |x(t)| < \infty. \tag{A.7}$$

The idea now is to use (A.7) and (A.6) to prove that A is totally bounded; since C is complete, it will follow that $\overline{A}$ is compact.

Let α be the supremum in (A.7). Given ε, choose a finite ε-net H in the interval $[-\alpha,\alpha]$ on the line, and choose k large enough that $w_x(1/k) < \varepsilon$ for all x in A. Take B to be the finite set consisting of the (polygonal) functions in C that are linear on each interval $I_{ki} = [(i-1)/k, i/k], 1 \le i \le k$, and take values in H at the end points. If $x \in A$, then $|x(i/k)| \le \alpha$, and therefore there is a point y in B such that $|x(i/k) - y(i/k)| < \varepsilon$ for $i = 0, 1, \ldots, k$. Now $y(i/k)$ is within 2ε of $x(t)$ for $t \in I_{ki}$, and similarly for $y((i-1)/k)$. Since $y(t)$ is a convex combination of $y((i-1)/k)$ and $y(i/k)$, it too is within 2ε of $x(t)$: $\rho(x,y) < 2\varepsilon$. Thus B is a finite 2ε-net for A. □

[‡] For fixed positive δ, we have $|w_x(\delta) - w_y(\delta)| \le 2\rho(x,y)$.

3 Meditation: the Problem, the Construction of the Chain and the General Procedure

From the four classical examples introduced in last section, we may have gotten a vague impression about what a *chain* is and how the *chaining arguments* proceeds. To get a clearer view of what kind of gaps that the chaining arguments fills in, we summarize the key conditions and the ultimate goals of the four examples in Section 2 in the Table A.1 below:

Table A.1. *The key conditions and the ultimate goals of the four theorems introduced in Section 2.*

Theorems	Conditions	Goals
Polyá Theorem	$\lvert F_n(x) - F(x)\rvert \to 0$ for fixed $x \in \mathbb{R}^1$	$\sup\limits_{x\in\mathbb{R}^1} \lvert F_n(x) - F(x)\rvert \to 0$
Glivenko-Cantelli Theorem	$\lvert F_n(x,\omega) - F(x)\rvert \to_{a.s.} 0$ for fixed $x \in \mathbb{R}^1$	$\sup\limits_{x\in\mathbb{R}^1} \lvert F_n(x,\omega) - F(x)\rvert \to_{a.s.} 0$
Uniform SLLN	$n^{-1} \sum\limits_{i=1}^{n} U(X_i,\theta) \to \mu(\theta)$ for fixed $\theta \in \Theta$	$\sup\limits_{\theta\in\Theta} \left\lvert n^{-1} \sum\limits_{i=1}^{n} U(X_i,\theta) - \mu(\theta)\right\rvert \to 0$
Arzelà-Ascolli Theorem	Uniformly bounded + Uniformly equicontinuous	Relative Compactness

In Table A.1, except for the last example for which we will uncover its connection to the chaining arguments later, all the theorems share the common feature that a point-wise result must be generalized to its uniform counterpart. Specifically, we may refer the set (e.g., $\mathbb{R}^1$, Θ) in the "Conditions" column as the *parameter space*. At the outset, an assertion S holds only for a fixed point in the parameter space (i.e., point-wisely), and the goal is to extend S to the whole parameter space. Such extensions are necessary as intermediate steps in establishing many asymptotic theorems (as may be seen from the examples in Section 4) and it is the chaining arguments that abridge the gap between them. Before giving a formal definition of a general chain, let's first find out what they are in the preceding four examples:

Polyá Theorem:

$$C = \{x_0, x_1, \ldots, x_k\}$$

Glivenko-Cantelli Theorem:

$$C = \{x_{m1}, x_{m2}, \ldots, x_{mm}\}$$

Uniform SLLN:

$$C = \{\theta_1, \ldots, \theta_m\}$$

Arzelà-Ascolli Theorem:

$$C = \{y \in C[0,1] : y \in B\}$$

In the above list, we use the generic symbol "C" to denote a chain, which is a *finite* set consisting of *representative* points from the parameter space. The key word "finite" is easily understood, whereas by saying "representative", we mean that the difference between suprema of the quantity of interest over the whole parameter space and C is negligible when the sample size n is large. This statement can be verified by carefully examining the proofs in Section 2. Now that we are clear on problems type that the chaining technique applies, it is time to give a formal definition of the chain and the chaining argument.

Definition A.1. Given a parameter space Θ whose cardinality is typically a continuum, suppose that for each $\theta \in \Theta$, the proposition $S(\theta)$ is true. The *chaining technique*, or the *chaining arguments* is a collection of mathematical statements that proves S is true uniformly on Θ. In the course of these statements, we need to construct a finite set $C = \{\theta_1, \ldots, \theta_m\} \subset \Theta$, called a *chain*, through which the final uniform assertion can be declared. The elements $\theta_i, i = 1, \ldots, m$ of C are called the *knots* of the chain.

In the above definition, everything is clear except the sentence "through which the final uniform assertion can be declared." This is due to we cannot include every case of how to achieve the uniform goal from the initial point-wise condition with the help of C in a generic definition. Nevertheless, in the case of establishing uniform convergence results where the chaining technique receives its popularity, there do exist some routine steps as we have illustrated in the first three examples in Section 2. Loosely speaking, in order to show $\sup\limits_{\theta \in \Theta} |D_n(\theta)| \to 0$ with any specified convergence mode, by cleverly constructing the chain C, we manage to show that

$$\sup_{\theta \in \Theta} |D_n(\theta)| \leq \sup_{\theta \in C} |D_n(\theta)| + \epsilon, \tag{A.8}$$

where ϵ is a quantity that can be made arbitrarily small in advance. In this way we successfully reduces the supremum over an uncountable set Θ to the supremum over a finite set C, thus the point-wise condition is immediately applicable.

There are, of course, much subtleties on determining C and establishing the inequality (A.8). In general, both of these two tasks closely rely on the properties of the parameter space Θ and that of the objects under investigation. For instance, the chain in the proof of Polyá Theorem depends on the ordering property of the real line and the continuity of F, whereas when showing (A.1), which is a special case of (A.8), the monotonicity of F must be employed. Consequently, it is important to keep in mind that the chaining technique may not be suitable for every problem that requires a uniform generalization — the structure and the condition of the problem does matter.

It is also remarkable that the point-wise condition in the proof is not always readily available, on the contrary, it's not unusual that establishing the point-wise condition is much harder than invoking the chaining arguments itself. To see this, let's just compare the difficulty of the proof of the strong law of large numbers and the chaining arguments used in the proof of Glivenko-Cantelli Theorem (the proof itself is short since we all take the validity of the point-wise SLLN for granted). This point is echoed in many papers in the quantile regression literature, where some advanced exponential inequalities must be called for to establish the point-wise preliminaries.

In summary, a complete chaining arguments can be formed by the following three steps:

Step 1: Establish the point-wise proposition.

Step 2: Based on the special structure of the problem at hand, construct a chain by which a passage from the point-wise result to the uniform goal is possible.

Step 3: Conclude the proof by establishing an inequality similar to (A.8).

So far, it remains to explain how the proof of the Arzelà-Ascolli Theorem connects the chaining philosophy introduced above. In that proof, instead of proving any uniform convergence result, we are contented to determine the chain itself. Therefore only the Step 2 in the above list is needed. Nevertheless, it fully demonstrates the technicalities during the process of constructing a chain and how a chain with only finitely many members can approximate any member in an uncountable family with any pre-specified accuracy. From this example, it may also be seen that the usage of the term “chain” is in fact quite random — the word “net” could be an equally descriptive term.

4 Modern Applications

In this section, after summarizing the general philosophy of the chaining arguments, we illustrate this important technique by reviewing some applications from journal papers.

Example 1. Our first example comes from Bickel (1975, Lemma 4.1). In this paper, for the linear model $Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i, 1 \le i \le n$, the author studied the asymptotic behavior of the *one-step estimator* $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ satisfying

$$\sum_{i=1}^{n} x_{ij}\psi(Y_i(\beta^*)) = \sum_{k=1}^{p}(\hat{\beta}_k - \beta_k^*)\sum_{i=1}^{n} x_{ik}x_{ij}\psi'(Y_i(\beta^*)),\ 1 \le j \le p.$$

where β^* is any given preliminary estimator, $Y_i(t) = Y_i - \sum_{j=1}^{p} x_{ij}t_j$ for $t = (t_1, \dots, t_p)'$. To show the asymptotic normality of $\hat{\beta}$, the crucial step is to show the small perturbation $T_n(t) - T_n(0)$ is uniformly negligible for t small enough. Here

$$T_n(t) := \frac{1}{\sqrt{n}}\sum_{i=1}^{n} c_i\{\psi(Y_i(t)) - E[\psi(Y_i(t))]\}$$

where $c_i = x_{i1}$. The asymptotic behavior of $T_n(t) - T_n(0)$ is summarized as the following lemma:

Lemma A.2. *Assume the following conditions hold:*

G *The matrices $n^{-1}X'X$ tend as $n \to \infty$ to a limit X_0 which is positive definite. Further, $\|X\|_\infty = \max_{i,j}|x_{ij}| = o(\sqrt{n})$.*

C *The function ψ is nondecreasing and satisfies*

$$\int_{-\infty}^{\infty}(\psi(x+h) - \psi(x))^2 dF(x) = O(1),$$

$$\int_{-\infty}^{\infty} (\psi(x+h) - \psi(x-h))^2 dF(x) = O(1)$$

as $h \to 0$.

In addition, assume $\beta = 0$. Then for a generic constant M,

$$\sup_{\|t\| \leq M/\sqrt{n}} |T_n(t) - T_n(0)| \to_P 0,$$

where we use $\|t\|$ to denote the maximum of the absolute values of the coordinates of t.

Proof. **Step 1:** Establish the point-wise convergence result for $T_n(t/\sqrt{n}) - T_n(0)$:

First show that for fixed t such that $\|t\| \leq M$, $T_n(t/\sqrt{n}) - T_n(0) \to_P 0$. To see this, compute

$$\begin{aligned}
&E\left[\left\{T_n(t/\sqrt{n}) - T_n(0)\right\}^2\right] \\
=&\frac{1}{n} \operatorname{Var}\left(\sum_{i=1}^{n} c_i[\psi(Y_i(t/\sqrt{n})) - \psi(Y_i(0))]\right) \\
\leq&\frac{1}{n}\sum_{i=1}^{n} c_i^2 E\left[\left\{\psi(\varepsilon_i - \sum_{j=1}^{p} x_{ij}t_j/\sqrt{n}) - \psi(\varepsilon_i)\right\}^2\right] \\
=&\frac{1}{n}\sum_{i=1}^{n} c_i^2 \int_{-\infty}^{\infty} [\psi(s - \sum_{j=1}^{p} x_{ij}t_j/\sqrt{n}) - \psi(s)]^2 f(s)\,\mathrm{d}s \\
\leq&\frac{1}{n}\sum_{i=1}^{n} c_i^2 \sup_{\|h\| \leq pM\|X\|_\infty/\sqrt{n}} \int_{-\infty}^{\infty} [\psi(s+h) - \psi(s)]^2 f(s)\,\mathrm{d}s \to 0
\end{aligned}$$

as $n \to \infty$, by Condition G and Condition C.

Step 2: Construct the chain, and argue that the value of the objective function at each point of the original parameter set can be well approximated by that of one knot of the chain.

Now, clearly, the "parameter set under investigation" is

$$\Theta = \{t \in \mathbb{R}^p : \|t\| \leq M/\sqrt{n}\}$$

For a given positive δ, we have $\Theta \subset K := \{t \in \mathbb{R}^p : \|t\| \leq (\lfloor \delta^{-1} \rfloor + 1)\delta M/\sqrt{n}\}$.

We will construct a finite "chain" for the slightly larger parameter set K. Intuitively, the chain consists of grid points that evenly divide the cube K. Mathematically, it is the finite set:

$$\boxed{C := \{(j_1\delta M/\sqrt{n}, \ldots, j_p\delta M/\sqrt{n}) : j_i \in \{0, \pm1, \ldots, \pm\lfloor\delta^{-1}\rfloor + 1\}, 1 \le i \le p\}.}$$

If $\|t\| \le M/\sqrt{n}$, let $P(t) \in C$ be the lowest vertex of the small cube containing t. For fixed δ, by **Step 1**:

$$\max_{\|t\|\le M/\sqrt{n}} |T_n(P(t)) - T_n(0)| \to_P 0. \tag{A.9}$$

On the other hand, let K_1 be any cube of the partition and let P_1 be its lowest vertex. Then, by monotonicity of ψ, it follows that

$$\begin{aligned}
&\sup_{t\in K_1} |T_n(t) - T_n(P_1)| \\
=&\sup_{t\in K_1} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n c_i\{[\psi(Y_i(P_1)) - \psi(Y_i(t))] - E[\psi(Y_i(P_1)) - \psi(Y_i(t))]\} \right| \\
\le&\sup_{t\in K_1} \frac{1}{\sqrt{n}} \sum_{i=1}^n |c_i|\{|\psi(Y_i(P_1)) - \psi(Y_i(t))| + E[|\psi(Y_i(P_1)) - \psi(Y_i(t))|]\} \\
\le&\frac{1}{\sqrt{n}} \sum_{i=1}^n |c_i|[\psi(Y_i(P_1) + M\delta S_i/\sqrt{n}) - \psi(Y_i(P_1) - M\delta S_i/\sqrt{n})] \\
&+ \frac{1}{\sqrt{n}} \sum_{i=1}^n |c_i|E[\psi(Y_i(P_1) + M\delta S_i/\sqrt{n}) - \psi(Y_i(P_1) - M\delta S_i/\sqrt{n})] \\
=:&I_{1n} + I_{2n},
\end{aligned}$$

where $S_i = \sum_{j=1}^p |x_{ij}|$.

To show $\sup_{t\in K_1} |T_n(t) - T_n(P_1)| = o_P(1)$, we first show that $I_{2n} = \delta O(1)$. By monotonicity of ψ, I_{2n} equals to

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^n |c_i| \int \left[\psi\left(s - \sum_{j=1}^p x_{ij}t_j^* + \frac{M\delta S_i}{\sqrt{n}}\right) - \psi\left(s - \sum_{j=1}^p x_{ij}t_j^* - \frac{M\delta S_i}{\sqrt{n}}\right)\right] \mathrm{d}F(s) \\
\le&\frac{1}{\sqrt{n}} \sum_{i=1}^n |c_i| \sup_{|q|\le M(\delta+1)S_j/\sqrt{n}} \int_{-\infty}^{\infty} \left[\psi\left(s + q + \frac{2M\delta S_i}{\sqrt{n}}\right) - \psi(s+q)\right] \mathrm{d}F(s)
\end{aligned}$$

$$
\begin{aligned}
&=\frac{2M\delta}{n}\sum_{i=1}^{n}|c_i|\sup_{\substack{|q|\le M(\delta+1)S_i/\sqrt{n}\\|h|\le 2M\delta S_i/\sqrt{n}}}\frac{1}{|h|}\int_{-\infty}^{\infty}|\psi\left(s+q+h\right)-\psi\left(s+q\right)|\,\mathrm{d}F(s)\\
&\le\frac{2M\delta}{n}\sum_{i=1}^{n}|c_i|\sup_{\substack{|q|\le M(\delta+1)p\|X\|_\infty/\sqrt{n}\\|h|\le 2M\delta p\|X\|_\infty/\sqrt{n}}}\frac{1}{|h|}\int_{-\infty}^{\infty}|\psi\left(s+q+h\right)-\psi\left(s+q\right)|\,\mathrm{d}F(s)\\
&=O(1)\delta\frac{1}{n}\sum_{i=1}^{n}|x_{i1}|\quad\text{(by Condition G and C, and the definition of } c_i\text{)}\\
&\le O(\delta)\frac{1}{n}\sum_{i=1}^{n}x_{i1}^2\quad\text{(by Cauchy-Schwarz inequality)}\\
&=O(1)\delta\quad\text{(by Condition G)}
\end{aligned}
$$

Here we use $(t_1^*,\dots,t_p^*)'$ to denote the coordinates of P_1.

Next, by the similar argument to Step 1, it can be shown that the variance of I_{1n} is bounded by

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n}c_i^2}{n}\int_{-\infty}^{\infty}\left[\psi\left(s-\sum_{j=1}^{p}x_{ij}t_j^*+\frac{M\delta S_i}{\sqrt{n}}\right)-\psi\left(s-\sum_{j=1}^{p}x_{ij}t_j^*-\frac{M\delta S_i}{\sqrt{n}}\right)\right]^2\mathrm{d}F(s)\\
\le&\frac{1}{n}\sum_{i=1}^{n}c_i^2\int_{-\infty}^{\infty}\left[\psi\left(s+\frac{M(\delta+1)S_i}{\sqrt{n}}\right)-\psi\left(s-\frac{M(\delta+1)S_i}{\sqrt{n}}\right)\right]^2\mathrm{d}F(s)\\
\le&\frac{1}{n}\sum_{i=1}^{n}c_i^2\sup_{\|h\|\le pM(\delta+1)\|X\|_\infty/\sqrt{n}}\int_{-\infty}^{\infty}[\psi(x+h)-\psi(x-h)]^2f(s)ds\to 0
\end{aligned}
$$

as $n\to\infty$. In summary, we showed that

$$\sup_{t\in K_1}|T_n(t)-T_n(P_1(t))|=o_P(1)+\delta O(1).$$

Clearly, the quantity $\sup_{t\in\Theta}|T_n(t)-T_n(P_1(t))|=\max_{K_1}\sup_{t\in K_1}|T_n(t)-T_n(P_1(t))|$ has the same order as the above expression.

Step 3: Use the arbitrariness of the length between knots, and finish the proof. Because the δ used in Step 2 can be arbitrarily small, in fact, we proved that

$$\sup_{t\in\Theta}|T_n(t)-T_n(P_1(t))|=o_P(1).\tag{A.10}$$

It then follows by (A.9) and (A.10) that

$$\begin{aligned}&\sup_{t\in\Theta}|T_n(t)-T_n(0)|\\ \le&\sup_{t\in\Theta}|T_n(t)-T_n(P_1(t))|+\max_{t\in\Theta}|T_n(P_1(t))-T_n(0)|\\ =&o_P(1)+o_P(1)=o_P(1).\end{aligned}$$

This completes the proof. □

The above example may be considered as a revival of the chaining arguments (although its conditions, and details of the proofs have minor mistakes, which have been all corrected in this report) and it soon receives wide popularity, particularly in the theoretical quantile regression literature. We will review several applications in the subsequent examples. Although these applications vary across different models and settings, the invocation of the chaining arguments are generally quite routine — it always follows the three steps summarized in Section 3, as demonstrated in Example 1.

In the quantile regression literature, the chaining arguments is needed when we want to establish the *uniform* Bahadur representation of the regression quantiles, for which a key step is to show that the subgradient function (a function in Δ) of the objective function can be approximated uniformly by an affine transformation of Δ. To achieve this goal, the chaining arguments is an indispensable tool. In the Table A.2 below, I listed some representative applications of the chaining arguments, from four different journal papers. For space restriction, the precise meanings of notations are not listed in the same table and may be found by consulting corresponding papers.

Although the models and the quantile estimates vary from paper to paper, it is important to note from the third column of Table A.2 that all these papers aim to show some "uniform negligibility in probability". In fact, the proofs of these goals are essentially similar to each other — the chaining technique must be invoked, and the three steps summarized in Section 3 are closely followed. In the remaining of this section, I will illustrate this procedure by going over the proof of Lemma 2.3 in Koenker and Zhao (1994). In its original proof, the author omitted the details of Step 2 of the chaining arguments. Here I supplemented this missing part, as well as corrected some notational typos appeared in the original proof.

Table A.2. *Four applications of the chaining technique from four journal papers. The first two papers concern with the homoscedastic linear models, while the third paper and the fourth paper study the location-scale model and the ARCH-type model, respectively. For each of these papers, the primary interest is to establish the Bahadur representation of the proposed quantile estimates, during which the key step is listed in the "Goal" column here.*

Paper	Model	Goal
Koenker and Portnoy (1987)	$y_i = x_i'\beta + u_i$	$\sup_{\tau\in[\varepsilon,1-\varepsilon]} \|\hat{\beta}(\tau) - \beta(\tau)\| = O_P(\sqrt{\log n/n})$
Gutenbrunner et al. (1993)	$y_i = x_i'\beta + u_i$	$\sup_{\substack{\|t\|\le C\sqrt{\log_2 n}, \\ \tau\in[\tau_n^*,1-\tau_n^*]}} \lvert r_n(t,\alpha)\rvert = o_P(1)$
Koenker and Zhao (1994)	$y_i = x_i'\beta + (x_i'\gamma)u_i$	$\sup_{D_{n\varepsilon}} \|\hat{V}(\Delta,\tau) - V(0,\tau) + f(F^{-1}(\tau))Q_2\Delta\| = O_P(n^{-1/4}\log n)$
Koenker and Zhao (1996)	$y_t = u_t(\gamma_0 + \sum_{\ell=1}^{q} \gamma_\ell \lvert y_{t-\ell}\rvert)$	$\sup_{\|\Delta\|\le M} \|V(\Delta) - V(0) + f(F^{-1}(\tau))G\Delta\| = o_P(1)$

Example 2. This example is Lemma 2.3 in Koenker and Zhao (1994).

Lemma A.3. *Define* $D_{n\varepsilon} = \{(\Delta,\tau) : \tau \in [\varepsilon, 1-\varepsilon], \|\Delta\| \le K\sqrt{\log n}\}$, $T(\Delta,\tau) = n^{-1/2}\sum_{i=1}^{n} \sigma_i^{-2} x_i x_i' \psi_\tau(u_i - F^{-1}(\tau) - n^{-1/2}\sigma_i^{-1}x_i'\Delta)$*, then under C1–C5 (see the paper for details),*

$$\sup_{D_{n\varepsilon}} \|T(\Delta,\varepsilon) - T(0,\tau) - E[T(\Delta,\tau)]\| = O_P(\log n).$$

Proof. Write $T(\Delta,\tau) = [T_{jk}(\Delta,\tau)]_{j=1,\ldots,p}^{k=1,\ldots,p}$, where

$$T_{jk}(\Delta,\tau) = n^{-1/2}\sum_{i=1}^{n} v_{ijk}\psi_\tau(u_i - F^{-1}(\tau) - n^{-1/2}\sigma_i^{-1}x_i'\Delta)$$

with $v_{ijk} = e_j' x_i x_i' e_k/\sigma_i^2$ is the (j,k)th entry of matrix $x_i x_i'/\sigma_i^2, i = 1,\ldots,n$. Here, as convention, $e_j = (0,\ldots,1,\ldots,0)'$ denotes the vector in $\mathbb{R}^p$ whose jth component is 1 and all the others are 0.

It is worth pointing out that since $T(\Delta,\tau)$ is a matrix of finite dimensions, the asymptotic behavior (such as tightness) of $T(\Delta,\tau)$ is implied by those of its individual entries $T_{jk}(\Delta,\tau)$. This is the standpoint we will take in the proof.

The first goal is to show that for any fixed $\Delta \in D_{n\varepsilon}$ and any $j, k \in \{1, \ldots, p\}$, $T_{jk}(\Delta, \tau) - T_{jk}(0, \tau) - E[T_{jk}(\Delta, \tau)] = O_P(\log n)$, which is implied by for any $\lambda > 0$,

$$P\left[|T_{jk}(\Delta, \tau) - T_{jk}(0, \tau) - E[T_{jk}(\Delta, \tau)]| \geq \lambda \log n\right] \leq 2e^{-\lambda \log n(1+o(1))}.$$

For ease of notation, denote $n^{1/2}(T_{jk}(\Delta, \tau) - T_{jk}(0, \tau) - E[T_{jk}(\Delta, \tau)])$ by $\tilde{T}_{jk} \equiv \tilde{T}_{jk}(\Delta, \tau)$. Then the goal becomes

$$P[|\tilde{T}_{jk}| \geq \lambda n^{1/2} \log n] \leq 2e^{-\lambda \log n(1+o(1))}. \tag{A.11}$$

Following the proof of Lemma A.2 in Koenker and Portnoy (1987), let $M_{jk}(t) \equiv E[\exp(t\tilde{T}_{jk})]$ denote the moment generating function of $\tilde{T}_{jk}$. Since the summands of $\tilde{T}_{jk}$ are mutually independent, it follows that $M_{jk}(t) = \prod_{i=1}^{n} M_{ijk}(t)$, where

$$\begin{aligned} M_{ijk}(t) = E\{&\exp[tv_{ijk}(\psi_\tau(u_i - F^{-1}(\tau) - n^{-1/2}\sigma_i^{-1}x_i'\Delta) - \psi_\tau(u_i - F^{-1}(\tau)) \\ &- E\psi_\tau(u_i - F^{-1}(\tau) - n^{-1/2}\sigma_i^{-1}x_i'\Delta)]\} \end{aligned}$$

For $\lambda_n > 0$, $t > 0$, by Markov's inequality (the introduction of the parameter t is noteworthy):

$$P[|\tilde{T}_{jk}| \geq \lambda_n] \leq e^{-t\lambda_n}(M_{jk}(t) + M_{jk}(-t)). \tag{A.12}$$

Since $\log M_{jk}(t) = \sum_{i=1}^{n} \log M_{ijk}(t)$, in order to bound $M_{jk}(t)$, first consider bounding $M_{ijk}(t)$ individually. We have the following three cases:

Case 1: $x_i'\Delta = 0$. This case is trivial, for which $M_{ijk}(t) = 1$.

Case 2: $x_i'\Delta > 0$. Denote by p_i the probability $P[F^{-1}(\tau) \leq u_i < F^{-1}(\tau) + n^{-1/2}\sigma_i^{-1}x_i'\Delta]$, then

$$\begin{aligned} &M_{ijk}(t) \\ =&E\left\{\exp\left[tv_{ijk}(p_i - I(F^{-1}(\tau) \leq u_i < F^{-1}(\tau) + n^{-1/2}\sigma_i^{-1}x_i'\Delta))\right]\right\} \\ =&p_i \exp(-tv_{ijk}(1 - p_i)) + (1 - p_i)\exp(tv_{ijk}p_i) \\ \leq&1 + 2p_i(tv_{ijk})^2 \exp(|tv_{ijk}|). \end{aligned}$$

Case 3: $x_i'\Delta < 0$. Denote by q_i the probability $P[F^{-1}(\tau) + n^{-1/2}\sigma_i^{-1}x_i'\Delta \le u_i < F^{-1}(\tau)]$. Similar calculation to case 2 yields

$$M_{ijk}(t) \le 1 + 2q_i(tv_{ijk})^2 \exp(|tv_{ij}|).$$

In a compact way, all of the three cases above can be summarized as $M_{ijk}(t) \le 1 + 2\,\mathrm{sgn}(x_i'\Delta)(P[u_i < F^{-1}(\tau) + n^{-1/2}\sigma_i^{-1}x_i'\Delta] - \tau)(tv_{ijk})^2 \exp(|tv_{ijk}|), t \in \mathbb{R}^1$. Regardless the sign of $x_i'\Delta$, by C5, there exists a constant c independent of i such that

$$|p_i| = |q_i| \le cn^{-1/2}\sigma_i^{-1}|x_i'\Delta|.$$

The inequality $\log(1 + x) \le x$ for $x \ge 0$ then implies

$$\log M_{ijk}(t) \le cn^{-1/2}\sigma_i^{-1}|x_i'\Delta|(tv_{ijk})^2 \exp(|tv_{ijk}|) \tag{A.13}$$

By Schwarz inequality and C4,

$$|v_{ijk}| = \frac{|e_j'x_ix_i'e_k|}{\sigma_i^2} \le \left\|\frac{x_i}{\sigma_i}\right\|^2 = O(n^{1/2}), \tag{A.14}$$

$$\exp(|tv_{ijk}|) \le \exp(|t|\|x_i/\sigma_i\|^2) \le \exp(Bn^{1/2}|t|) \tag{A.15}$$

uniformly in i, for some constant $B > 0$. (A.13), (A.14), (A.15), and C3 then implies

$$\begin{aligned}\log M_{jk}(t) &\le cn^{-1/2}\sum_{i=1}^n \sigma_i^{-1}|x_i'\Delta|(tv_{ijk})^2 \exp(|tv_{ijk}|)\\ &\le cn^{-1/2}\sum_{i=1}^n \sigma_i^{-1}\|x_i\|\|\Delta\||t|^2\|x_i/\sigma_i\|^4 \exp(Bn^{1/2}|t|)\\ &\le c'\|\Delta\|\sum_{i=1}^n \|x_i/\sigma_i\|^3|t|^2 \exp(Bn^{1/2}|t|)\\ &\le c''n\sqrt{\log n}|t|^2 \exp(Bn^{1/2}|t|),\end{aligned}$$

where c', c'' are constants do not depend on n.

Note the above bound holds for all $t \in \mathbb{R}^1$, then recall our goal (A.11), by taking

$\lambda_n = \lambda n^{1/2}\log n, t = n^{-1/2}$ in (A.12) yields that

$$\begin{aligned} P[|\tilde{T}_{jk}| \geq \lambda n^{1/2}\log n] &\leq 2\exp(-\lambda \log n + c''\sqrt{\log n} e^B) \\ &= 2\exp(-\lambda \log n(1+o(1))). \end{aligned} \tag{A.16}$$

Hence (A.11) follows.

Having proved (A.11), the next step is to generalize the point-wise result to its uniform counterpart (4.7). The approach is the classical "chaining argument", which is originated from the seminal paper Bickel (1975).

In the following, we use $a, a_1, a_2, \ldots$ to denote generic constants.

For clarity, denote the interval $[\varepsilon, 1-\varepsilon]$ by I_ε, $\{\Delta \in \mathbb{R}^p : \|\Delta\| \leq K\sqrt{\log n}\}$ by H_n so that $D_{n\varepsilon} = I_\varepsilon \times H_n$. Since we use L_2 norm throughout the paper, the set H_n is a ball in the Euclidean space $\mathbb{R}^p$. Accordingly, let $I_n := [-K\sqrt{\log n}, K\sqrt{\log n}] \times \cdots \times [-K\sqrt{\log n}, K\sqrt{\log n}]$ be the smallest rectangle containing H_n. As the beginning step of chaining arguments, we arrange a finite set of grid points to decompose I_n into small cube. In detail, since it is desirable that the diameter of each small cube is not greater than n^{-3}, we set

$$P_n := \{j\delta_n : j \in \{0, \pm 1, \ldots, \pm\lfloor K\sqrt{\log n}/\delta_n\rfloor\}\} \cup \{-K\sqrt{\log n}, K\sqrt{\log n}\},$$

where $\delta_n = n^{-3}/\sqrt{p}$. In the same manner, decompose I_ε by small intervals whose lengths are not greater than n^{-3} by letting

$$P_0 := \{\varepsilon + i/n^{-3} : i = 0, \ldots, \lfloor(1-2\varepsilon)n^3\rfloor\} \cup \{1-2\varepsilon\}.$$

We then define the set of vertices (grid points) that "supporting" the set $D_{n\varepsilon}$ as:

$$V := (P_0 \times P_n \times \cdots \times P_n) \cap D_{n\varepsilon}$$

It is easily seen that the cardinality of V, which we shall denote by N, is bounded above by

$$(\lfloor(1-2\varepsilon)n^3\rfloor + 2) \times (2\lfloor K\sqrt{\log n}/\delta_n\rfloor + 3)^p \leq a n^{3p+3}(\log n)^{p/2}$$

for some constant a independent of n.

By (A.16), we have

$$\begin{aligned} &P\left[\sup_{(\tau,\Delta)\in V}|\tilde{T}_{jk}(\Delta,\tau)| \geq (3p+5)n^{1/2}\log n\right] \\ \leq &an^{3p+3}(\log n)^{p/2} \times 2\exp(-(3p+5)\log n(1+o(1))) \to 0 \end{aligned} \tag{A.17}$$

as $n \to \infty$.

On the other hand, let $\mathscr{C}$ be the collection of disjoint cubes whose vertices belong to $P_n \times \cdots \times P_n$ so that for each $C \in \mathscr{C}$, diam $C \leq n^{-3}$. For any $(\tau, \Delta) \in D_{n\varepsilon}$, let $P(\tau)$ be the greatest number in P_0 such that $P(\tau) \leq \tau$ and $P(\Delta)$ be the lowest vertex of $C \in \mathscr{C}$ that contains Δ (without loss of generality, we assume for every Δ, $P(\Delta) \in D_{n\varepsilon}$.). For simplicity, write $\tau^* = P(\tau)$, $\Delta^* = P(\Delta)$, $\tilde{\Delta} = n^{-1/2}\Delta$, $\tilde{\Delta}^* = n^{-1/2}\Delta$. Also without lost of generality, assume $\sigma_i \equiv 1$. Then

$$|\tilde{T}_{jk}(\Delta,\tau) - \tilde{T}_{jk}(\Delta^*,\tau^*)| \leq T_1 + T_2 + T_3 + T_4,$$

where

$$\begin{aligned} T_1 &= \sum_{i=1}^{n} |v_{ijk}||I(u_i < F^{-1}(\tau^*) + x_i'\tilde{\Delta}^*) - I(u_i < F^{-1}(\tau) + x_i'\tilde{\Delta})| \\ T_2 &= \sum_{i=1}^{n} |v_{ijk}||I(u_i < F^{-1}(\tau^*)) - I(u_i < F^{-1}(\tau))| \\ T_3 &= \sum_{i=1}^{n} |v_{ijk}||\tau^* - \tau| \\ T_4 &= \sum_{i=1}^{n} |v_{ijk}||F(F^{-1}(\tau^*) + x_i'\tilde{\Delta}^*) - F(F^{-1}(\tau) + x_i'\tilde{\Delta})|. \end{aligned}$$

By construction, also recall $|v_{ijk}| = O(n^{1/2})$, $T_3 = O(n \cdot n^{1/2} \cdot n^{-3}) = O(n^{-3/2})$.

Since $\tau > \tau^*$, we have $E[T_2] = \sum_{i=1}^{n} |v_{ijk}|(F(F^{-1}(\tau^*)) - F(F^{-1}(\tau)) = O(n^{-3/2})$. The last equality holds because by Lagarange mean value theorem, there exists $\xi \in (\tau^*, \tau)$ such that

$$F(F^{-1}(\tau^*)) - F(F^{-1}(\tau)) = s'(\xi)(\tau - \tau^*) = O(n^{-3}).$$

In above we also used assumption C5. By Markov inequality, it follows that $T_2 = O_P(n^{-3/2})$.

To deal with T_4, decompose $F(F^{-1}(\tau) + x_i'\tilde{\Delta}) - F(F^{-1}(\tau^*) + x_i'\tilde{\Delta}^*)$ as $J_1 + J_2$, where

$$\begin{aligned} J_1 &= F(F^{-1}(\tau) + x_i'\tilde{\Delta}) - F(F^{-1}(\tau) + x_i'\tilde{\Delta}^*) \\ J_2 &= F(F^{-1}(\tau) + x_i'\tilde{\Delta}^*) - F(F^{-1}(\tau^*) + x_i'\tilde{\Delta}^*) \end{aligned}$$

By Taylor's theorem (with integral remainder), it follows that

$$\begin{aligned} J_1 &= f(F^{-1}(\tau) + x_i'\tilde{\Delta}^*)x_i'(\tilde{\Delta} - \tilde{\Delta}^*) \\ &\quad + \int_0^1 (1-s)f'(F^{-1}(\tau) + x_i'\tilde{\Delta}^* + sx_i'(\tilde{\Delta} - \tilde{\Delta}^*))\,\mathrm{d}s \cdot [x_i'(\tilde{\Delta} - \tilde{\Delta}^*)]^2 \end{aligned}$$

It can be shown that under C5, both f and f' are bounded on $[F^{-1}(\varepsilon), F^{-1}(1-\varepsilon)]$, this fact and that $x_i'\tilde{\Delta}^* = o(1)$ implies that there exists a constant a independent of n such that (to be perfectly rigorous, we may slightly strengthen assumption C5 so that under consideration is an interior point of $[F^{-1}(\varepsilon), F^{-1}(1-\varepsilon)]$.)

$$|J_1| \le a(\|x_i\|\|\tilde{\Delta} - \tilde{\Delta}^*\| + \|x_i\|^2\|\tilde{\Delta} - \tilde{\Delta}^*\|^2) = O(n^{-11/4}).$$

Similarly, for J_2, we have

$$\begin{aligned} J_2 &= f(F^{-1}(\tau^*) + x_i'\tilde{\Delta}^*)(F^{-1}(\tau) - F^{-1}(\tau^*)) \\ &\quad + \int_0^1 (1-s)f'(F^{-1}(\tau^*) + x_i'\tilde{\Delta}^* + sx_i'(F^{-1}(\tau) - F^{-1}(\tau^*))))\,\mathrm{d}s \\ &\cdot[F^{-1}(\tau) - F^{-1}(\tau^*)]^2 \end{aligned}$$

Once more application of mean value theorem yields that $F^{-1}(\tau) - F^{-1}(\tau^*) = O(n^{-3})$, hence the above expansion implies that $J_2 = O(n^{-3})$. Collecting these results, we found that $I_3 = O(n \cdot n^{1/2} \cdot n^{-11/4}) = O(n^{1/4})$.

In the same line with getting the order for T_2 from T_3, it can be verified by using the order of T_4 that $T_1 = O_P(n^{1/4})$.

In summary,

$$\tilde{T}_{jk}(\Delta, \tau) - \tilde{T}_{jk}(\Delta^*, \tau^*) = O_P(n^{1/4})$$

uniformly in (τ, Δ) (the uniformity comes from that all the upper bounds obtained above are independent of (Δ, τ)). In other words, there exists a large M such that (we retrieve the notation $P(\tau)$ and $P(\Delta)$ here):

$$P\left[\sup_{D_{n\varepsilon}} |\tilde{T}_{jk}(\Delta, \tau) - \tilde{T}_{jk}(P(\Delta), P(\tau))| \geq Mn^{1/4}\right] \to 0 \tag{A.18}$$

as $n \to \infty$.

Hence $\sup_{D_{n\varepsilon}} |\tilde{T}_{jk}(\Delta, \tau)| = O_P(n^{1/2} \log n)$ follows by combining (A.17) and (A.18). □

5 Further Notes

Regarding the term "chain" and "chaining", there exists some slightly different, yet spiritually similar definition, see Section 3 of Pollard (1990). On the other hand, the chaining arguments summarized in this chapter may be viewed as a shortcut to prove uniform results without diving into the empirical processes framework. The similar concept embodied under the empirical processes framework is *stochastic equicontinuity*, see Andrews (1994b) and Andrews (1994a) for details.

Bibliography

Donald W.K. Andrews. Non-strong mixing autoregressive processes. *Journal of Applied Probability*, **21**(4):930–934, 1984.

Donald W.K. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72, 1994a.

Donald W.K. Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, **4**:2247–2294, 1994b.

F. J. Anscombe. Examination of residuals. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 1–36, Berkeley, Calif., 1961. University of California Press.

I. Barrodale and F.D.K. Roberts. Solution of an overdetermined system of equations in the l_1 norm. *Communications of the ACM*, **17**(6):319–320, 1974.

Douglas M. Bates and Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. John Wiley & Sons, Inc., 1988.

Peter J. Bickel. One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, **70**(350):428–434, 1975.

Peter J. Bickel. Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *The Annals of Statistics*, **6**(2):266–291, 1978.

Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, Inc., third edition, 1995.

Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, Inc., second edition, 1999.

Jelena Bradic, Jianqing Fan, and Weiwei Wang. Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**(3):325–349, 2011.

Raymond J. Carroll and David Ruppert. A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. *Journal of the American Statistical Association*, **77**(380):878–882, 1982.

Raymond J. Carroll and David. Ruppert. *Transformation and Weighting in Regression*. Monographs on Statistics and Applied Probability No. 30. CRC Press, 1988.

Colin Chen. A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, **16**(1):136–164, 2007.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**(4):987–1007, 1982.

Jianqing Fan and Qiwei Yao. *Nonlinear Time Series: Parametric and Nonparametric Approaches*. New York: Springer-Verlag, 2003.

Jianqing Fan, Lingzhou Xue, and Hui Zou. Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics*, **42**(3):819–849, 2014.

Jianqing Fan, Lingzhou Xue Xue, and Hui Zou. Multitask quantile regression under the transnormal model. *Journal of the American Statistical Association*, **111** (516):1726–1735, 2016.

Thomas S. Ferguson. *A Course in Large Sample Theory*, volume **54**. Springer Science+Business Media, B.V., 1996.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, **3**(1):100–118, 1975.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

Christoph Gutenbrunner and Jana Jurečková. Regression rank scores and regression quantiles. *The Annals of Statistics*, **20**(1):305–330, 1992.

Christoph Gutenbrunner, Jana Jurečková, Roger Koenker, and Stephen Portnoy. Tests of linear hypotheses based on regression rank scores. *Journal of Nonparametric Statistics*, **2**(4):307–331, 1993.

Xuming He and Qi-Man Shao. A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, **24**(6):2608–2630, 1996.

Nils Lid Hjort and David Pollard. Asymptotics for minimisers of convex processes. *arXiv preprint arXiv:1107.3806*, 2011.

David R. Hunter and Kenneth Lange. Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, **9**(1):60–77, 2000.

Jana Jurečková. Asymptotic relations of M-estimates and R-estimates in linear regression model. *The Annals of Statistics*, **5**(3):464–472, 1977.

Jana Jurečková. Robust estimators of location and their second-order asymptotic relations. In *A Celebration of Statistics*, pages 377–392. Springer, 1985.

Jana Jurečková and Bohumír Procházka. Regression quantiles and trimmed least squares estimator in nonlinear regression model. *Journal of Nonparametric Statistics*, **3**:201–222, 1994.

Bo Kai, Runze Li, and Hui Zou. Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1):49–69, 2010.

Jeankyung Kim and David Pollard. Cube root asymptotics. *The Annals of Statistics*, **18**(1):191–219, 1990.

Keith Knight. Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, **26**(2):755–770, 1998.

Keith Knight and Wenjiang Fu. Asymptotics for Lasso-type estimators. *The Annals of Statistics*, **28**(5):1356–1378, 2000.

Roger Koenker. A note on L-estimates for linear models. *Statistics & Probability Letters*, **2**(6):323–325, 1984.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs No. 38. Cambridge University Press, 2005.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica*, **46**(1): 33–50, 1978.

Roger Koenker and Gilbert Bassett Jr. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, **46**(1):43–61, 1982.

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

Roger Koenker and Stephen Portnoy. L-estimation for linear models. *Journal of the American Statistical Association*, **82**(399):851–857, 1987.

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

Roger Koenker and Quanshui Zhao. L-estimation for linear heteroscedastic models. *Journal of Nonparametric Statistics*, **3**(3-4):223–235, 1994.

Roger Koenker and Quanshui Zhao. Conditional quantile estimation and inference for ARCH models. *Econometric Theory*, **12**(5):793–813, 1996.

Hongcheng Liu, Tao Yao, Runze Li, et al. Global solutions to folded concave penalized nonconvex learning. *The Annals of Statistics*, **44**(2):629–659, 2016.

Sanjay Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal of Optimization*, **2**(4):575–601, 1992.

Walter Oberhofer. The consistency of nonlinear regression minimizing the L_1-norm. *The Annals of Statistics*, **10**(1):316–319, 1982.

Walter Oberhofer and Harry Haupt. Asymptotic theory for nonlinear quantile regression under weak dependence. *Econometric Theory*, **32**(3):686–713, 2016.

David Pollard. Asymptotics via empirical processes. *Statistical Science*, **4**(4): 341–354, 1989.

David Pollard. Empirical processes: Theory and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–86. JSTOR, 1990.

David Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**(2):186–199, 1991.

G. Polyá. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, **8**:171–181, 1920.

Stephen Portnoy and Roger Koenker. Adaptive L-estimation for linear models. *The Annals of Statistics*, **17**(1):362–381, 1989.

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

Walter Rudin. *Principles of Mathematical Analysis*. New York: McGraw-Hill, Inc., third edition, 1976.

David Ruppert and Raymond J. Carroll. Trimmed least squares estimation in the linear model. *Journal of the American Statistical Association*, **75**(372):828–838, 1980.

Bernard W. Silverman. *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability No. 26. CRC Press, 1986.

William F. Stout. *Almost Sure Convergence*. New York: Academic Press, 1974.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, fourth edition, 2003.

Jinde Wang. Asymptotic normality of L_1-estimators in nonlinear regression. *Journal of Multivariate Analysis*, **54**(2):227–238, 1995.

Alan H. Welsh. Robust estimation of smooth regression and spread functions and their derivatives. *Statistica Sinica*, **6**(2):347–366, 1996.

Alan H. Welsh, Raymond J. Carroll, and David Ruppert. Fitting heteroscedastic regression models. *Journal of the American Statistical Association*, **89**(425): 100–116, 1994.

Margaret H. Wright. Interior methods for constrained optimization. *Acta Numerica*, **1**:341–407, 1992.

Margaret H. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, **42**(1):39–56, 2005.

Wei Biao Wu. Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(40): 14150–14154, 2005.

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

Quanshui Zhao. Asymptotically efficient median regression in the presence of heteroskedasticity of unknown form. *Econometric Theory*, **17**(4):765–784, 2001.

Zhibiao Zhao and Zhijie Xiao. Efficient regressions via optimally combining quantile information. *Econometric Theory*, **30**(6):1272–1314, 2014.

Kenneth Q. Zhou and Stephen L. Portnoy. Statistical inference on heteroscedastic models based on regression quantiles. *Journal of Nonparametric Statistics*, **9**(3): 239–260, 1998.

Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, **36**(3):1108–1126, 2008.

Vita

Zhanxiong Xu

Zhanxiong Xu was born in Quzhou, China, on October 12, 1986. After finishing high school in 2004, he entered the University of Science and Technology of China, major in mathematics. Since September 2007, he began to study statistics in the department of statistics and finance at the same university, from where he received his bachelor degree in 2009 and master degree in 2012, both in statistics. He started his Ph.D. at Penn State University since August 2012 and is expected to receive his Ph.D. in statistics in August 2017. During this period, his main research interest is the large-sample theory in quantile regression.