5-2013

Robust and Efficient Regression

Qi Zheng
Clemson University, qiz@clemson.edu

ROBUST AND EFFICIENT REGRESSION

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Statistics

by
Qi Zheng
May 2013

Accepted by:
Dr. Colin Gallagher, Committee Chair
Dr. Karunarathna B. Kulasekera, co-advisor
Dr. Chanseok Park
Dr. Xiaoqian Sun
Dr. Robert Taylor

Abstract

This dissertation aims to address two problems in regression analysis. One problem is the model selection and robust parameter estimation in high dimensional linear regressions. The other is concerning developing a robust and efficient estimator in nonparametric regressions.

In Chapter 1, we introduce the robust and efficient regression analysis, discuss those two interesting problems and our motivations, and present several exciting results.

We propose a novel robust penalized method for high dimensional linear regression in Chapter 2. Asymptotic properties are established and a data-driven procedure is developed to select adaptive penalties. We show it is the very first estimator to achieve desired oracle properties with certainty for high dimensional linear regression. Extensive simulations have been conducted and demonstrate the usefulness of the new technique.

A new local polynomial nonparametric regression is developed in Chapter 3. It minimizes a convex combination of several weighted loss functions simultaneously. The optimal weights are selected by a proposed procedure and adapt to the tails of the error distribution resulting in a procedure which is both robust and resistant. The asymptotic properties have been investigated. We show the resulting estimators are at least as efficient as those provided by existing procedures, but can be much more efficient for many distributions. Its excellent finite sample performance is presented through simulations under a variety of settings. A real data analysis exhibits the usefulness of the proposed methodology.

Dedication

I dedicate this work to my loving parents. It is their love and support that made this work a complete one.

Acknowledgments

I would like to express my gratitude to many individuals who helped me in many ways during this work.

First and my foremost, I am indebted to my co-advisors Drs. C. Gallagher and K.B.Kulasekera for their guidance and support. Their mathematical insights inspired me and helped me make this a success.

I would also like to thank Drs. C. Park, X. Sun, and R. Taylor for their insightful suggestions.

Last but not least, I would like to thank the Department of Mathematical Sciences for providing me financial support during my Ph.D studies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression analysis is a fundamental statistical tool for the investigation of the relationships between a dependent variable and one or more independent variables. Of interest is to estimate a function called the regression function, which describes the expected behaviors or characteristics of the dependent variables given the independent variables. A great number of techniques for carrying out regression analysis have been developed and applied to various scientific domains for centuries, and some of them even have become standard procedures for statistical data analysis, such as linear regression and ordinary least squares. However, there are still a lot of open problems in this area, and thus abundant active research is still being conducted. For instance, the least square technique is inefficient if errors are heavy-tailed. More importantly, it could fail to provide reliable estimates if the data set is contaminated by some outliers. Therefore, in recent decades, a great amount of literature has been devoted to develop robust and efficient regression analysis.

Linear regression can be used to detect which among the independent variables are related to the dependent variable, and to measure the effect of independent variables upon the dependent variables. This usage is well known as “model selection and parameter estimation”. Although this has been well studied for the case where the number of independent variables is finite, it is still a big challenge in the situation where the number of independent variables is much larger than the number of observations. One of our aims is to propose a robust estimation which can not only select the true model with probability converging to 1, but also estimate the parameters accurately.

Another attractive problem is so called nonparametric regression, in which we do not assume that the underlying regression function is of any particular parametric form. Instead, it is simply required to be a smooth function. Model misspecification can be avoided by using nonparametric regression. Moreover, nonparametric regression has larger flexility to explain data for which parametric regression models are incapable of capturing the characteristics of the conditional expectation function. We attempt to develop a kernel-based local polynomial estimator to achieve both robustness and efficiency under different errors.

1.1 High dimensional linear regression

Over the last decade, many new applications arising in biometrics, image processing, econometrics, and many other fields, have created an increasing demand for high-dimensional data analysis to deal with a large number of variables. In many situations, e.g. climate studies, business intelligence, DNA microarry data, pattern identification problems in social network, the dimensionality, p, exceeds the number of observations, n. The dimensionality p can possibly be of order $O(\exp(n^\alpha))$, for some $0 < \alpha < 1$; and may increase with the growth of sample size. But in most situations, only s of them are significant, where $s = o(n)$. Classical subset selection procedures are incapable to provide reliable estimates of regression model parameters due to extremely heavy computational intensity and instability suffering from 'the curse of dimensionality'. Some of my doctoral research concentrated on developing model selection and parameters estimation procedures for high-dimensional linear models (HDLMS).

Some literature (e.g. Fan and Peng (2004), Huang et.al.(2008a), Huang et.al.(2008b)) has considered modifying penalized estimators (e.g. LASSO, SCAD) for HDLMS, and established the desired oracle property. Here the oracle property of a method means it can correctly select the significant variables with probability converging to 1, and provide asymptotically normal estimators which in the limit perform as well as those fitted with all insignificant variables excluded in advance. However, those techniques are based on the squared loss and hence require stringent moment conditions on the unobservable error sequence, $\{\epsilon_i\}$, which may not be satisfied by many econometric data sets. To achieve robustness, Candes and Tao (2007) and Belloni and Chernozhukov (2011) attempted to integrate the quantile regression into the penalty framework. However, both estimators only achieve the $\sqrt{n/(s\log(p))}$ consistency rate, which is slower than the oracle rate $\sqrt{\frac{n}{s}}$ from He and Shao (2000).

Motivated by the results of Wang et al. (2007), Zou and Yuan (2008), and others in the literature, we explore robust quantile techniques with fully adaptive penalties to produce an adaptively penalized quantile regression, which can simultaneously select the model and estimate regression regression coefficients. We relax strong moment conditions imposed on $\{\epsilon_i\}$, propose a new data-driven procedure to select penalties which completely adapt to magnitudes of regression coefficients, and demonstrate that the adaptively penalized quantile regression is not only robust but also possesses the desired oracle property. This is an advancementd from the existing quantile regression methods for HDLMS. To our best knowledge, this is the first quantile regression estimator to enjoy the oracle property with certainty for high dimensional linear models.

1.2 Adaptively weighted kernel regression

A general nonparametric regression model is defined as $y_i = m(x_i) + \sigma(x_i)\epsilon_i$. Since the regression function $m(\cdot)$, does not take any predetermined form, this model has a much larger flexibility to explain data, and hence can be applied in many areas including: group testing, environmental science, social science, etc.

Numerous procedures have been proposed to estimate $m(\cdot)$. Most of them are constructed as local polynomial approximation with two types of loss functions: the least squares (LS) and the quantile check functions. Compared to local LS, the local quantile regression is robust and more efficient for heavy-tailed errors, but may be inefficient for short-tailed errors. Kai et al. (2010) proposed a local composite of quantile regression (CQR), and showed that the local CQR can significantly improve the estimation efficiency of its local LS counterpart for common non-normal errors. However, the loss in efficiency compared to the local LS still exists in many scenarios. In addition to that, it is unclear how many quantiles should be used in the local CQR. Even for a huge data set, increasing the number of quantiles does not necessarily improve the efficiency of estimates (See Kai et al. (2010)).

We develop a new local polynomial methodology for nonparametric regression based on optimizing a linear combination of several loss functions. We propose a simple data-driven procedure to select weights for the convex combination and establish the asymptotic properties of the resulting estimator. We show that the proposed method combines the strengths of LS and quantile regression to gain both efficiency and robustness. It performs at least as well as the local LS or a local CQR for any error distribution, and can improve the estimation efficiency upon LS and CQR for many distributions. The proposed method is also quite robust and works well even if the error distribution does not have a finite variance. Moreover, we demonstrate we can increase the number of quantiles for a larger data set to achieve more efficiency. Our simulation experiment and real data analysis also exhibited the proposed estimator compete favorably with other methods.

Chapter 2

Adaptive Penalized Quantile Regression for High Dimensional Data

In this chapter, we propose a new adaptive L_1 penalized quantile regression estimator for high-dimensional sparse regression models with heterogeneous error sequences. We show that under weaker conditions compared with alternative procedures, the adaptive L_1 quantile regression selects the true underlying model with probability converging to one, and the unique estimates of nonzero coefficients it provides have the same asymptotic normal distribution as the quantile estimator which uses only the covariates with non-zero impact on the response. Thus, the adaptive L_1 quantile regression enjoys oracle properties. We propose a completely data driven choice of the penalty level λ_n, which ensures good performance of the adaptive L_1 quantile regression. Extensive Monte Carlo simulation studies have been conducted to demonstrate the finite sample performance of the proposed method.

2.1 Introduction

Consider the high dimensional sparse regression model

$$y_i = \beta_0^* + \beta_1^* z_{i1} + \cdots + \beta_p^* z_{ip} + \epsilon_i, \qquad i = 1, \cdots, n \tag{2.1}$$

where $\{y_i\}$'s are random variables, $\{\mathbf{z}_i\}$'s are $p \times 1$ independent random covariate vectors, and $\{\epsilon_i\}$ are independent random error terms with $P(\epsilon_i \leq 0|\mathbf{z}_i) = \tau$ for some quantile index τ. We allow the dimension of the covariate vector to be very large, possibly of order $O(\exp(n^\alpha))$, for some constant $0 < \alpha < 1$; but the regression parameter β^* is sparse in the sense that only $s << p$ of its components are non-zero. Of interest is to identify the nonzero regressors and estimate their regression coefficients as well. Such models have attracted great attention due to the demand for data analysis created by many new applications arising in genetics, signal processing, machine learning, climate change point detection and other fields with high-dimensional data sets available.

Various methods have been developed to identify the unknown model and estimate the corresponding coefficients simultaneously for the high dimensional sparse model (see Fan and Peng 2004; Huang *et.al.* 2008a; Huang *et.al.* 2008b), which mostly focus on the penalized least squares regression. Although some of them enjoy desirable oracle properties (Fan and Li 2001), they generally require stringent moment assumptions (Cramér condition) on the unobservable homoscedastic random errors, $\{\epsilon_i\}$. Therefore, they are not robust and may no be applicable in practice. Compared with least squares, another important statistical method, quantile regression (Koenker and Bassett 1978), is robust and allows relaxation of moment conditions on the heterogeneous error sequence. The advantage of quantile regression goes beyond that: it can provide a more complete model of the relationship between predictors and response variables. (e.g. Koenker 2005), it owns excellent computational properties. (e.g. Portnoy and Koenker 1997), and it has widespread applications, (e.g. Yu *et.al.* 2003, Chernozhukov 2005). Belloni and Chernozhukov (2011) integrate general quantile regression into an L_1 penalty framework for the high-dimensional sparse model. Another interesting estimator, the Dantzig selector, considered by Candes and Tao (2007), can be considered as a penalized median regression. However, both of these estimators achieve the $\sqrt{n/(s\log(p))}$ consistency rate, which is slower than the oracle rate $\sqrt{n/s}$ from He and Shao(2000). Wang *et al.* (2012) proposed a quantile regression with SCAD penalty. Since the objective function is not convex, the solutions are not unique. To our best knowledge, the desirable oracle properties have not been achieved by any penalized quantile regression for the high-dimensional sparse model.

we attempt to overcome the limitations of the existing quantile regression techniques by combining quantile regression with a fully adaptive L_1 penalty function to produce adaptive L_1 quantile regression, which can simultaneously select the model and provide a robust estimator possessing oracle properties. Exploiting the ideas of Wang *et.al.* (2007) and Zou and Yuan (2008), we use the consistent estimator from Belloni and Chernozhukov (2011) to determine adaptive weights. Since we are using quantile loss functions, we do not require the *Cramér* condition on the error sequence. Our contributions are summarized as follows:

- First, we show that under mild conditions, the adaptive L_1 quantile regression will select the correct model with probability converging to 1, and for any quantile index in a compact set in (0,1), the unique adaptive L_1 quantile regression estimates are consistent with the oracle rate $\sqrt{n/s}$. This is an advancement from the existing quantile regrssion methods for the high-dimensional sparse model.
- Second, any linear combination of the estimates is asymptotically normal with the same asymptotic variance as that of the oracle estimator.
- Third, in deriving the aforementioned oracle properties, we propose a new data-driven procedure to select the penalty level and show that it satisfies the requirements to achieve the oracle rate.

The rest of the chapter is organized as follows. In Section 2, we define the adaptive L_1 quantile regression procedure. In section 3, we study the asymptotic properties of the L_1 quantile regression estimator and discuss the choice of penalty level λ_n. Numerical studies are presented in Section 4. We give concluding remarks in Section 5, and relegate the technical proofs to the Appendix A.

2.2 The adaptive L_1 quantile regression

We start with introducing notations. We implicitly index all parameter values by the sample size n, but we omit the index whenever this does not cause confusion. We use the notation $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We denote the l_2-norm by $\| \cdot \|$, and the l_0-"norm" (the number of nonzero components)by $\|\cdot\|_0$. Given a vector $\delta \in \mathbb{R}^{p+1}$, and a set of indices $T \subset \{0, 1, \cdots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. And q^* is the τth quantile of ϵ.

In order to define the adaptive L_1 quantile regression, let us briefly review quantile regression and L_1 penalized quantile regression. Let $\mathbf{x}_i = (1, \mathbf{z}_i^T)^T$. Quantile regression estimator of β^* can be obtained by solving

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta), \tag{2.2}$$

where $\rho_\tau(t) = \tau 1(t > 0)t - (1 - \tau)1(t \leq 0)t$ is the check function.

Without loss of generality, we assume that the first $s + 1$ elements of β^* are nonzero, and the rest are zero. For simplicity, write $\beta^* = (\beta_a^{*T}, \beta_b^{*T})^T$, where β_a^* is a $(s+1) \times 1$ vector and β_b^* is a $(p - s) \times 1$ vector of zeroes. Similarly, we decompose $\mathbf{x}_i$ as $(\mathbf{x}_{ia}^T, \mathbf{x}_{ib}^T)^T$.

Belloni and Chernozhukov (2011) proposed a penalized L_1 quantile regression estimator $\tilde{\beta}$, which minimizes:

$$\tilde{Q}_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + \frac{\lambda_n \sqrt{\tau(1-\tau)}}{n} \sum_{i=1}^{p} \hat{\sigma}_j |\beta_j| \tag{2.3}$$

where $\hat{\sigma}_j = \sum_{i=1}^{n} x_{ij}^2 / n, j = 1, \cdots, p$ and obeys $P(\max_{1 \leq j \leq p} |\hat{\sigma}_j - 1| \leq 1/2) \geq 1 - \alpha \to 1$. Here λ_n is the penalty parameter. Ideally, a penalty function should be adaptive in the sense that it penalizes insignificant variables enough to force estimates of their regression coefficients to be zero, but does not overpenalize significant variables, so that the correct model can be identified and hence oracle properties can be attained. However, it can be seen that the penalty for each variable in (3) is of the same order, λ_n / n, and hence not quite adaptive. A similar issue appears in the estimator proposed by Candes and Tao (2007).

To improve the quantile regression for the high-dimensional sparse model, we attempt to assign fully adaptive weights to different variables and propose the adaptive L_1 quantile regression estimator $\hat{\beta}$, which is a minimizer of the objective function

$$Q_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + \lambda_n \sum_{j=1}^{p} \omega_j |\beta_j| \tag{2.4}$$

where $\omega \in \mathbb{R}^p$ is weights vector chosen to be $|\tilde{\beta}|^{-1} \wedge \sqrt{n}$, for any $\sqrt{n/(s \log(n \vee p))}$-consistent estimator $\tilde{\beta}$ of β^*. For example, we can take the estimator from Belloni and Chernozhukov (2011) as $\tilde{\beta}$, which under conditions A1-A3 given below will converge at a sufficiently fast rate. The formulation (2.4) includes the LAD-Lasso proposed by Wang *et. al.* (2007) as a special case that the dimensionality p is fixed.

2.3 Asymptotic Properties

In this section, we state primitive regularity conditions and then establish the asymptotic properties of the adaptive L_1 quantile regression estimator.

2.3.1 Regularity Conditions

The following regularity conditions are assumed throughout the rest of this chapter.

A1 (Sampling and smoothness). For any value $\mathbf{x}$ in the support of $\mathbf{x}_i$, the conditional density $f_{\epsilon|\mathbf{z}}(\epsilon|\mathbf{z})$ is continuously differentiable at each $y \in \mathbb{R}$, and $f_{\epsilon|\mathbf{x}}(\epsilon|\mathbf{x})$ and $\frac{\partial}{\partial\epsilon} f_{\epsilon|\mathbf{x}}(\epsilon|\mathbf{x})$ are bounded in absolute value by constants $\bar{f}$ and $\bar{f}'$ uniformly in $\epsilon \in \mathbb{R}$ and $\mathbf{x}$ in the support of $\mathbf{x}_i$. Moreover, the conditional density of $\epsilon|\mathbf{x}$ evaluated at the conditional quantile $q^*_{\mathbf{x}}$ is bounded away from 0 uniformly for any $\mathbf{x}$ in the support of $\mathbf{x}_i$. That is, there exists a constant $\underline{f}$, such that $f_{\epsilon|\mathbf{x}}(q^*_{\mathbf{x}}|\mathbf{x}) > \underline{f} > 0$ uniformly.

A2 (Restricted identifiability and nonlinearity). Define $T = \{0, 1, \cdots, s\}$, and $\bar{T}(\delta, m) \subset \{0, 1, \cdots, p\} \backslash T$ as the support of the m largest in absolute value components of the vector. For some constants $m \geq 0$ and $c \geq 0$, the matrix $E[\mathbf{x}_i \mathbf{x}_i']$ satisfies

$$\kappa_m^2 := \inf_{\delta \in A_{\mho}, \delta \neq 0} \frac{\delta' E[\mathbf{x}_i \mathbf{x}_i']\delta}{\|\delta_{T \bigcup \bar{T}(\delta,m)}\|^2} > 0$$

where $A := \{\delta \in \mathbb{R}^{p+1} : \|\delta_{T^c}\| \leq c_0\|\delta_T\|, \|\delta_{T^c}\|_0 \leq n\}$ and $\kappa_0^2 \leq C_f$ for some constant C_f. Moreover,

$$q := \frac{3}{8} \frac{\underline{f}^{3/2}}{\bar{f}'} \inf_{\delta \in A, \delta \neq 0} \frac{E[|\mathbf{x}_i^T \delta|^2]^{3/2}}{E[|\mathbf{x}_i^T \delta|^3]} > 0$$

A3 (Growth rate of covariates) The growth rate of significant variables and all variables allowed is assumed to satisfy $s^3(\log(n \vee p))^{2+\gamma}/n \to 0$, for some $\gamma > 0$.

A4 (Moments of covariate) Covariates satisfy the Cram$\acute{e}r$ condition $E[|z_{ij}|^k] \leq 0.5 C_m M^{k-2} k!$ for some constantd C_m, M, all $k \geq 2$ and all $j = 1, \cdots, p$

A5 (Well separated regression coefficients) We assume that there exists a $b_0 > 0$, such that for all $j \leq s$, $|\beta_j^*| > b_0$. We note b_0 could still be unknown to us.

Conditions A1-A5 are commonly assumed in the literature (see e.g. Fan and Peng 2004; Huang *et.al.* 2008a; Huang *et.al.* 2008b, Belloni and Chernozhukov 2011). Condition A1 is slightly different from Condition D.1 in Belloni and Chernozhukov (2011). The assumption D.1 in Belloni and Chernozhukov (2011), requiring the conditional density at the conditional quantile is uniformly bounded away from 0, can be replaced by a more general condition. In fact, we only need that the conditional density is nonvanishing. Condition A2 requires that there exists a constant C_f, such that $\kappa_0^2 \leq C_f$. This along with the fact that κ_m^2 is nonincreasing in m, immediately entails that the smallest eigenvalue of the covariance matrix $\Sigma_s := E[\mathbf{x}_{ia}\mathbf{x}_{ia}']$ is finite and bounded away from 0.

Condition A3 seems to be a strong assumption at first glance, because it limits the size of significant variables to be less than $n^{1/3}$, rather than $n^{2/3}$ as shown in Portnoy (1984). However, this assumption is in accord with Welsh (1989), in which the author showed that if the score function is discontinuous, the growth rate for covariates, $p^3(\log(n))^{2+\gamma}/n \to 0$ is sufficient to obtain the consistency and asymptotic normality under the full model. Since we deal with the high-dimensional sparse model, the growth rate would be expected to obey $s^3(\log(n \vee p))^{2+\gamma}/n \to 0$. Condition A4 is important for us to apply Bernstein's inequality, and hence to establish the sparsity property of the adaptive L_1 quantile estimator. In addition, A5 also implies $\sum_{i=1}^n E\|\mathbf{x}_{ia}\|^2 \sim O(ns)$, which is essential for establishing the oracle consistency property. Condition A5 is also required in Huang *et. al.* (2008b). It assumes that the nonzero coefficients are uniformly bounded away from 0; in other words, the parameter values of the true model are well separated from zero. This assumption can be relaxed to that $\min_{j \leq s} |\beta_j^*|$ goes to 0 at a suitable rate, at the cost of more complicated technical proofs.

2.3.2 Oracle Properties

We show that the adaptive L_1 quantile regression estimator enjoys oracle properties.

Theorem 2.3.1 *231 Suppose that assumptions A1-A5 are satisfied. Furthermore, if λ_n satisfies $\lambda_n s/\sqrt{n} \to 0$ and $\lambda_n/(\sqrt{s}\log(n \vee p)) \to \infty$, then the adaptive L_1 quantile regression estimator $\hat{\beta}$ must satisfy the following three properties:*

1. *Variable selection consistency:*

$$P(\hat{\beta}_b = 0) \geq 1 - 6\exp\{-\frac{\log(n \vee p)}{4}\}$$

2. *Estimation consistency:*

$$\|\hat{\beta} - \beta^*\| = O_p(\sqrt{\frac{s}{n}})$$

3. *Asymptotic Normality: Let $u_s^2 = \alpha^T \Sigma_s \alpha$ for any vector $\alpha \in \mathbb{R}^s$ satisfying $\|\alpha\| < \infty$. Then*

$$n^{1/2} u_s^{-1} \alpha^T (\hat{\beta}_a - \beta_a^*) \xrightarrow{D} N(0, \frac{\tau(1-\tau)}{f^2(q^*)})$$

Remark 2.3.1 *$\tilde{\beta}$ must be at least $\sqrt{n/(s\log(n \vee p))}$-consistent. If $\tilde{\beta}$ is a consistent estimator of β^* with some faster rate, that is, there is a sequence of a_n such that $a_n\|\tilde{\beta} - \beta^*\| \sim O_p(1)$ and $\sqrt{n/(s\log(n \vee p))} \sim o(a_n)$, the oracle properties can still be achieved if $\lambda_n s/\sqrt{n} \to 0$ and $\lambda_n a_n / \sqrt{n \log(n \vee p)} \to \infty$.*

Remark 2.3.2 *The asymptotic normality of any linear combination $u_s^{-1}\alpha(\hat{\beta}_a - \beta_a^*)$ is a substitute for the traditional asymptotic normality. Convergence of the finite-dimensional distributions ensures convergence in sequence space. In practice, hypothesis tests and confidence intervals would be constructed using linear combinations.*

2.3.3 The choice of λ_n

The regularization parameter, λ_n, plays a crucial role for the adaptive L_1 quantile estimator. It controls the overall magnitude of the adaptive weights and should be chosen so that insignificant variables' regression coefficient estimates shrink to zero, while significant variables are not overpenalized.

Procedures, which are commonly used to select λ_n, such as k-fold cross-validation, generalized cross-validation (Tibshirani 1996; Fan and Li 2001), and so on, can be applied to choose λ_n with some appropriate modification. However, using them may have several drawbacks. First, p, the number of variables in the full model, is increasing as the sample size grows. This factor results in an unpleasant issue in that the number of potential models goes to infinity very quickly, which makes computation much too expensive. Second, their statistical properties are not clearly understood for (ultra)high-dimensional regression. For example, there is no guarantee that K-fold cross-validation would provide a choice of λ_n with a proper rate. Third, their statistical properties are still uncharted under the heavy-tailed errors, where quantile regressions are often applied.

Wang and Leng (2007) developed a BIC criterion to select the tuning parameter λ_n for least square approximation (LSA) procedure, and its theoretical model selection consistency property has been demonstrated in Wang *et.al.* (2007) for fixed dimensionality and in Wang *et. al.* (2009) for high-dimensional regression. However, two limitations make such a BIC criterion less favorable in this ultra-high dimensional problem. The first limitation is that one of the requirements in Wang *et. al.* (2009) is $p < n$, which may not be satisfied in the ultra-high dimensional problem. The other limitation is that there is no efficient path-finding algorithm for quantile regression. Thus, we need to search all possible subsets to find the minimum BIC. This could potentially exhaust our computation. One might be able to use the LSA to approximate the quantile regression, and then implement least angle regression slicing (LARS) algorithm to find a solution path in an easier manner, as pointed out in Wang and Leng (2007). However, this would require obtaining a reliable estimate of the inverse of the covariance matrix (see Wang and Leng 2007), which is a difficult problem in the ultra-high dimensional case. Instead we consider an alternative method for selecting λ_n.

According to Theorem 2.3.1, a proper λ_n must satisfy two conditions: $\lambda_n s/\sqrt{n} \to 0$ and $\lambda_n/(\sqrt{s}\log(n \vee p)) \to \infty$. We can see that $O(\sqrt{s}\log(n \vee p)(\log n)^{\gamma/2})$ is a suitable choice of λ_n under the condition A5. However, the obstacle is that we do not know the true dimension s. Hence, a natural problem is can we find a good estimate of s, or at least get a quantity of order $O(s)$? Belloni and Chernozhukov (2011) show that their estimator $\|\tilde{\beta}_\tau\|_0 \sim O_p(s)$. If the parameter values of the minimal true model are well separated from zero as condition A7 assumes, then $\|\tilde{\beta}\|_0 \sim O_p(s)$. Since $\tilde{\beta}$ is consistent, $\|\tilde{\beta}_\tau\|_0$ is of order s with a large probability. Therefore, we can use $\tilde{\beta}_\tau$ not only to adjust weights for each regression coefficient, but also to get a quantity used to construct a good choice of λ_n. In practice, we choose $\lambda_n = 0.25\sqrt{\|\tilde{\beta}\|_0}\log(n \vee p)(\log n)^{0.1/2}$ and it works well in our simulation studies.

2.4 Numerical analysis

To evaluate the finite sample performance of the proposed estimator, we conducted Monte Carlo simulations. We compare the performance of the oracle quantile estimator, the L_1 penalized, post L_1 penalized quantile estimators (Belloni and Chernozhukov 2011), and the proposed adaptive estimator. The post L_1 penalized quantile estimator is obtained by applying ordinary quantile regression to the model selected by the L_1 penalized quantile regression.

We adopt the simulation settings used in Belloni and Chernozhukov (2011). Consider the regression model 1:

$$y_i = \mathbf{x}_i^T \beta + \epsilon,$$

where $\beta = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \cdots, 0)^T$ and $\mathbf{x}_i = (1, \mathbf{z}_i^T)^T$ consists of an intercept and covariates $\mathbf{z}_i \sim N(0, \Sigma)$, and the errors ϵ are independently and identically distributed $\epsilon \sim N(0, \sigma^2)$. The dimension p of covariate is 500, and the true dimensiona s is 6. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. We apply the median regression and choose $\lambda_n = 0.25\sqrt{\|\tilde{\beta}\|_0} \log(n \vee p)(\log n)^{0.1/2}$. We consider three levels of noise $\sigma = 1, 0.5$ and 0.1. 100 training data sets are generated, each consisting of 100 observations. We assess model selection by calculating N1: the number of covariates selected by each estimator $\hat{\beta}$, N2: the correct number of covariates selected by each estimaor, and the percentage of underfitted, correctly fitted, and overfitted. We evaluate the estimation accuracy by computing the norm of the bias and the empirical risk $[E[\mathbf{x}_i^T(\hat{\beta} - \beta)]^2]^{1/2}$. The results are summarized in the Table 1. We can see that although the proposed estimator may

Table 2.1: Simulation results for model 1

	Average N1	Average N2	Underfitted	Correctly fitted	Overfitted	Bias	Empirical risk
				σ=1			
Oracle	6	6	0	1	0	0.03	0.31
L_1	3.21	3.21	1	0	0	0.77	1.09
Post L_1	3.21	3.21	1	0	0	0.30	0.59
Adaptive	4.04	4.04	1	0	0	0.22	0.43
				σ=0.5			
Oracle	6	6	0	1	0	0.02	0.15
L_1	4.41	4.40	0.98	0.02	0	0.49	0.69
Post L_1	4.41	4.40	0.98	0.02	0	0.21	0.31
Adaptive	5.05	5.04	0.73	0.26	0.01	0.16	0.25
				σ=0.1			
Oracle	6	6	0	1	0	0	0.03
L_1	5.93	5.93	0.07	0.93	0	0.15	0.20
Post L_1	5.93	5.93	0.07	0.93	0	0.01	0.04
Adaptive	6.05	5.99	0.01	0.95	0.04	0.01	0.03

still fail to select some significant variables when σ is large due to the ultra-high dimensionality, it significantly improves the performance of quantile regression in both model selection and estimation, compared with the L_1 penalized , post L_1 penalized quantile estimators. Notice that the proposed estimator does not necessarily treat 0 as an absorbing status even when the initial L_1 penalized estimator provides a zero estimate. This is the advantage of using $\omega_j = |\tilde{\beta}|^{-1} \wedge \sqrt{n}$, which provides another opportunity to select the significant regressors, and hence provides better results.

Following Wang *et. al* (2012), we consider model 2, which is a heterogenous version model 1.

$$y_i = \mathbf{x}_i^T \beta + \Phi(x_{i2})\epsilon$$

where $\Phi(\cdot)$ is the standard normal cumulative density function. We consider $\sigma = 1$ and $\sigma = 0.5$. And the results are presented in Table 2. Similar conclusions can be drawn from Table 2. All three methods are able to work for regression models with heterogenous errors. However, as observed from Table 2, the adaptive penalized quantile regression drastically outperformed the L_1 penalized , post L_1 penalized quantile estimators in both model selection and estimation.

Table 2.2: Simulation results for model 2

	Average N1	Average N2	Underfitted	Correctly fitted	Overfitted	Bias	Empirical risk
				σ=1			
Oracle	6	6	0	1	0	0.02	0.11
L_1	4.36	4.35	0.96	0.04	0	0.53	0.74
Post L_1	4.36	4.35	0.96	0.04	0	0.20	0.31
Adaptive	5.08	5.06	0.75	0.25	0	0.14	0.22
				σ=0.5			
Oracle	6	6	0	1	0	0	0.05
L_1	5.35	5.34	0.62	0.38	0	0.33	0.46
Post L_1	5.35	5.34	0.62	0.38	0	0.12	0.15
Adaptive	5.88	5.85	0.15	0.85	0	0.05	0.08

2.5 Conclusion

In this chapter, the adaptive L_1 quantile regression is introduced and studied for high-dimensional sparse models. It is shown that such an adaptive robust estimator enjoys the oracle properties. In the case of quantile regression we can relax the moment conditions and the constant variance assumption on the error sequence from those used to prove oracle properties of penalized least squares loss methods for high-dimensional data. Our simulation results demonstrate that the proposed estimator owns satisfactory finite sample performances. Although the oracle properties of a single quantile index τ are presented here, the result can be easily extended to a finite composite quantile regression [Zou and Yuan (2008)].

Chapter 3

Adaptively Weighted Kernel Regression

In this chapter, we develop a new kernel-based local polynomial methodology for nonparametric regression based on optimizing a linear combination of several loss functions. Optimal weights for least squares and quantile loss functions can be chosen to provide maximum efficiency and these optimal weights can be estimated from data. The resulting estimators are at least as efficient as those provided by existing procedures, but can be much more efficient for many distributions. The data based weights adapt to the tails of the error distribution resulting in a procedure which is both robust and resistant. Furthermore, the assumption of homogeneous error variance is not required. The method is used to model the change of global temperature anomolies over the last 100 years.

3.1 Introduction

Consider a general nonparametric model

$$Y = m(X) + \sigma(X)\epsilon, \tag{3.1}$$

where Y is the response variable, X is the explanatory variable, $m(\cdot)$ is a smooth nonparametric regression function, $\sigma(X)$ is a smooth function and ϵ is random error with a p.d.f. symmetric of 0. Without loss of generality, we assume $E[\epsilon_i^2] = 1$ if $E[\epsilon_i^2] < \infty$.

Various methods have been developed to fit this type of model [see e.g. Watson (1964); Wahba (1990); Fan and Gijbels (1996)]. It is fairly common to fit the model weighted least squares (LS) with local polynomial approximation [e.g. Fan and Gijbels (1992)]. However, least squares fitting can be very sensitive to heavy-tailed errors and severe outliers. Consequently least squares based local polynomial regression could fail to produce reliable estimates in some cases. In contrast, quantile regression [Koenker and Bassett (1978)] is robust against outliers, can provide a more complete model of the relationship between predictors and response variables [e.g. Koenker (2005)], owns excellent computational properties. [e.g. Portnoy and Koenker (1997)], and has widespread applications [e.g. Yu *et.al.* (2003); Chernozhukov (2005)]. As a result, a substantial amount of literature has been devoted to study local polynomial quantile regression [see e.g. Fan *et al.* (1994); Welsh (1996); Yu and Jones (1998)]. For some error structures, local polynomial quantile regression can be more efficient than local least squares polynomial regression. For example, if the error follows a Laplacian distribution, the local median polynomial regression has been demonstrated to be the most efficient [Fan *et al.* (1994); Welsh (1996); Yu and Jones (1998)]. In other cases, local quantile regression could be arbitrarily less efficient than local LS polynomial regression (e.g., for normal data), resulting from the fact that loss functions of quantile regressions penalize residuals of small magnitude too strongly.

To improve the performance of quantile regression, Koenker and Portnoy (1987) considered L-estimation for linear models. An L-estimator is a weighted average of quantile estimators, which can achieve high efficiency for non-normal data. Bickel (1973) and Koenker (1984) demonstrate that as the number of quantiles used increases, the optimally weighted L-estimator is as efficient as the maximum likelihood estimator. However, it is difficult to find the optimal weights [see Portnoy and Koenker (1989)], and the computational cost increases dramatically with the number of quantiles. Instead, Zou and Yuan (2008) introduced composite quantile regression (CQR), which equally weights quantile loss functions. Kai *et al.* (2010) adapted composite quantiles to the local polynomial framework. They showed that the local polynomial CQR can significantly improve the estimation efficiency of its local LS counterpart for common non-normal errors. However, the loss in efficiency compared to the LS polynomial regression still exists in many scenarios. In addition to that, it is unclear how many quantiles should be used in the local polynomial CQR. Even for a huge data set, increasing the number of quantiles does not necessarily improve the efficiency of estimates [See Kai *et al.* (2010)]. Since neither L-estimators nor CQR estimators incorporate the least squares loss, these estimators can require a large number of quantiles to achieve efficiency especially when the magnitude of errors is small. Bradic *et al.* (2011) attempted to minimize composite loss functions simultaneously which results in a robust and efficient estimator for high dimensional linear regression.

In this chapter, we attempt to embed the usage of a convex combination of loss functions into nonparametric kernel regression to obtain a robust estimator with significant improvement in efficiency. Different from the combination of LS and least absolute deviation (LAD) for errors with finite errors or the combination of quantile losses for symmetric errors discussed in Bradic *et al.* (2011), we combine the least squares loss with quantile loss functions for symmetric errors and picking weights to optimize asymptotic efficiency results in a method which inherits all strengths from least squares and quantile regression methods. We establish the asymptotic properties of the resulting estimator and show that it performs at least as well as the local LS polynomial estimator or a local polynomial CQR for any error distribution, and can improve the estimation efficiency for many distributions. Furthermore it achieves the same efficiency as the optimally weighted L-estimator and can achieve higher efficiency than the equally weighted CQR of Kai *et al.* (2010). We propose a simple data-driven procedure to select weights for the convex combination and show that the aforementioned asymptotic properties can be achieved by this adaptively weighted local polynomial regression estimator. The adaptively weighted local polynomial estimator is quite robust and works well even if the error distribution does not have a finite variance.

The rest of this chapter is organized as follows. In Section 2, we define the adaptive weighted local polynomial regression estimator. In Section 3, we study theoretical properties of the proposed estimator. Section 4 presents our simulation studies and the analysis of a real data set. We give concluding remarks in Section 5, and relegate technical proofs to the Appendix B.

3.2 The adaptively weighted local polynomial regression

We start by setting up notations. Let $\rho_\tau(t) = \tau 1(t > 0)t - (1 - \tau)1(t \leq 0)t$ be the check function with quantile index τ. Let $\tau_k = \frac{k}{q+1}, k = 1, \cdots, q$ be equally spaced quantile indices between 0 and 1. We denote the τ_kth quantile by $q_{\tau_k}, k = 1, \cdots, q$. In particular, let $\tau_0 = 0$, $q_{\tau_0} = 0$ and $\rho_{\tau_0}(t) = t^2$. We use $F(\cdot)$ and $f(\cdot)$ to denote the cumulative distribution function and probability density function of ϵ_i, respectively. $g_X(\cdot)$ is the marginal density of X. $K(\cdot)$ is a classical kernel function. We also use the following notations $\tau_{k,k'} = \tau_{k\wedge k'} - \tau_k\tau_{k'}$ where $k \wedge k' = \min\{k, k'\}$, and $\tau_{0,k} = E[\epsilon_i 1(\epsilon_i \leq q_{\tau_k})]$, for $k = 1, \cdots, q$.

In order to define the adaptively weighted local polynomial regression, let us briefly review the local LS polynomial regression, the local quantile polynomial regression and the local polynomial CQR.

Let (x_i, y_i) be n independently and identically distribution observations. Of interest is to estimate the value of $m(X)$ at x_0. Suppose $m(X)$ is smooth enough to be approximated by a pth order polynomial in a neighborhood of x_0, that is, $m(x) \approx \sum_{j=0}^{p} \frac{1}{j!} m^{(j)}(x_0)(x - x_0)^j$. The local LS polynomial regression estimator of $(m(x_0), m^{(1)}(x_0), \cdots, m^{(p)}(x_0))$ is defined as the minimizer of the following objective function

$$\min_{a_0, a_1, \cdots, a_p} \sum_{i=1}^{n} \rho_{\tau_0} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K\left(\frac{x_i - x_0}{h}\right) \tag{3.2}$$

where h is a smoothing parameter. Fan and Gijbels (1992) demonstrated that the local LS polynomial regression owns several desirable properties: it adapts to a wide variety of design densities, significantly reduces bias at boundary points, and attains high minimax efficiency.

However, the local LS polynomial regression suffers from outliers and heavy-tailed errors. Motivated by its robustness and other good features, several authors [Fan *et al.* (1994); Welsh (1996); Yu and Jones (1998)] advocated local quantile polynomial regression

$$\min_{a_0, a_1, \cdots, a_p} \sum_{i=1}^{n} \rho_{\tau} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K\left(\frac{x_i - x_0}{h}\right) \tag{3.3}$$

for some quantile index τ. Although the local quantile polynomial regression can be applied for more general error structures, it can be arbitrarily inefficient compared to the local LS polynomial regression. To improve the efficiency of the local quantile polynomial regression while maintaining the robustness, Kai *et al.* (2010) proposed the local polynomial CQR as follows

$$\min_{a_{01}, \cdots, a_{0q}, a_1, \cdots, a_p} \sum_{i=1}^{n} \sum_{k=1}^{q} \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=1}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) K\left(\frac{x_i - x_0}{h}\right) \tag{3.4}$$

They showed that the local polynomial CQR can significantly improve efficiency compared to the local quantile polynomial regression. However, the loss of efficiency of the local polynomial CQR still exists for some commonly seen distributions.

We consider combining CQR and LS to produce an efficient and robust regression estimator. Let $\theta = (a_{01}, \cdots, a_{0q}, a_0, a_1, \cdots, a_p)$ and denote the solution to the objective function

$$\min_{\theta} \sum_{i=1}^{n} \left[\sum_{k=0}^{q} \beta_k \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) \right] K\left(\frac{x_i - x_0}{h} \right) \tag{3.5}$$

by $\hat{\theta}_\beta = (\hat{a}_{01}, \cdots, \hat{a}_{0q}, \hat{a}_0, \hat{a}_1, \cdots, \hat{a}_p)$. Here $a_{00} = 0$ and $\beta_0, \cdots, \beta_q$ are well-chosen non-negative weights which adapt to the error structures. The details about how to choose those weights are presented in Section 3.2. The adaptively weighted local polynomial regression estimator is defined as

$$\begin{aligned} \hat{m}_\beta(x_0) &= \frac{\frac{1}{2\sigma}\sum_{k=1}^{q} \beta_k f(q_{\tau_k}) \hat{a}_{0k}}{\beta_0 + \frac{1}{2\sigma}\sum_{k=1}^{q} \beta_k f(q_{\tau_k})} + \hat{a}_0 \\ \hat{m}_\beta^{(j)}(x_0) &= \hat{a}_j, \qquad j = 1, \cdots, p \end{aligned} \tag{3.6}$$

For identification purposes, we set $\sigma\beta_0 + \sum_{k=1}^{q} \beta_k f(q_{\tau_k})/2 = 1$. The estimator becomes

$$\hat{m}_\beta(x_0) = \frac{1}{2} \sum_{k=1}^{q} \beta_k f(q_{\tau_k}) \hat{a}_{0k} + \hat{a}_0. \tag{3.7}$$

This formulation actually provides an advantage. In the following section, it can be seen that the variances of $\hat{m}(x_0)$ and $\hat{m}^{(j)}(x_0)$, $1 \le j \le p$ are of similar forms. Consequently, minimizing them separately still produces the same optimal weights vector β.

3.3 Asymptotic Properties

In this section, we state primitive regularity conditions and then establish the asymptotic properties of the adaptively weighted local polynomial regression estimator.

3.3.1 Regularity Conditions

To study the asymptotic properties of the adaptively weighted local polynomial regression estimator, the following regularity conditions are assumed throughout the rest of this paper.

(A) $m(\cdot)$ has continuous $(p+2)$th derivative in the neighbourhood of x_0

(B) f is symmetric about 0 and belongs to the domain of attraction of some stable distribution S.

(C) f is continuous and positive.

(D) $g_X(\cdot)$ is positive and differentiable in the neighborhood of x_0.

(E) $K(\cdot)$ is a symmetric kernel function with a compact support $[-M, M]$, and satisfies

(a) $|K(u)| < C_k$

(b) $\int_{-M}^{M} K(u)du = 1$

(c) $\int_{-M}^{M} u^j K(u)du = \mu_j$, $\int_{-M}^{M} u^j K^2(u)du = \nu_j$, $j \geq 0$. In particular, $\mu_j = \nu_j = 0$ for odd j.

Regularity conditions A, C, D, E are commonly assumed in the literature [see e.g. Fan (1992), Yu and Jones (1998), Kai *et al.* (2010)]. As is pointed out elsewhere, the assumption that $K(\cdot)$ has a compact support can be relaxed at the cost of more complicated technical proofs. In simulation studies, we exhibit the excellent performance of the proposed estimator with the classical normal kernel. The assumption that f is symmetric about 0 is required in Kai *et al.* (2010). Although weighted CQR for asymmetric errors was considered recently in Sun *et al.* (2013), we still maintain the symmetric assumption to simplify the complicated proof that the impact of the LS part is negligible when $E[\epsilon_i^2]$ does not exist. However, our estimator can be generalized to asymmetric distributions following Sun *et al.* (2013).

Up front, under the assumption that $E[\epsilon_i^2] < \infty$, we establish the asymptotic properties of the adaptively weighted local polynomial regression estimator to demonstrate that it is more efficient and hence is favorable to other polynomial regression estimators. Next we consider $E[\epsilon_i^2] = \infty$ and show that the impact of the LS part in the adaptively weighted local polynomial regression estimator is asymptotically negligible, while the efficiency is preserved under this infinite variance scenario. Therefore, the proposed estimator is a robust and efficient alternative to other polynomial regression estimators.

To avoid the complicated statements, we first illustrate our ideas via the i.i.d error models,

$$Y = m(X) + \sigma\epsilon$$

and then generalize it to heterogeneous error models.

3.3.2 Asymptotic Properties when $E[\epsilon_i]^2$ exists

Throughout this subsection we assume $E[\epsilon_i]^2 < \infty$. To state the asymptotic properties of the adaptively weighted local polynomial regression estimator, we need to introduce the following notations:

Define

$$S(\beta) = \begin{pmatrix} S_{11}(\beta) & S_{12}(\beta) \\ S_{21}(\beta) & S_{22}(\beta) \end{pmatrix}$$

where $S_{11}(\beta)$ is a $q \times q$ diagonal matrix with diagonal elements $\beta_k f(q_{\tau_k})/(2\sigma)$, for $k = 1, \cdots, q$, S_{22} is a $(p+1) \times (p+1)$ matrix with (j, j')-entry $\mu_{(j+j'-2)}$, for $j, j' = 1, \cdots, p+1$, and $S_{12}(\beta) = S_{21}(\beta)^T$ is a $q \times (p+1)$ matrix with (k, j)-entry $\beta_k f(q_{\tau_k})/(2\sigma)\mu_{j-1}$, for $k = 1, \cdots, q; j = 1, \cdots, p+1$.

Let

$$V_\beta = 4\beta_0^2\sigma^2 - 4\beta_0 \sum_{k=1}^{q} \beta_k \sigma \tau_{0,k} + \sum_{k,k'=1}^{q} \beta_k \beta_{k'} \tau_{k,k'},$$

and we define

$$\Sigma(\beta) = \begin{pmatrix} \Sigma_{11}(\beta) & \Sigma_{12}(\beta) \\ \Sigma_{21}(\beta) & \Sigma_{22}(\beta) \end{pmatrix}$$

where $\Sigma_{11}(\beta)$ is a $q \times q$ matrix with (k, k')-entry $\beta_k \beta_{k'} \nu_0 \tau_{k,k'}$, for $k, k' = 1, \cdots, q$, $\Sigma_{22}(\beta)$ is a $(p+1) \times (p+1)$ matrix with (j, j')th element $V_\beta \nu_{(j+j'-2)}$, for $j, j' = 1, \cdots, p+1$, and $\Sigma_{12}(\beta) = \Sigma_{21}^T(\beta)$ is a $q \times (p+1)$ matrix with (k, j)-entry $(-2\beta_0\beta_k\sigma\tau_{0,k} + \beta_k \sum_{k'=1}^{q} \beta_{k'}\tau_{k,k'})\nu_{(j-1)}$, for $k = 1, \cdots, q; j = 1, \cdots, (p+1)$.

Let $r_{i,p} = m(x_i) - \sum_{j=0}^{p} m^{(j)}(x_0)(x_i - x_0)^j/j!$ be the residual of the Taylor expansion of $m(x_i)$ at x_0, and $\xi_{\beta,i} = -2\beta_0(\sigma\epsilon_i + r_{i,p}) + \sum_{k=1}^{q} \beta_k[1(\epsilon_i \leq (\sigma q_{\tau_k} - r_{i,p})/\sigma) - \tau_k]$. We define $\mathbf{W}_{\beta,n} = (w_{\beta,01}, \cdots, w_{\beta,0q}, w_{\beta,0}, w_{\beta,1}, \cdots, w_{\beta,p})^T$, where

$$w_{\beta,0k} = \beta_k \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)[1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k], \qquad k = 1, \cdots, q$$

$$w_{\beta,j} = \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\left(\frac{x_i - x_0}{h_n}\right)^j \xi_{\beta,i}, \qquad j = 0, \cdots, p.$$

Then the asymptotic properties of the adaptively weighted local polynomial regression estimator can be established in the following theorem:

Theorem 3.3.1 *Suppose assumptions A,B,C,D, and E are satisfied. Furthermore, we assume*

$E[\epsilon_i^2] < \infty$. If $h_n \to 0$ and $nh_n \to \infty$, then for any nonnegative weights vector $\beta = (\beta_0, \cdots, \beta_q)^T$,

$$\sqrt{nh_n} S(\beta) A_{h_n} (\hat{\theta}_\beta - \theta^*) + \frac{1}{2g(x_0)} E[\mathbf{W}_{\beta,n}] \xrightarrow{L} N\left(\mathbf{0}, \frac{1}{4g(x_0)} \Sigma(\beta)\right)$$

where $\theta^* = (q_{\tau_1}, \cdots, q_{\tau_q}, m(x_0), m^{(1)}(x_0), \cdots, m^{(p)}(x_0))^T$ is a vector of true parameters and A_{h_n} is a $(q+1+p) \times (q+1+p)$ diagonal matrix with diagonal elements $(1, \cdots, 1, h_n^0/0!, \cdots, h_n^p/p!)$.

As special cases, two corollaries follow immediately.

Corollary 3.3.1 *Under the same assumptions as Theorem 3.1, if $p = 1$, we have*

$$\sqrt{nh_n}\left[\hat{m}_\beta(x_0) - m(x_0) - \frac{m^{(2)}(x_0)}{2}\mu_2 h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_0 \sigma^2}{4g(x_0)} V_\beta\right)$$

and the mean squared error of $\hat{m}(x_0)$ is

$$MSE(\hat{m}_\beta(x_0)) = \left(\frac{m^{(2)}(x_0)}{2}\mu_2\right)^2 h_n^4 + \frac{\nu_0\sigma^2}{4g(x_0)}\frac{V_\beta}{nh_n} + o_p\left(h_n^4 + \frac{1}{nh_n}\right) \tag{3.8}$$

Corollary 3.3.2 *Under the same assumptions as Theorem 3.1, if $p = 1$, then*

$$\sqrt{nh_n}\left[\hat{m}_\beta^{(1)}(x_0) - m^{(1)}(x_0) - \left(\frac{m^3(x_0)}{6} + \frac{m^{(2)}(x_0)g^{(1)}(x_0)}{2g(x_0)}\right)\frac{\mu_4}{\mu_2}h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_2\sigma^2}{4g(x_0)h_n^2\mu_2^2}V_\beta\right)$$

and

$$MSE(\hat{m}_\beta^{(1)}(x_0)) = \left(\frac{m^{(3)}(x_0)}{6} + \frac{m^{(2)}(x_0)g^{(1)}(x_0)}{2g(x_0)}\right)^2 \frac{\mu_4^2}{\mu_2^2}h_n^4 + \frac{\nu_2\sigma^2}{4g(x_0)\mu_2^2}\frac{V_\beta}{nh_n^3} + o_p\left(h_n^4 + \frac{1}{nh_n^3}\right) \tag{3.9}$$

if $p = 2$, then

$$\sqrt{nh_n}\left[\hat{m}_\beta^{(1)}(x_0) - m^{(1)}(x_0) - \frac{m^{(3)}(x_0)\mu_4}{6\mu_2}h_n^2\right] \xrightarrow{L} N\left(0, \frac{\nu_2\sigma^2}{4g(x_0)h_n^2\mu_2^2}V_\beta\right)$$

and

$$MSE(\hat{m}_\beta^{(1)}(x_0)) = \left(\frac{m^{(3)}(x_0)}{6}\right)^2 \frac{\mu_4^2}{\mu_2^2}h_n^4 + \frac{\nu_2\sigma^2}{4g(x_0)\mu_2^2}\frac{V_\beta}{nh_n^3} + o_p\left(h_n^4 + \frac{1}{nh_n^3}\right) \tag{3.10}$$

Corollary 3.1 indicates that the bias of $\hat{m}(x_0)$ relies on β through C_β, which is assumed to be 1 for model identification. Thus, the optimal weights vector β in the sense of minimizing the MSE of

$\hat{m}(x_0)$ can be chosen by minimizing V_β:

$$\beta_{opt} = \underset{\beta \geq 0, \alpha^T\beta = 1}{\operatorname{argmin}} \; V_\beta \tag{3.11}$$

where $\alpha = (1, f(q_{\tau_k})/(2\sigma), \cdots, f(q_{\tau_q})/(2\sigma))^T$.

Remark 3.3.1 *When $q \to \infty$, if we set $\beta_0 = 0$ and minimize V_β with respect to β, the resulting covariance matrix is the same as that of the nonparametric polynomial L-estimation with optimal weights. Therefore, the proposed method is also as efficient as the maximum likelihood, when $q \to \infty$.*

Noting that the mean squared error of $\hat{m}^{(1)}_\beta(x_0)$ only depends on β by V_β as well, then β_{opt} is also optimal for estimating $m^{(1)}(x_0)$. For most practical interests, estimating $m(x_0)$ is the main focus. However, it can be shown that β_{opt} is optimal for estimating all $m^{(j)}(x_0)$ for $j \leq p$.

Since equation (3.11) is a constrained quadratic minimization problem, the closed form solution for the optimal weights can be difficult to obtain. However, in some cases, optimal weights can be explicitly found. We provide several examples to show the availability of the optimal weights.

Example 3.3.1 *Let $q = 1, \tau_1 = \frac{1}{2}$ and $p \geq 1$. In other words, we consider the combination of LS and LAD, then the optimal weights are*

$$\beta_{0,opt} = \begin{cases} 0 & \text{if } \frac{1-2f(0)E[|\epsilon_i|]}{4f(0)^2-4f(0)E[|\epsilon_i|]+1} < 0, \\ 1 & \text{if } \frac{1-2f(0)E[|\epsilon_i|]}{4f(0)^2-4f(0)E[|\epsilon_i|]+1} > 1, \\ \frac{1}{\sigma}\frac{1-2f(0)E[|\epsilon_i|]}{4f(0)^2-4f(0)E[|\epsilon_i|]+1} & \text{otherwise.} \end{cases} \quad \text{and} \quad \beta_{1,opt} = \frac{2(1-\beta_{0,opt})}{f(0)}$$

Example 3.3.2 *If $\epsilon_i \sim N(0, \lambda^2)$, then $\beta_{0,opt} = 1$ and $\beta_{k,opt} = 0$, for all $1 \leq k \leq q$.*

Example 3.3.3 *If $\epsilon_i \sim$ Laplace$(0, \lambda)$ and q is odd, then $\beta_{l,opt} = 2f(0)$ for $l = (q+1)/2$, and $\beta_{0,opt} = \beta_{k,opt} = 0$, for all $k \neq l$.*

Both the local polynomial LS and the local polynomial CQR estimators are special cases of weighted local polynomial regression. Regardless of the error distribution, the efficiency achieved by choosing the theorectically optimal weights can be no less than that gained by either of those methods. Moreover, the proposed estimator can be more efficient than the local LS polynomial regression estimator and the local polynomial CQR for some distributions, as Example 3.1 and 3.2 demonstrate.

Theorem 2 in Kai *et al.* (2010) indicates that as the number of quantiles increases, the asymptotic relative efficiency between CQR and LS converges to 1. For the proposed weighted estimator, increasing the number of quantiles does not impact the efficiency in this way, but can in fact improve the asymptotic efficiency of the estimator.

Although V_β is typically unobservable, we can replace them with consistent estimators. Let $\tilde{\zeta}_i$ be residuals of a $\sqrt{nh_n}$-consistent preliminary estimation, $i = 1, \cdots, n$. For example, we could use the residuals from local polynomial median regression, or residuals from local polynomial CQR, or the residuals from local LS polynomial regression, if the error terms $\{\epsilon_i\}$ have a finite second moment. We use the notation $\tilde{T}$ to denote the empirical estimate of T by using $\tilde{\zeta}_i$, for some statistic T. Then $\tilde{V}_\beta = 4\beta_0^2\tilde{\sigma}^2 - 4\beta_0 \sum_{k=1}^q \beta_k \tilde{\sigma}\tilde{\tau}_{0,k} + \sum_{k,k'=1}^q \beta_k \beta_{k'} \tau_{k,k'}$, and we can obtain the practically optimal weights vector

$$\hat{\beta} = \underset{\beta \geq 0, \tilde{\alpha}^T\beta = 1}{\operatorname{argmax}} \tilde{V}_\beta \tag{3.12}$$

where $\tilde{\alpha} = (\tilde{\sigma}, \frac{1}{2}\tilde{f}(\tilde{q}_{\tau_1}), \cdots, \frac{1}{2}\tilde{f}(\tilde{q}_{\tau_q}))^T$. The consistency of $\hat{\beta}$ can easily be verified. We have the following corollary:

Corollary 3.3.3 *Under the same assumptions as Theorem 3.1,*

$$\sqrt{nh_n} S(\beta_{opt}) A_{h_n}(\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)} E[\mathbf{W}_{\beta_{opt},n}] \xrightarrow{L} N\left(\mathbf{0}, \frac{1}{4g(x_0)}\Sigma(\beta_{opt})\right)$$

The proposed estimator, using $\hat{\beta}$ obtained from (3.12) does not suffer from any loss of efficiency.

Notice that in $\tilde{V}_\beta$, q_{τ_k}, $\tau_{0,k}$, and $f_{q_{\tau_k}}$ need to be estimated. Therefore, the number of quantiles depends on the sample size. When the sample size is small we recommend using only a few quantiles to avoid the impact by introducing too many parameters. On the other hand, if the sample size is large, more quantiles should be adopted. In practice, the cross-validation, AIC, BIC type estimators can be applied to choose the number of quantiles.

3.3.3 Asymptotic properties when $E[\epsilon_i^2]$ does not exist

Since least squares may not provide reliable estimates when heavy-tailed errors or outliers appear, in this case one might use a weight of zero ($\beta_0 = 0$) for the LS part of objective function (3.5). In practice we do not know if the variance is finite and we propose picking weights using a numerical solution to the constrained quadratic minimization problem (3.12). So $\hat{\beta}_0$ is not necessarily 0. We would like to find out if the proposed estimator can still be applied.

The following theorem answers the aforementioned question.

Theorem 3.3.2 *Suppose assumptions A,B,C,D, and E are satisfied. Furthermore, we assume $E[\epsilon_i^2]$ does not exist. If $h_n \to 0$ and $nh_n \to \infty$, then*

$$\sqrt{nh_n} S(\beta_{opt}) A_{h_n} (\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)} E[\mathbf{W}_{\beta_{opt},n}] \xrightarrow{L} N\left(\mathbf{0}, \frac{1}{4g(x_0)} \Sigma(\beta_{opt})\right)$$

This theorem indicates that $\hat{\beta}_0$ converges to 0 fast enough to make the instability caused by LS negligible. Theorem 3.2 coupled with Theorem 3.1 imply that the adaptively weighted local polynomial regression can be applied universally. It is a very safe alternative to other estimators. In addition, because $\hat{\beta}$ is chosen to adapt to different error distributions, the resulting local polynomial regression estimator is asymptotically more efficient than the local polynomial CQR. Those features make the proposed estimator very appealing in practice.

3.3.4 Heterogeneous errors

In the foregoing sections, we exhibit the desirable theoretical properties of the adaptively weighted local polynomial regression estimator under the homogeneous model. An interesting question naturally arises: "Can this method be applied to regression problems of which the error sequences are heterogeneous?"

The essential idea of the proposed procedure is to use the residuals from some preliminary method to select approximately optimal weights for the different loss functions. If the error sequences are homogeneous, then all residuals can be employed to establish the error structure. On the other hand, if the errors are heterogeneous, residuals of observations with covariate values closer to x_0, the point of interest, should contribute more to the local error structure estimation. Hence we can use weighted residuals to estimate the local error structure at x_0, where weights are assigned by a kernel function. Take the uniform kernel as an illustration, as $n \to \infty$, the number of observations falling into $[x_0 - h_n, x_0 + h_n]$ is of order nh_n. Therefore, the asymptotic efficiency should not be impacted by doing local error structure estimation. In practice, the pilot fit also provides initial bandwidths so that we can manipulate observations falling into the smoothing window to approximate the error structure locally.

Theorem 3.3.3 *Under model (1), suppose assumptions A,B,C,D, and E are satisfied. Furthermore, if $h_n \to 0$ and $nh_n \to \infty$, then*

$$\sqrt{nh_n} S(\beta_{opt}) A_{h_n}(\hat{\theta}_{\hat{\beta}} - \theta^*) + \frac{1}{2g(x_0)} E[\mathbf{W}_{\beta_{opt},n}] \xrightarrow{L} N(\mathbf{0}, \frac{1}{4g(x_0)} \Sigma(\beta_{opt}))$$

where $\hat{\beta}$ is obtained from (12) by using weighted local residuals.

Notice for heterogeneous cases, V_β varies at different x. However, it can be shown that for the theoretical optimal weights β_{opt} at different x, $(\beta_{1,opt}, \cdots, \beta_{q,opt})$'s and $\sigma(x)\beta_{0,opt}$'s are are constant. Thus, the weights are smooth functions of x, and so are the resulting estimators. Those simple relationships in fact facilitate our computation to obtain the optimal weights $\hat{\beta}$ at different x. We can randomly select some points in the support X, and calculate $\hat{\beta}$ at each point. Then we can first average obtained $(\hat{\beta}_{1,opt}, \cdots, \hat{\beta}_{q,opt})$ as practically global optimal weights for quantiles. Moreover, we can acquire a basis taking an average of $\hat{\beta}_0/\hat{\sigma}(x)$'s. For a given x, the product of the basis and a consistent estimate of $\sigma(x)$ yields optimal $\hat{\beta}_0$.

3.3.5 Bandwidth Selection

The performance of local polynomial regression estimators depends crucially on the smoothing parameter h. Obtaining a good bandwidth is very important for the success of the adaptively weighted local polynomial regression estimator. Given a weights vector β, the optimal bandwidth in the sense of minimizing MSE$(\hat{m}_\beta(x_0))$ is

$$h_{\beta,opt}(x_0) = \left[\frac{1}{(m^{(2)}(x_0))^2} \frac{\nu_0 \sigma^2(x_0)}{4g(x_0)\mu_2^2} V_\beta\right]^{\frac{1}{5}} n^{-\frac{1}{5}},$$

and the optimal bandwidth for the local linear regression estimator is

$$h_{LS}(x_0) = \left[\frac{1}{(m^{(2)}(x_0))^2} \frac{\nu_0 \sigma^2(x_0)}{g(x_0)\mu_2^2}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}.$$

It follows that

$$h_{\beta,opt}(x_0) = \left(\frac{V_\beta}{4}\right)^{1/5} h_{LS}(x_0) \tag{3.13}$$

As suggested in Kai *et al.* (2010), when $E[\epsilon_i^2]$ exists we can exploit this simple relationship to select the optimal bandwidth for the proposed estimator using existing bandwidth selectors for the local linear estimator. When $E[\epsilon_i^2]$ does not exist, we can similarly select the bandwidth via the relationship between the proposed estimator and the local LAD linear estimator. In both cases, we can infer $V_{\tilde{\beta}}$ from preliminary estimates.

3.4 Numerical Studies and Applications

In this section, we conduct a simulation study which evaluates the finite sample performance of the adaptively weighted local polynomial regression estimator. We then apply the proposed estimator to a real data set as a demonstration of its practical use.

3.4.1 Simulations

In our simulation studies, we adopt the settings used in Kai *et al.* (2010). We consider two simulation models.

1. $Y = \sin(2X) + 2\exp(-16X^2) + 0.5\epsilon$, where $X \sim N(0, 1)$

2. $Y = X\sin(2\pi X) + (1/5 + \cos(2\pi X)/10)\epsilon$, where $X \sim Unif(0, 1)$

In each model, we consider various distributions for ϵ: $N(0, 1)$ and Unif$(-1/2, 1/2)$ represent light-tailed errors; Laplace$(0, 1)$ represents moderate-tailed errors; a t_3-distribution represents heavy-tailed errors; a mixture of two normal distributions $0.95N(0, 1) + 0.05N(0, \sigma^2)$ with $\sigma = 3, 10$ represent errors with light and severe outliers, respectively; and Cauchy$(0, 1)$ represent distributions without finite second moments. They belong to the domain of attraction of some stable distribution, respectively. For each combination, we simulated 400 independent training data sets, each consisting of 200 observations.

Since heavy-tailed errors and contaminated data sets are taken into account in the studies, we use a local polynomial median regression as a safe preliminary fit to get a consistent estimator $\tilde{V}_\beta$ for V_β. The local polynomial median regression can be conveniently obtained via *quantreg* package in R.

We compare the proposed method with the classical local linear estimator and the local polynomial CQR via evaluating the integrated mean squared errors (IMSE), which is a summation of mean squared errors at L equally spaced grid points over the interval at which the regression function is estimated. For model 1, we estimate $m(x)$ over [-1.5,1.5] with $L = 200$ and for model 2,

Table 3.1: Simulation results for model 1

	RIMSE		x=0.75			RIMSE		x=0.75	
	mean	standard deviation	bias	standard deviation		mean	standard deviation	bias	standard deviation
$N(0,1)$					LS	-	-	0.0026	0.1851
CQR$_5$	0.9517	0.1007	0.0054	0.1957	AW$_5$	0.9831	0.0636	0.0042	0.1896
CQR$_9$	0.9720	0.0683	0.0028	0.1914	AW$_9$	0.9818	0.0513	0.0026	0.1886
CQR$_{19}$	0.9824	0.0489	0.0034	0.1897	AW$_{19}$	0.9861	0.0460	0.0027	0.1864
$Unif(-1/2,1/2)$					LS	-	-	0.0000	0.0706
CQR$_5$	0.9005	0.0620	0.0010	0.0755	AW$_5$	1.0335	0.1012	0.0010	0.0727
CQR$_9$	0.9565	0.0579	0.0000	0.0740	AW$_9$	1.1485	0.2035	0.0010	0.0756
CQR$_{19}$	0.9939	0.0625	0.0010	0.0731	AW$_{19}$	1.1905	0.2912	-0.0011	0.1149
$Laplace(0,1)$					LS	-	-	0.0009	0.2572
CQR$_5$	1.1315	0.2079	0.0009	0.2555	AW$_5$	1.2235	0.4206	0.0078	0.2738
CQR$_9$	1.0780	0.1417	0.0027	0.2570	AW$_9$	1.2077	0.3935	0.0093	0.2735
CQR$_{19}$	1.0390	0.0855	-0.0011	0.2559	AW$_{19}$	1.1856	0.3617	0.0013	0.2730
t_3					LS	-	-	0.0039	0.2725
CQR$_5$	1.4379	0.9505	-0.0039	0.2395	AW$_5$	1.4949	0.7788	-0.0039	0.2314
CQR$_9$	1.3137	0.6906	-0.0039	0.2666	AW$_9$	1.5286	0.8823	-0.0042	0.2321
CQR$_{19}$	1.1541	0.3561	-0.0042	0.2706	AW$_{19}$	1.5003	0.7874	-0.0044	0.2597
$0.95N(0,1)+0.05N(0,9)$					LS	-	-	-0.0017	0.2239
CQR$_5$	1.1006	0.2235	0.0003	0.2335	AW$_5$	1.0889	0.1742	0.0018	0.2334
CQR$_9$	1.0836	0.1618	0.0013	0.2286	AW$_9$	1.0815	0.1445	0.0014	0.2279
CQR$_{19}$	1.0499	0.1005	-0.0023	0.2303	AW$_{19}$	1.0778	0.1274	0.0013	0.2300
$0.95N(0,1)+0.05N(0,100)$					LS	-	-	0.0092	0.5413
CQR$_5$	2.5579	1.4642	0.0085	0.6297	AW$_5$	2.5791	1.4620	-0.0044	0.6204
CQR$_9$	1.9259	1.0100	-0.0033	0.6631	AW$_9$	2.2399	1.0012	-0.0035	0.6433
CQR$_{19}$	1.3574	0.4989	-0.0011	0.6721	AW$_{19}$	2.2495	1.0090	0.0014	0.6399
Cauchy$(0,1)$					LS	-	-	0.1093	5.1507
CQR$_5$	584.02	3458.39	-0.0526	0.2923	AW$_5$	974.05	5156.48	-0.0354	0.2000
CQR$_9$	314.52	2180.07	-0.0707	0.3941	AW$_9$	992.61	5416.02	-0.0330	0.2028
CQR$_{19}$	59.97	379.52	-0.1104	0.9891	AW$_{19}$	999.87	5390.16	0.0325	0.2036

we estimate $m(x)$ over [0,1] with $L = 200$. We consider $q = 5, 9, 19$ for the local polynomial CQR and the adaptively weighted local polynomial regression estimator. We use the normal kernel and select h_{LS} via a plug-in bandwidth selector, dpill, proposed by Ruppert *et al* (1995). For the proposed estimator we select the bandwidth using equation (15). The bandwidths for CQR are calculated using their relationship to LS. We summarize our simulation results using RIMSE: the ratio of the IMSE of the local linear estimator over the IMSE of other estimators. We also evaluate the performance of the proposed estimator at a specific point. The results are presented in Table 1 and Table 2, where CQR$_5$, CQR$_9$, CQR$_{19}$ denote the local polynomial CQR with $q = 5, 9, 19$ respectively, and likewise AW$_5$, AW$_9$, AW$_{19}$ denote the adaptively weighted local polynomial regression estimator with $q = 5, 9, 19$, respectively.

It appears that the proposed method adapts well to the different error distributions. In Table 1 we can see that the adaptively weighted local polynomial regression estimators outperform LS and CQR counterparts for most of the distributions considered. The proposed estimator shows

Table 3.2: Simulation results for model 2

	RIMSE		x=0.4			RIMSE		x=0.4	
	mean	standard deviation	bias	standard deviation		mean	standard deviation	bias	standard deviation
$N(0,1)$					LS	-	-	-0.0078	0.1851
CQR_5	0.9360	0.1681	-0.0055	0.1463	AW_5	0.9638	0.1148	-0.0077	0.1448
CQR_9	0.9640	0.1199	-0.0066	0.1436	AW_9	0.9769	0.0966	-0.0082	0.1444
CQR_{19}	0.9786	0.0869	-0.0065	0.1432	AW_{19}	0.9800	0.0909	-0.0070	0.1450
$Unif(-1/2,1/2)$					LS	-	-	-0.0019	0.0487
CQR_5	0.8240	0.0839	-0.0012	0.0532	AW_5	1.0088	0.1785	0.0012	0.0484
CQR_9	0.8971	0.0703	-0.0015	0.0519	AW_9	1.1705	0.3793	0.0010	0.0480
CQR_{19}	0.9493	0.0688	-0.0017	0.0516	AW_{19}	1.1912	0.4840	-0.0012	0.0481
$Laplace(0,1)$					LS	-	-	-0.0189	0.1841
CQR_5	1.2161	0.5549	-0.0157	0.1694	AW_5	1.2809	0.7334	0.0158	0.1808
CQR_9	1.1399	0.1417	-0.0159	0.1740	AW_9	1.2659	0.7234	0.0165	0.1827
CQR_{19}	1.0882	0.2635	0.0175	0.1719	AW_{19}	1.2145	0.6254	0.0179	0.1858
t_3					LS	-	-	0.0130	0.2069
CQR_5	1.4906	1.1322	-0.0097	0.1869	AW_5	1.6056	1.4276	-0.0089	0.1800
CQR_9	1.3753	1.1081	-0.0091	0.1860	AW_9	1.6445	1.4496	-0.0084	0.1786
CQR_{19}	1.2087	0.4249	-0.0095	0.1957	AW_{19}	1.6336	1.2998	-0.0093	0.1807
$0.95N(0,1)+0.05N(0,9)$					LS	-	-	-0.0165	0.1681
CQR_5	1.1394	0.4607	-0.0220	0.2270	AW_5	1.1260	0.3810	-0.0220	0.2271
CQR_9	1.1296	0.3960	-0.0241	0.2313	AW_9	1.1275	0.3664	-0.0235	0.2304
CQR_{19}	1.0769	0.2372	-0.0231	0.2309	AW_{19}	1.1196	0.3118	-0.0236	0.2291
$0.95N(0,1)+0.05N(0,100)$					LS	-	-	-0.0193	0.2872
CQR_5	3.4597	3.3701	-0.0219	0.2059	AW_5	3.4498	3.3139	-0.0219	0.2014
CQR_9	2.7592	2.2671	-0.0223	0.2360	AW_9	3.0444	2.3735	-0.0219	0.2111
CQR_{19}	1.6473	0.9416	-0.0189	0.2530	AW_{19}	3.1194	2.4475	-0.0205	0.2152
Cauchy$(0,1)$					LS	-	-	0.0999	2.3582
CQR_5	814.34	6309.01	-0.0523	0.0436	AW_5	1351.13	10207.48	-0.0366	0.0339
CQR_9	629.06	4816.19	-0.0636	0.0556	AW_9	1409.42	10972.55	-0.0359	0.0338
CQR_{19}	328.93	3012.81	-0.0803	0.0872	AW_{19}	1423.53	10795.12	-0.0365	0.0368

significant improvement over CQR in terms of RIMSE and also in terms of estimating the function at the point $x_0 = 0.75$. The first section of Table 1 shows little loss in efficiency relative to LS, when the error distribution is normal. Although the RIMSEs of the proposed estimators are slightly less than 1 for the normal distribution, the asymptotic ratio will get closer to 1 as sample sizes increase. The results in Table 2 indicate that the proposed adaptively weighted estimator still performs well in the presence of heteroskedasticity. In this case it appears that using too many quantiles can result in a loss of efficiency, but the proposed method tends to outperform CQR under the simulation set up.

3.4.2 A real data analysis

To illustrate its practical use, we apply the adaptively weighted local polynomial regression to the global temperature data set to study the modern temperature trend. The data set consisting of weighted global temperatures from 1911 to 2011 is available through U.S. national climatic data center (NCDC). Since the residuals from a local linear median regression have no significant autocorrelation according to the Ljung-Box test, the independence assumption of the error terms is not outlandish. We first use the local linear method, the local CQR and the proposed estimator

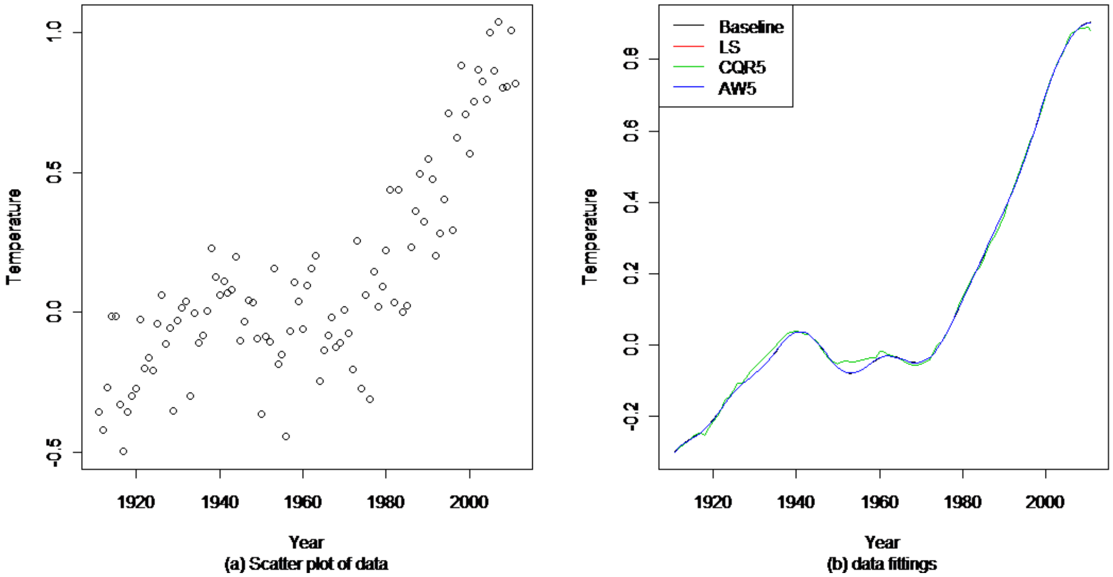
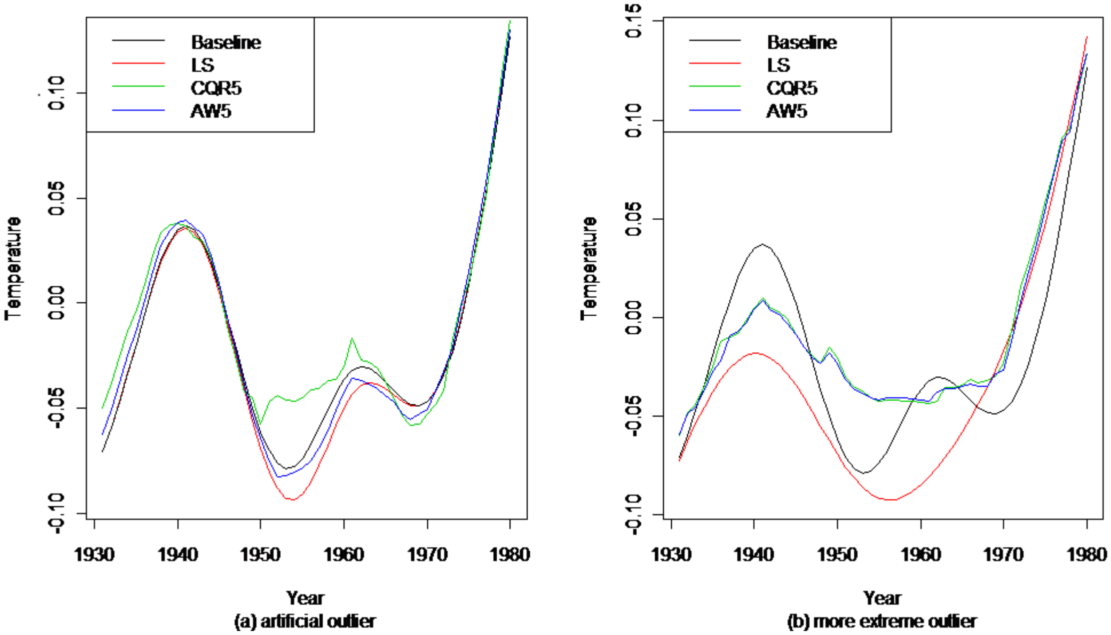


Figure 3.1: Scatter plot and three fittings

with 5 quantiles to fit the regression model. In the following analysis, we choose the local linear fit as the baseline fit. From Figure 3.1, we can see that all three procedures provide similar fits, and the local linear fit and the AW_5 fit are almost identical. The interesting part is the right end of the plots. It seems that the local linear fit and AW_5 support the claim that the global temperature is still increasing, while the CQR_5 indicates there is a change point around 2010.

Although there is no outlier in the data set, we artificially create one to examine robustness properties, we move the observation of 1956 from -0.4431 to -0.6647. In Figure 3.2 (a) we only depict the fits from 1931 to 1980 , since for other years the outlier has no effect on the estimates. Comparing with Figure 3.1 (b), we note that the local CQR_5 and AW_5 still maintain similar patterns, while the local linear estimator starts to deviate from the baseline. Moreover, we move the observation of 1956 from -0.4431 to -1.35 to simulate a severe outlier, and the fits are displayed in Figure 3.2 (b). Although all three procedures are affected by this severe outlier, the local linear changes drastically, whereas the local CQR_5 and AW_5 are much less affected.



(a) artificial outlier

(b) more extreme outlier

Figure 3.2: Outliers

3.5 Conclusion

In this chapter, we combine the strength of the least squares and quantile regression to propose the adaptively weighted local polynomial estimator for nonparametric regression. The novelty of the method is that it adapts to the distribution of the error terms in a regression model. We have explicitly described how data can be used to select weights as well as the bandwidth parameter. It appears that even when the weights are selected from the data, the estimators perform nearly as well as the optimal choice. For example, if the distribution is normal the method is nearly as efficient as LS, but the method still works well if the errors follow a t-distribution with 3 degrees of freedom. The estimators compete favorably with equally weighted composite quantile regression. The idea of weighting different objective functions and using asymptotic efficiency to select the optimal weights can be extended to other situations.

Appendices

Appendix A Proofs for Chapter 2

Appendix A-1: Consistency and sparsity

Define the score function of $\rho_\tau(\cdot)$ by $\varphi_\tau(\cdot)$, i.e. $\varphi_\tau(t) = \tau 1(t \geq 0) - (1-\tau)1(t < 0)$. $\hat{\beta}_\tau$ is the minimizer of the objective function

$$Q_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^T \beta) + \lambda_n \sum_{j=0}^{p} \omega_j |\beta_j|$$

Throughout $\tilde{\beta}$ is a $\sqrt{n/(s\log(n \vee p))}$-consistent estimator of β^*.

Lemma A.1 *Under assumptions A1-A5, if $\lambda_n/(\sqrt{s}\log(n \vee p)) \to \infty$ and $\omega_i = |\tilde{\beta}_{\tau j}|^{-1}$ for $1 \leq j \leq p$, then the adaptive L_1 quantile regression estimator $\hat{\beta}_\tau$ satisfies $\hat{\beta}_{\tau b} = 0$ with probability tending to 1.*

Proof: It can be seen that the objective function $Q_\tau(\beta)$ is piecewise linear. According to Theorem 1 in Bloomfield and Steiger (1983, page 7), the minimum of $Q_\tau(\beta)$ can be achieved at some breaking point $\breve{\beta}$, where $\rho_\tau(y_i - \mathbf{x}_i^T \breve{\beta}) = 0$ for some values of $i = 1, \cdots, n$.

Take the first derivative of $Q(\beta)$ at any differential point $\breve{\beta} \in R^{p+1}$ with respect to $\beta_j, j = s+1, \cdots, p$, and we obtain that

$$\frac{\partial Q(\beta)}{\partial \beta_j}|_{\breve{\beta}} = -\sum_{i=1}^{n} \varphi(y_i - \mathbf{x}_i^T \breve{\beta}) x_{ij} + \lambda_n \omega_j \mathrm{sgn}(\breve{\beta}_j) \qquad \text{(A.1)}$$

Let

$$D(\breve{\beta}, \beta^*) = \sum_{i=1}^{n} \varphi(y_i - \mathbf{x}_i^T \breve{\beta}) x_{ij} - \sum_{i=1}^{n} \varphi(y_i - \mathbf{x}_i^T \beta^*) x_{ij}$$

Note that,

$$\begin{aligned} D(\check{\beta}, \beta^*) = & \sum_{\epsilon_i \geq q^*_{\mathbf{x}_i}, \epsilon_i \geq q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta}-\beta^*)} [\tau x_{ij} - \tau x_{ij}] \\ & + \sum_{\epsilon_i \geq q^*_{\mathbf{x}_i}, \epsilon_i < q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta}-\beta^*)} [-(1-\tau) x_{ij} - \tau x_{ij}] \\ & + \sum_{\epsilon_i < q^*_{\mathbf{x}_i}, \epsilon_i \geq q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta}-\beta^*)} [\tau x_{ij} + (1-\tau) x_{ij}] \\ & + \sum_{\epsilon_i < q^*_{\mathbf{x}_i}, \epsilon_i < q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta}-\beta^*)} [-(1-\tau) x_{ij} + (1-\tau) x_{ij}]. \end{aligned}$$

where $q^*_{\mathbf{x}_i}$ is the conditional τth quantile of $\epsilon_i | \mathbf{x}_i$. For $K_1 = \{i : q^*_{\mathbf{x}_i} \leq \epsilon_i < q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta} - \beta^*)\}$ and $K_2 = \{i : q^*_{\mathbf{x}_i} > \epsilon_i \geq q^*_{\mathbf{x}_i} + \mathbf{x}_i^T(\check{\beta} - \beta^*)\}$,

$$D(\check{\beta}, \beta^*) = -\sum_{K_1} x_{ij} + \sum_{K_2} x_{ij}.$$

Hence,

$$\begin{aligned} & |\sum_{i=1}^n \varphi(y_i - \mathbf{x}_i^T \check{\beta}) x_{ij}| \\ = & |\sum_{i=1}^n \varphi(y_i - \mathbf{x}_i^T \beta^*) x_{ij} + D(\check{\beta}, \beta^*)| \\ \leq & |\sum_{i=1}^n \varphi(y_i - \mathbf{x}_i^T \beta^*) x_{ij}| + |\sum_{K_1} x_{ij}| + |\sum_{K_2} x_{ij}| \\ =: & I_1 + I_2 + I_3 \end{aligned}$$

Consider I_1 first. Let $\xi_i = \varphi(y_i - \mathbf{x}_i^T \beta^*) = \tau 1(\epsilon_i \geq q^*_{\mathbf{x}_i}) - (1-\tau) 1(\epsilon_i < q^*_{\mathbf{x}_i})$. Conditional on $\mathbf{x}_i$, it is easy to verify that $E[\xi_i x_{ij}] = 0$ and $\xi_i x_{ij}, i = 1, \cdots, n$ satisfy the Cramér condition. As a result, applying Bernstein's inequality yields

$$\begin{aligned} P(|\sum_{i=1}^n \xi_i x_{ij}| > \sqrt{5 C_m n \log(n \vee p)})) & \leq 2 \exp\{-\frac{5 C_m \log(n \vee p)}{2[C_m + M\sqrt{5 C_m}\frac{\sqrt{\log(n \vee p)}}{\sqrt{n}}]}\} \\ & \leq 2 \exp\{-\frac{5 \log(n \vee p)}{4}\} \end{aligned}$$

Let

$$\Omega_1 = \{ \max_{s+1 \le j \le p} | \sum_{i=1}^{n} \xi_i x_{ij} | \le \sqrt{5 C_m n \log(n \vee p)} \}$$

Then

$$P(\Omega_1) \ge 1 - 2 \exp\{ \log(p-s) - \frac{5 \log(n \vee p)}{4} \} \ge 1 - \nu_1$$

where $\nu_1 = 2 \exp\{ - \log(n \vee p)/4 \} \to 0$ as $n \to \infty$. Applying Bernstein's inequality to I_2 yields

$$P(| \sum_{K_1} x_{ij} | > \sqrt{5 C_m \log(n \vee p)})) \le 2 \exp\{ - \frac{5 C_m \log(n \vee p)}{2[\frac{|K_1| C_m}{n} + M \sqrt{5 C_m} \frac{\sqrt{\log(n \vee p)}}{\sqrt{n}}]} \}.$$

Define

$$\Omega_2 = \{ \max_{s+1 \le j \le p} | \sum_{i \in K_1} x_{ij} | \le \sqrt{5 C_m n \log(n \vee p)} \}$$

We obtain $P(\Omega_2) \ge 1 - \nu_1$. A similar argument will show that $P(\Omega_3) \ge 1 - \nu_1$, where

$$\Omega_3 = \{ \max_{s+1 \le j \le p} | \sum_{i \in K_2} x_{ij} | \le \sqrt{5 C_m n \log(n \vee p)} \}.$$

Note that $\Omega_1 \bigcup \Omega_2 \bigcup \Omega_3 \subset \{ |\varphi(y_i - \mathbf{x}_i^T \check{\beta}) x_{ij}| \le 3 \sqrt{5 C_m n \log(n \vee p)} \}$. Therefore,

$$P(|\varphi(y_i - \mathbf{x}_i^T \check{\beta}) x_{ij}| \le 3 \sqrt{5 C_m n \log(n \vee p)}) \ge 1 - 3\nu_1$$

Since $||\tilde{\beta}|| \sim O_p(\sqrt{s \log(n \vee p)/n})$, for n sufficiently large with probability approaching 1,

$$\frac{\lambda_n \omega_j}{3 \sqrt{5 C_m n \log(n \vee p)}} > 1.$$

With probability at least $1 - 3\nu_1$, we have

$$\frac{|\varphi(y_i - \mathbf{x}_i^T \check{\beta}) x_{ij}|}{3 \sqrt{5 C_m n \log(n \vee p)}} \le 1 < \frac{\lambda_n \omega_j}{3 \sqrt{5 C_m n \log(n \vee p)}},$$

for all $j > s$. This implies that with probability tending to 1

$$\frac{\partial Q(\beta)}{\partial \beta_j}|_{\check{\beta}} = \begin{cases} > 0 & \text{if } \check{\beta}_j > 0 \\ < 0 & \text{if } \check{\beta}_j < 0 \end{cases}$$

Since $Q(\beta)$ is a continuous function, $\hat{\beta}$, the minimizer of $Q(\beta)$ must satisfy $\hat{\beta}_b = 0$.

Lemma A.2 *Under the assumptions A1-A5, if $\lambda_n s/\sqrt{n} \to 0$ and $\omega_i = |\tilde{\beta}_{\tau j}|^{-1}$ for $0 \le j \le p$, then the adaptive L_1 quantile regression estimator is $\sqrt{n/s}$-consistent.*

Proof: We want to show that for any $\epsilon > 0$, there exists a sufficiently large constant, such that

$$P\{\inf_{\|\delta_{\mathbf{a}}\|=C} Q_a(\beta_a^* + \sqrt{\frac{s}{n}}\delta_a) > Q_a(\beta_a^*)\} > 1 - \epsilon \tag{A.2}$$

where $Q_a(\cdot)$ is the objective function restricted to the true underlying model, $\delta_a \in \mathbb{R}^s$ and $\|\delta\| = C$. Since the objective function $Q_a(\beta_a)$ is strictly convex, the inequality (A.2) implies, with probability at least $1-\epsilon$, the oracle quantile estimator lies in the shrinking ball $\{\beta^* + \sqrt{s/n}\delta_a : \delta_a \in \mathbb{R}^{s+1}, \|\delta_a\| \le C\}$. This provides the consistency result immediately.

$$\begin{aligned} Q_a(\beta_a^* + \sqrt{\frac{s}{n}}\delta_a) - Q_a(\beta_a^*) = & \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_{ia}^T(\beta_a^* + \sqrt{\frac{s}{n}}\delta_a)) - \rho(y_i - \mathbf{x}_{ia}^T\beta_a^*) \\ & + \lambda_n \sum_{j=0}^{s} \omega_j(|\beta_{\tau j}^* + \sqrt{\frac{s}{n}}\delta_{aj}| - |\beta_{\tau j}^*|) \end{aligned} \tag{A.3}$$

According to Knight(1998), for any $x \neq 0$, we have

$$|x - y| - |x| = -y[1(x > 0) - 1(x < 0)] + 2\int_0^y [1(x < t) - 1(x < 0)]dt$$

Then we have

$$\rho(x - y) - \rho(x) = y[1(x < 0) - \tau] + 2\int_0^y [1(x < t) - 1(x < 0)]dt$$

Hence, (A.3) can be written as

$$\begin{aligned}
&\sqrt{\frac{s}{n}}\sum_{i=1}^{n}\mathbf{x}_{ia}^T\delta_a[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)-\tau]\\
&+\sum_{i=1}^{n}\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt\\
&+\lambda_n\sum_{j=0}^{s}\omega_j(|\beta_{\tau j}^*+\sqrt{\frac{s}{n}}\delta_{aj}|-|\beta_{\tau j}^*|)\\
&:=\sqrt{\frac{s}{n}}T_1+T_2+T_3
\end{aligned}$$

Using independence and the Cauchy-Schwarz inequality,

$$\begin{aligned}
E[T_1^2]&=E[(\sum_{i=1}^{n}\mathbf{x}_{ia}^T\delta_a[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)-\tau])^2]\\
&=E[\sum_{i=1}^{n}(\mathbf{x}_{ia}^T\delta_a[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)-\tau])^2]\\
&\le n\tau(1-\tau)E[\|\mathbf{x}_{ia}\|^2\|\delta_a\|^2]\\
&\le ns\tau(1-\tau)C_mC^2.
\end{aligned}$$

Using Chebychev's inequality, we see that for any constant k

$$P\left(\sqrt{\frac{s}{n}}T_1>ksC^2\right)\le\frac{\tau(1-\tau)C_m}{C^2}. \tag{A.4}$$

Next, we deal with T_2. The goal is to show that $T_2\overset{p}{\ge}0.5s\underline{f}\kappa_0^2C^2$. Using independence and the fact that $V(X)\le EX^2$,

$$\begin{aligned}
V[T_2]&=V\left[\sum_{i=1}^{n}\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt\right]\\
&\le nE\left[\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt\right]^2
\end{aligned}$$

Given an $\eta > 0$ we have

$$
\begin{aligned}
&nE\left[\left(\left(\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt)^2 1(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|>\eta\right)\right]\\
\le& 4sE\left[(\mathbf{x}_{ia}^T\delta_a)^2 1\left(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|>\eta\right)\right]\\
\le& 4sE[|\mathbf{x}_{ia}^T\delta_a|^3]^{2/3}\left(P(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|>\eta)\right)^{1/3}, \qquad \text{(A.5)}
\end{aligned}
$$

where the last line follows from Holder's inequality. Under condition A4,

$$
E[|\mathbf{x}_{ia}^T\delta_a|^3] \le \frac{3}{8}\frac{\underline{f}^{3/2}}{\bar{f}'}\frac{E[|\mathbf{x}_{ia}^T\delta_a|^2]^{3/2}}{q}. \qquad \text{(A.6)}
$$

Applying Bernstein's inequality (Lemma 2.2.11 of Van Der Vaart and Wellner (1996)),

$$
P(|\mathbf{x}_{ia}^T\delta_a|>\eta\frac{\sqrt{n}}{\sqrt{s}}) \le 2\exp\left\{\frac{-\eta^2 n}{2s(C^2C_m+MC\eta\frac{\sqrt{n}}{\sqrt{s}})}\right\}. \qquad \text{(A.7)}
$$

Combining bounds (A.6) and (A.7) yields :

$$
\begin{aligned}
\text{RHS of (A.5)} &\le 4s\left(\frac{3}{8}\frac{\underline{f}^{3/2}}{\bar{f}'}\frac{E[|\mathbf{x}_{ia}^T\delta_a|^2]^{3/2}}{q}\right)^{2/3}\left(2\exp\left\{\frac{-\eta\sqrt{n}}{2MC\sqrt{s}}\right\}\right)^{1/3}\\
&\le 3^{2/3}2^{1/3}\frac{\underline{f}}{(\bar{f}'q)^{2/3}}C_mC^2s^2\exp\left\{\frac{-\eta\sqrt{n}}{6MC\sqrt{s}}\right\}\\
&= 3^{2/3}2^{1/3}\frac{\underline{f}}{(\bar{f}'q)^{2/3}}C_mC^2\exp\left\{2\log(s)-\frac{\eta\sqrt{n}}{6MC\sqrt{s}}\right\},
\end{aligned}
$$

which converges to 0 if η satisfies (C1): $\log(s) \sim o\left(\eta\sqrt{n}/(12MC\sqrt{s})\right)$ and (C2): $\eta\sqrt{n}/\sqrt{s} \to \infty$. On the other hand,

$$
\begin{aligned}
&nE\left[\left(\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt\right)^2 1\left(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|\le\eta\right)\right]\\
\le& 2n\eta E\left[\left(\int_0^{\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|}[1(y_i-\mathbf{x}_{ia}^T\beta_a^*<t)-1(y_i-\mathbf{x}_{ia}^T\beta_a^*<0)]dt\right)1\left(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|<\eta\right)\right]\\
=& 2n\eta E\left[\left(\int_0^{\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|}[F_{\epsilon|\mathbf{x}_i}(q_{\mathbf{x}_i}^*+t)-F_{\epsilon|\mathbf{x}_i}(q_{\mathbf{x}_i}^*)]dt\right)1\left(\sqrt{\frac{s}{n}}|\mathbf{x}_{ia}^T\delta_a|<\eta\right)\right] \qquad \text{(A.8)}
\end{aligned}
$$

If η is close to 0, then $F(t) - F(0) \leq \bar{f}t, \forall |t| < \eta$. Thus, we obtain

$$(A.7) \leq \bar{f}t\eta n E[(\int_0^{\sqrt{\frac{s}{n}}|x_{ia}^T\delta_a|} tdt)1(\sqrt{\frac{s}{n}}|x_{ia}^T\delta_a| < \eta)] \leq \bar{f}t\eta^3 n$$

which converges to 0, if η satisfies (C3): $\eta^3 n \to 0$. If η satisfies conditions C1, C2 and C3, then as $n \to \infty$ $V(T_2) \to 0$. By Chebyshev's inequality, we have

$$T_2 - nE\{\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a} [1(y_i - \mathbf{x}_{ia}^T\beta_a^* < t) - 1(y_i - \mathbf{x}_{ia}^T\beta_a^* < 0)]dt\} \xrightarrow{p} p$$

Using Cauchy-Schwartz inequality and a similar argument as in the proof of $V(T_2) \to 0$, we can show that for n sufficiently large

$$nE\left\{\int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a} [1(y_i - \mathbf{x}_{ia}^T\beta_a^* < t) - 1(y_i - \mathbf{x}_{ia}^T\beta_a^* < 0)]dt\right\} \geq \frac{1}{2}\underline{f}\kappa_0^2 C^2 s$$

Finally for T_3, we have

$$|\lambda_n \sum_{j=0}^{s} \omega_j(|\beta_j^* + \sqrt{\frac{s}{n}}\delta_{aj}| - |\beta_j^*|)| \leq \lambda_n \sum_{j=1}^{s} \omega_j \sqrt{\frac{s}{n}}|\delta_{aj}| \leq \lambda_n \frac{s}{\sqrt{n}} \max_{1\leq j\leq s} \frac{1}{|\beta_j^*|} C \to 0$$

Combining the fact that T_3 converges to zero in probability with (A.4), we see that for sufficiently large C, (A.3) is positive with probability at least $1 - \epsilon$ and (A.2) is satisfied.

Appendix A-2: Asymptotic Normality

Proof of Theorem 2.3.1: As in the foregoing proofs, we see that with probability at least $1 - 3\nu_1$, $\hat{\beta} = \check{\beta}$. Therefore, properties (1) and (2) are achieved automatically. We know that $\check{\beta} = ((\beta^* + \sqrt{s/n}\check{\delta}_a)^T, 0)^T$ where $\sqrt{s/n}\check{\delta}_a$ is the minimizer of the following function:

$$\begin{aligned} Q_a(\beta_a^* + \sqrt{\frac{s}{n}}\delta_a) - Q_a(\beta_a^*) &= \sqrt{\frac{s}{n}} \sum_{i=1}^{n} \mathbf{x}_{ia}^T\delta_a[1(\epsilon_i < q_{\mathbf{x}_i}^*) - \tau] \\ &+ \sum_{i=1}^{n} \int_0^{\sqrt{\frac{s}{n}}\mathbf{x}_{ia}^T\delta_a} [1(\epsilon_i < q_{\mathbf{x}_i}^* + t) - 1(\epsilon_i < q_{\mathbf{x}_i}^*)]dt \\ &+ \lambda_n \sum_{j=0}^{s} \omega_j(|\beta_j^* + \sqrt{\frac{s}{n}}\delta_{aj}| - |\beta_j^*|) \\ &:= J_1 + J_2 + J_3 \end{aligned}$$

And with probability at least $1-\epsilon$, $\check{\delta}_a$ locates in a ball $B_\epsilon := \{\delta_a : \|\delta\| \leq C\}$ for some constant C that implicitly depends on ϵ. For any $\delta_a \in B_\epsilon$, using the argument as in the proof of consistency, we can show that

$$E|J_1/s|^2 \leq C_m\|\delta_a\|^2, \qquad J_2 \xrightarrow{p} \frac{1}{2}f(q^*)s\delta_a^T\Sigma_S\delta_a,$$

and

$$|J_3| \leq \|\delta_a\|O(\sqrt{s}(\log(n))^{\gamma/2}\log(n\vee p))\frac{s}{\sqrt{n}}\max_{1\mathrm{J}\leq s}\frac{1}{|\tilde{\beta}_j|} = o(1).$$

Thus, with probability at least $1-3\nu_1-\epsilon$, minimizing $Q_a(\beta_a^* + \sqrt{s/n}\delta_a) - Q_a(\beta_a^*)$ is equivalent to minimizing

$$\sqrt{\frac{s}{n}}\sum_{i=1}^n \mathbf{x}_{ia}^T\delta_a[1(\epsilon_i < q_{\mathbf{x}_i}^*) - \tau] + \frac{1}{2}f(q^*)s\delta_a^T\Sigma_S\delta_a,$$

which provides

$$\check{\delta}_a = \frac{\sum_{i=1}^n \Sigma_s^{-1}\mathbf{x}_{ia}[1(\epsilon_i < q_{\mathbf{x}_i}^*) - \tau]}{f(q^*)\sqrt{ns}}$$

Therefore, with probability at least $1-3\nu_1-\epsilon$

$$\sqrt{n}u_s^{-1}\alpha^T(\hat{\beta}_a - \beta_a^*) = \sqrt{n}\frac{\sum_{i=1}^n u_s^{-1}\alpha^T\Sigma_s^{-1}\mathbf{x}_{ia}[1(\epsilon_i < q_{\mathbf{x}_i}^*) - \tau]}{f(q^*)n}$$

Denote ζ_i by $u_s^{-1}\alpha^T\Sigma_s^{-1}\mathbf{x}_{ia}[1(\epsilon_i < q_{\mathbf{x}_i}^*) - \tau]$ for $i = 1, \cdots, n$. Then $E[\zeta_i] = 0$ and $\mathrm{Var}[\zeta_i] = \tau(1-\tau)$ Therefore, we have have

$$\sqrt{n}\frac{\sum_{i=1}^n \zeta_i}{f(q^*)n} \xrightarrow{d} N(0, \frac{\tau(1-\tau)}{f^2(q^*)})$$

. which completes the proof.

Appendix B Proofs for Chapter 3

Proof of Theorem 3.3.1:

We sketch the proofs in this section. Detailed proofs are provided in Appendix C.

Let

$$Q_\beta(\theta) = \sum_{i=1}^{n} \left[\sum_{k=0}^{q} \beta_k \rho_{\tau_k} \left(y_i - a_{0k} - \sum_{j=0}^{p} \frac{1}{j!} a_j (x_i - x_0)^j \right) \right] K\left(\frac{x_i - x_0}{h} \right)$$

We can see that minimizing $Q_\beta(\theta)$ with respect to θ is equivalent to minimizing $Q_\beta(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\beta(\theta^*)$ with respect to $\mathbf{u}$, where $\mathbf{u} = (u_{01}, \cdots, u_{0q}, u_0, u_1, \cdots, u_p)^T$ is a $q+1+p$ vector.

Let $\Delta_{0,i} = \sum_{j=0}^{p} \frac{1}{j!} u_j (x_i - x_0)^j$, and $\Delta_{k,i} = \Delta_{0,i} + u_{0k}$, for $k = 1, \cdots, q$. Applying the identity [Knight (1998)],

$$\rho_\tau(x-y) - \rho_\tau(x) = y[1(x \le 0) - \tau] + \int_0^y \{1(x \le z) - 1(x \le 0)\} dz$$

yields

$$\begin{aligned}
&= \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \left\{ -2\beta_0 \Delta_{0,i}(\sigma\epsilon_i + r_{i,p}) + \sum_{k=1}^{q} \beta_k \left[1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) - \tau_k \right] \Delta_{k,i} \right\} \\
&+ \frac{\beta_0}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \Delta_{0,i}^2 \\
&+ \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n} \right) \sum_{k=1}^{q} \beta_k \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma} \right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) \right\} dz \\
&:= I_1 + I_2 + I_3
\end{aligned}$$

By some algebra, we can show that $Var[I_3] \to 0$. Applying Chebyshev's inequality yields $I_3 - E[I_3] \xrightarrow{p} 0$ and

$$E[I_3] = n \sum_{k=1}^{q} \beta_k E\left[Z_{n,k,i}(\mathbf{u}) 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| \le \eta \right) \right] + n \sum_{k=1}^{q} \beta_k E\left[Z_{n,k,i}(\mathbf{u}) 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| > \eta \right) \right] \quad \text{(B.1)}$$

where $Z_{n,k,i}(\mathbf{u}) = K\left(\frac{x_i - x_0}{h_n} \right) \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma} \right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) \right\} dz$. Using the same argument as in the proof of $E[I_3^2] \to 0$, we can show that

$$n \sum_{k=1}^{q} \beta_k E\left[Z_{n,k,i}(\mathbf{u}) 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| \le \eta \right)\right]$$

$$= \frac{1}{2} \sum_{k=1}^{q} \beta_k f(q_{\tau_k}) g(x_0) \int_{-M}^{M} K(t_i) \left(u_{0k} + \sum_{j=0}^{p} \frac{u_j}{j!} h_n^j t_i^j \right)^2 dt_i + o(\|A_{h_n}\mathbf{u}\|^2) \qquad \text{(B.2)}$$

where $t_i = (x_i - x_0)/h_n$.

Applying the Cauchy-Schwartz inequality, we have

$$n \sum_{k=1}^{p} \beta_k E\left[Z_{n,k,i}(\mathbf{u}) 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| > \eta \right)\right] = o(\|A_{h_n}\mathbf{u}\|^2) = o(1) \qquad \text{(B.3)}$$

Therefore,

$$I_3 \xrightarrow{p} \frac{1}{2} \sum_{k=1}^{p} \beta_k f(q_{\tau_k}) g(x_0) \int_{-M}^{M} K(t_i) \left[u_{0k} + \sum_{j=0}^{p} \frac{u_j}{j!} h_n^j t_i^j \right]^2 dt_i \qquad \text{(B.4)}$$

According to the law of large numbers, we know

$$I_2 \xrightarrow{a.s.} \beta_0 g(x_0) \int_{-M}^{M} K(t_i) \left(\sum_{j=0}^{p} \frac{u_j}{j!} h_n^j t_i^j \right)^2 dt_i \qquad \text{(B.5)}$$

Since I_1 can be written as $\mathbf{W}_{\beta,n}^T A_{h_n} \mathbf{u}$, then

$$Q_\beta(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\beta(\theta^*) = g(x_0)\mathbf{u}^T A_{h_n} S(\beta) A_{h_n} \mathbf{u} + \mathbf{W}_{\beta,n}^T A_{h_n} \mathbf{u} + o_p(\|A_{h_n}\mathbf{u}\|)$$

Let $\hat{\mathbf{u}}$ denote the minimizer of $Q(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q(\theta^*)$, we have

$$S(\beta) A_{h_n} \hat{\mathbf{u}} = -\frac{1}{2g(x_0)} \mathbf{W}_{\beta,n} + o_p(1)$$

According to the definition of $\mathbf{W}_{\beta,n}$, applying CLT yields

$$\frac{\alpha^T \mathbf{W}_{\beta,n} - E[\alpha^T \mathbf{W}_{\beta,n}]}{\sqrt{\mathrm{Var}[\alpha^T \mathbf{W}_{\beta,n}]}} \xrightarrow{L} N(0,1)$$

for any nonzero $(1+q+p)\times 1$ vector α. Thus, The Cramer-Wald device provides us

$$[\mathrm{Cov}(\mathbf{W}_{\beta,n})]^{-\frac{1}{2}}(\mathbf{W}_{\beta,n}-E[\mathbf{W}_{\beta,n}]) \xrightarrow{L} N(\mathbf{0},\mathbf{I}_{(1+q+p)\times(1+q+p)})$$

where $\mathrm{Cov}(\mathbf{W}_{\beta,n})$ is the covariance matrix of $\mathbf{W}_{\beta,n}$. It is easy to check

$$\mathrm{Cov}(\mathbf{W}_{\beta,n})] \xrightarrow{p} g(x_0)\Sigma(\beta)$$

Therefore, we have

$$S(\beta)A_{h_n}\hat{\mathbf{u}}+\frac{1}{2g(x_0)}E[\mathbf{W}_{\beta,n}] \xrightarrow{L} N(\mathbf{0},\frac{1}{4g(x_0)}\Sigma(\beta)) \tag{B.6}$$

This completes the proof of Theorem 3.3.1. □

Corollary 3.3.1 and Corollary 3.3.2 are special cases of Theorem 3.1. We omit the proofs here. Complete proofs can be seen in the supplemental material.

Proof of Corollary 3.3.3:

Consider $Q_{\hat{\beta}}(\theta^*+(nh_n)^{-1/2}\mathbf{u})-Q_{\hat{\beta}}(\theta^*)$. We can write it as:

$$Q_\gamma(\theta^*+(nh_n)^{-1/2}\mathbf{u})-Q_\gamma(\theta^*)+Q_{\beta_{opt}}(\theta^*+(nh_n)^{-1/2}\mathbf{u})-Q_{\beta_{opt}}(\theta^*)$$

where $\gamma=\hat{\beta}-\beta_{opt}$. Since β_{opt} is a fixed vector given the error structure, then using the same arguments as in Theorem 3.1, we obtain that

$$Q_{\beta_{opt}}(\theta^*+(nh_n)^{-1/2}\mathbf{u})-Q_{\beta_{opt}}(\theta^*)=g(x_0)\mathbf{u}^TA_{h_n}S(\beta_{opt})A_{h_n}\mathbf{u}+\mathbf{W}^T_{\beta_{opt},n}A_{h_n}\mathbf{u}+o_p(\|A_{h_n}\mathbf{u}\|^2)$$

And we have

$$
\begin{aligned}
& Q_\gamma(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\gamma(\theta^*) \\
= & \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \left\{ -2\gamma_0(\sigma\epsilon_i + r_{i,p})\Delta_{0,i} + \sum_{k=1}^{q} \gamma_k \left[1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k \right] \Delta_{k,i} \right\} \\
+ & \frac{\gamma_0}{nh_n} \Delta_{0,i}^2 \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \\
+ & \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \sum_{k=1}^{q} \gamma_k \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) \right\} dz \qquad \text{(B.7)}
\end{aligned}
$$

Since

$$
\begin{aligned}
& \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) \right\} dz \\
\xrightarrow{p} & \frac{1}{2\sigma} f(q_{\tau_k}) g(x_0) \int_{-M}^{M} K(t_i) \left[u_{0k} + \sum_{j=0}^{p} \frac{u_j}{j!} h_n^j t_i^j \right]^2 dt_i
\end{aligned}
$$

and

$$
\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \Delta_{0,i}^2 \xrightarrow{a.s.} g(x_0) \int_{-M}^{M} K(t_i) \left[\sum_{j=0}^{p} \frac{u_j}{j!} h_n^j t_i^j \right]^2 dt_i
$$

then the last two terms of (B.7) are $o(\|A_{h_n}\mathbf{u}\|^2)$. Applying Slutsky's Theorem, we can show that the first term of (B.7) is $o_p(A_{h_n}\mathbf{u})$. Therefore,

$$
Q_{\hat{\beta}}(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_{\hat{\beta}}(\theta^*) = g(x_0)\mathbf{u}^T A_{h_n} S(\beta_{opt}) A_{h_n}\mathbf{u} + \mathbf{W}^T_{\beta_{opt},n} A_{h_n}\mathbf{u} + o_p(\|A_{h_n}\mathbf{u}\|^2)
$$

This completes the proof of Corollary 3.3.3. □

In order to prove Theorem 3.3.2, the following lemmas are needed.

Lemma B.1 *If assumptions B,C,D and E are satisfied, and $E[\epsilon_i^2]$ does not exist, then*

(a) $\frac{\sum_{i=1}^{n} |\epsilon|}{a_n\sqrt{n}} \to 0$

(b) $\frac{\sum_{i=1}^{n}\epsilon_i^2}{a_n^2} \to O_p(1)$

(c) $\frac{\sum_{i=1}^{n}\frac{1}{\sqrt{h_n}}K(\frac{X_i-x_0}{h_n})\epsilon_i}{a_n} \leq O_p(1)$

where $\{a_n\}$ *is the norming sequence for* ϵ_i*, such that* $\frac{\sum_{i=1}^{n}\epsilon_i}{a_n} \xrightarrow{L} S$

Lemma B.2 *Let* $\hat{\beta}_{0p} = \frac{1}{\tilde{\sigma}}\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}1(\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1} \geq 0)$*, where* $\tau_l = 1/2$*. Then* $\hat{\beta}_0 \leq \hat{\beta}_{0p}$ *almost surely.*

We omit the proofs here. They can be found in the supplementary material.

Proof of Theorem 3.3.2:

Let $\tilde{\zeta}_i = \sigma\tilde{\epsilon}_i, i = 1, \cdots, n$ denote the residuals of a $\sqrt{nh_n}$-consistent fit. Since σ and ϵ_i are unknown, we use $\tilde{\sigma}^2 = \sum_{i=1}^{n}\tilde{\zeta}_i^2/n$ to denote the sample variance of $\sigma\epsilon_i$. Then we have $\tilde{\tau}_{0,l} = -\sum_{i=1}^{n}|\tilde{\zeta}_i|/(2n\tilde{\sigma})$ and $\tilde{f}(0) = \frac{1}{2nb}\sum_{i=1}^{n}\tilde{\sigma}1\left(|\sigma\tilde{\epsilon}_i| < b\right)$, for some $b \sim O(n^{-1/5})$.

The key of the proof is to show that the impact from the least square part is negligible. Since $\hat{\beta}_0 \xrightarrow{p} \beta_{0opt} = 0$, we need to show

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{x_i - x_0}{h_n}\right)\hat{\beta}_0\sigma\epsilon_i = o_p(1) \tag{A.8}$$

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{k=1}^{q}\hat{\beta}_k\sum_{i=1}^{n}K\left(\frac{x_i - x_0}{h_n}\right)\left[1\left(\epsilon_i < q_{\tau_k}\right) - \tau_k\right] = O_p(1) \tag{A.9}$$

According to Lemma B.2, if we can show that $\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h_n}\right)\hat{\beta}_{0p}\epsilon_i = o_p(1)$, then (A.8) can be directly inferred.

$$\begin{aligned}
&\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{x_i - x_0}{h_n}\right)\hat{\beta}_{0p}\sigma\epsilon_i \\
= &\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}1\left(\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1} \geq 0\right)\frac{a_n}{\sqrt{n}}\frac{1}{a_n\sqrt{h_n}}\sum_{i=1}^{n}K\left(\frac{x_i - x_0}{h_n}\right)\sigma\epsilon_i
\end{aligned}$$

Since $E[\epsilon_i^2]$ does not exist, by Lemma B.1, we have

$$\frac{1}{a_n\sqrt{h_n}}\sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right)\sigma\epsilon_i \leq O_p(1)$$

and

$$\begin{aligned}
& \frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}^2(0)+8\tilde{f}(0)\tilde{\tau}_{0,l}+1}\frac{a_n}{\sqrt{n}} \\
= & \frac{1-\frac{4\sum_{i=1}^{n}\tilde{\sigma}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{2n\tilde{\sigma}}}{4\left(\frac{\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\right)^2\tilde{\sigma}^2-\frac{8\sum_{i=1}^{n}\tilde{\sigma}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{2n\tilde{\sigma}}+1}\frac{a_n}{\sqrt{n}} \\
= & \frac{1-\frac{2\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{a_n\sqrt{n}}\frac{a_n}{\sqrt{n}}}{4\left(\frac{\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\right)^2\frac{\tilde{\epsilon}_i^2}{a_n^2}\frac{a_n^2}{n}-\frac{4\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{a_n\sqrt{n}}\frac{a_n}{\sqrt{n}}+1}\frac{a_n}{\sqrt{n}} \\
= & \frac{\frac{\sqrt{n}}{a_n}-\frac{2\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{a_n\sqrt{n}}}{4\left(\frac{\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\right)^2\frac{\tilde{\epsilon}_i^2}{a_n^2}-\frac{4\sum_{i=1}^{n}1(|\sigma\tilde{\epsilon}_i|<b)}{2nb}\frac{\sum_{i=1}^{n}|\sigma\tilde{\epsilon}_i|}{a_n\sqrt{n}}\frac{\sqrt{n}}{a_n}+\frac{n}{a_n^2}} \\
\xrightarrow{p} & \, 0
\end{aligned}$$

Consequently,

$$\tilde{\sigma}\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n} K(\frac{x_i - x_0}{h_n})\hat{\beta}_{0p}\epsilon_i = o_p(1)$$

From the above proof, we can see that as $n \to \infty$,

$$\sum_{k=1}^{q}\frac{\tilde{f}(\tilde{q}_{\tau_k})}{\tilde{\sigma}}\tilde{\sigma}\hat{\beta}_k = \sum_{k=1}^{q}\tilde{f}(\tilde{q}_{\tau_k})\hat{\beta}_k = 1 - \tilde{\sigma}\hat{\beta}_0 \to 1$$

Since $\tilde{f}(\tilde{q}_{\tau_k})/\tilde{\sigma}$ is bounded for $1 \leq k \leq q$, then $\sum_{k=1}^{q}\tilde{\sigma}\hat{\beta}_k$ is bounded away from 0. There, (A.9) can be inferred.

This completes the proof of Theorem 3.3.2 $\square$

The proof of Theorem 3.3.3 is essentially the same as for Theorem 3.3.2. Thus, we omit it here.

Appendix C Supplemental Material

Proof of Theorem 3.3.1:

$$
\begin{aligned}
& Q_\beta(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\beta(\theta^*) \\
&= \beta_0 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!}(m^{(j)}(x_0) + \frac{u_j}{\sqrt{nh_n}})(x_i - x_0)^j \right)^2 K\left(\frac{x_i - x_0}{h_n}\right) \\
&- \beta_0 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \frac{1}{j!} m^{(j)}(x_0)(x_i - x_0)^j \right)^2 K\left(\frac{x_i - x_0}{h_n}\right) \\
&+ \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - \sigma q_{\tau_k} - \frac{u_{0k}}{\sqrt{nh_n}} - \sum_{j=0}^{p} \frac{1}{j!}(m^{(j)}(x_0) + \frac{u_j}{\sqrt{nh_n}})(x_i - x_0)^j \right) K\left(\frac{x_i - x_0}{h_n}\right) \\
&- \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} \rho_{\tau_k} \left(y_i - \sigma q_{\tau_k} - \sum_{j=0}^{p} \frac{1}{j!} m^{(j)}(x_0)(x_i - x_0)^j \right) K\left(\frac{x_i - x_0}{h_n}\right) \\
&:= T_1 + T_2
\end{aligned}
$$

Consider T_1 first. Let $\Delta_{0,i} = \sum_{j=0}^{p} u_j (x_i - x_0)^j / j!$, $\Delta_{k,i} = \Delta_{0,i} + u_{0k}$, and $r_{i,p} = m(x_i) - \sum_{j=0}^{p} m^{(j)}(x_0)(x_i - x_0)^j / j!$, then

$$
\begin{aligned}
T_1 &= -\beta_0 \sum_{i=1}^{n} \frac{\Delta_{0,i}}{\sqrt{nh_n}} \left(2\sigma\epsilon_i + 2r_{i,p} - \frac{\Delta_{0,i}}{\sqrt{nh_n}} \right) K\left(\frac{x_i - x_0}{h_n}\right) \\
&= -\frac{2\beta_0}{\sqrt{nh_n}} \sum_{i=1}^{n} \left\{ \Delta_{0,i}(\sigma\epsilon_i + r_{i,p}) K\left(\frac{x_i - x_0}{h_n}\right) \right\} + \frac{\beta_0}{nh_n} \sum_{i=1}^{n} \Delta_{0,i}^2 K\left(\frac{x_i - x_0}{h_n}\right)
\end{aligned}
$$

We consider T_2 next. Applying the identity [Knight (1998)], for $x \neq 0$

$$
\rho_\tau(x - y) - \rho_\tau(x) = y[1(x \le 0) - \tau] + \int_0^y \{1(x \le z) - 1(x \le 0)\} dz
$$

yields

$$
\begin{aligned}
T_2 &= \frac{1}{\sqrt{nh_n}} \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} \Delta_{k,i} \left[1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) - \tau_k \right] K(\frac{x_i - x_0}{h_n}) \\
&+ \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma} \right) - 1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma} \right) \right\} dz K\left(\frac{x_i - x_0}{h_n}\right)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& T_1 + T_2 \\
&= \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \left\{ -2\beta_0 \Delta_{0,i}(\sigma\epsilon_i + r_{i,p}) + \sum_{k=1}^{q} \beta_k \left[1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k \right] \Delta_{k,i} \right\} \\
&+ \frac{\beta_0}{nh_n} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \Delta_{0,i}^2 \\
&+ \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h_n}\right) \sum_{k=1}^{q} \beta_k \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) \right\} dz \\
&:= I_1 + I_2 + I_3
\end{aligned}
$$

Denote $K\left(\frac{x_i - x_0}{h_n}\right) \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq (\sigma q_{\tau_k} - r_{i,p} + z)/\sigma\right) - 1\left(\epsilon_i \leq (\sigma q_{\tau_k} - r_{i,p})/\sigma\right) \right\} dz$ by $Z_{n,k,i}(\mathbf{u})$. Then

$$
\begin{aligned}
& E[I_3^2] \\
&= nE\left[K^2\left(\frac{x_i - x_0}{h_n}\right) \left\{ \sum_{k=1}^{q} \beta_k \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) \right\} dz \right\}^2 \right] \\
&\leq nE\left[K^2\left(\frac{x_i - x_0}{h_n}\right) \sum_{k=1}^{q} q\beta_k^2 \left\{ \int_0^{\frac{\Delta_{k,i}}{\sqrt{nh_n}}} \left\{ 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p} + z}{\sigma}\right) - 1\left(\epsilon_i \leq \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) \right\} dz \right\}^2 \right] \\
&\leq nq \sum_{k=1}^{q} \beta_k^2 E[Z_{n,k,i}^2(\mathbf{u})]
\end{aligned}
$$

For any $\eta > 0$, we have

$$
\begin{aligned}
& nqE\left[Z_{n,k,i}^2(\mathbf{u}) 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| > \eta \right) \right] \\
&\leq nqE\left[K^2\left(\frac{x_i - x_0}{h_n}\right) \left(\int_0^{\frac{|\Delta_{k,i}|}{\sqrt{nh_n}}} 2dz \right)^2 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| > \eta \right) \right] \\
&\leq 4qE\left[K^2(\frac{x_i - x_0}{h_n}) \frac{\Delta_{k,i}^2}{h_n} 1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}| > \eta \right) \right] \qquad \text{(S.1)}
\end{aligned}
$$

Then (S.1) converges to 0, as $\sqrt{nh_n}\eta \to \infty$.

Choose η to be close to 0, and condition on x_i's, we obtain

$$
\begin{aligned}
& nqE\left[Z^2_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right] \\
\le\, & 2nq\eta E\left[K^2\left(\frac{x_i-x_0}{h_n}\right)\int_0^{|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|}\left\{1\left(\epsilon_i\le\frac{\sigma q_{\tau_k}-r_{i,p}+z}{\sigma}\right)-1\left(\epsilon_i\le\frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma}\right)\right\}dz\right. \\
& \left.1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right] \\
\le\, & 2nq\eta E\left[K^2\left(\frac{x_i-x_0}{h_n}\right)\left\{\int_0^{|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|}\left\{f\left(\frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma}\right)\frac{z}{\sigma}+o(z)\right\}dz\right\}1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right] \\
\le\, & q\eta E\left[K^2\left(\frac{x_i-x_0}{h_n}\right)f\left(\frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma}\right)\frac{\Delta^2_{k,i}}{\sigma h_n}+o\left(\frac{\Delta^2_{k,i}}{h_n}\right)\right] \\
=\, & q\eta\int_{-\infty}^{\infty}\left\{K^2\left(\frac{x_i-x_0}{h_n}\right)(f(q_{\tau_k})+o(1))(g(x_0)+o(1))\frac{\Delta^2_{k,i}}{\sigma h_n}+o\left(\frac{\Delta^2_{k,i}}{h_n}\right)\right\}dx_i \\
=\, & \frac{q\eta}{\sigma}f(q_{\tau_k})g(x_0)\int_{-M}^{M}K^2(t_i)\left(u_{0k}+\sum_{j=0}^{q}\frac{u_j}{j!}h_n^jt_i^j\right)^2dt_i+o(q\eta\|A_{h_n}\mathbf{u}\|^2) \qquad \text{(S.2)}
\end{aligned}
$$

where $t_i=(x_i-x_0)/h_n$ and A_{h_n} is a $(q+1+p)\times(q+1+p)$ diagonal matrix with diagonal elements $(1,\cdots,1,h_n^0/0!,\cdots,h_n^p/p!)$.

Since we can choose η arbitrarily small, (S.2) implies

$$
nqE\left[Z^2_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right]=o(\|A_{h_n}\mathbf{u}\|^2)=o(1) \qquad \text{(S.3)}
$$

(S.1) and (S.3) together indicate that $E[I_3^2]=o(1)$. Applying Chebyshev's inequality yields $I_3-E[I_3]\xrightarrow{p}0$ and

$$
E[I_3]=n\sum_{k=1}^{q}\beta_kE\left[Z_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right]+n\sum_{k=1}^{q}\beta_kE\left[Z_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|>\eta\right)\right]
$$

Using the same argument in the proof of $E[I_3^2]\to 0$, we can show that

$$\begin{aligned}
&n\sum_{k=1}^{q}\beta_k E\left[Z_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|\le\eta\right)\right]\\
=&\frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})g(x_0)\int_{-M}^{M}K(t_i)\left(u_{0k}+\sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right)^2 dt_i+o(\|A_{h_n}\mathbf{u}\|^2)
\end{aligned}$$

Applying Cauchy-Schwartz inequality, we have

$$n\sum_{k=1}^{p}\beta_k E\left[Z_{n,k,i}(\mathbf{u})1\left(|\frac{\Delta_{k,i}}{\sqrt{nh_n}}|>\eta\right)\right]=o(\|A_{h_n}\mathbf{u}\|^2)=o(1)$$

Therefore,

$$I_3\xrightarrow{p}\frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})g(x_0)\int_{-M}^{M}K(t_i)\left[u_{0k}+\sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right]^2 dt_i \qquad \text{(S.4)}$$

According to the law of large number, we know

$$I_2\xrightarrow{a.s.}\beta_0 g(x_0)\int_{-M}^{M}K(t_i)\left(\sum_{j=0}^{p}\frac{u_j}{j!}h_n^j t_i^j\right)^2 dt_i \qquad \text{(S.5)}$$

Let $\xi_{\beta,i}=-2\beta_0(\sigma\epsilon_i+r_{i,p})+\sum_{k=1}^{q}\beta_k[1(\epsilon_i\le(\sigma q_{\tau_k}-r_{i,p})/\sigma)-\tau_k]$. We define $\mathbf{W}_{\beta,n}=(w_{\beta,01},\cdots,w_{\beta,0q},w_{\beta,0},w_{\beta,1},\cdots,w_{\beta,p}$ where

$$\begin{aligned}
w_{\beta,0k}&=\beta_k\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h_n}\right)[1\left(\epsilon_i\le\frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma}\right)-\tau_k], && k=1,\cdots,q\\
w_{\beta,j}&=\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{x_i-x_0}{h_n}\right)\left(\frac{x_i-x_0}{h_n}\right)^j\xi_{\beta,i}, && j=0,\cdots,p.
\end{aligned}$$

Let

$$S(\beta)=\begin{pmatrix} S_{11}(\beta) & S_{12}(\beta)\\ S_{21}(\beta) & S_{22}(\beta)\end{pmatrix}$$

where $S_{11}(\beta)$ is a $q\times q$ diagonal matrix with diagonal elements $\beta_k f(q_{\tau_k})/(2\sigma)$, for $k=1,\cdots,q$, $S_{22}(\beta)$ is a $(p+1)\times(p+1)$ matrix with (j,j')-entry $(\beta_0+\sum_{k=1}^{q}\beta_k f(q_{\tau_k})/(2\sigma))\mu_{(j+j'-2)}$, i.e. $\mu_{(j+j'-2)}$, for $j,j'=1,\cdots,p+1$, and $S_{12}(\beta)=S_{21}(\beta)^T$ is a $q\times(p+1)$ matrix with (k,j)-entry $\beta_k f(q_{\tau_k})/(2\sigma)\mu_{j-1}$, for $k = 1, \cdots, q; j = 1, \cdots, p+1$.

Since I_1 can be written as $\mathbf{W}_{\beta,n}^T A_{h_n}\mathbf{u}$, then

$$Q_\beta(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\beta(\theta^*) = g(x_0)\mathbf{u}^T A_{h_n} S(\beta) A_{h_n}\mathbf{u} + \mathbf{W}_{\beta,n}^T A_{h_n}\mathbf{u} + o_p(\|A_{h_n}\mathbf{u}\|^2)$$

Then minimizing $Q_\beta(\theta^* + (nh_n)^{-1/2}\mathbf{u}) - Q_\beta(\theta^*)$ with respect to $\mathbf{u}$ yields that

$$S(\beta)A_{h_n}\hat{\mathbf{u}} = -\frac{1}{2g(x_0)}\mathbf{W}_{\beta,n} + o_p(1)$$

We define $\mathbf{W}_{\beta,n}^* = (w_{\beta,01}^*, \cdots, w_{\beta,0q}^*, w_{\beta,0}^*, w_{\beta,1}^*, \cdots, w_{\beta,p}^*)^T$, where

$$w_{\beta,0k}^* = \beta_k \frac{1}{\sqrt{nh_n}} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)[1(\epsilon_i \le q_{\tau_k}) - \tau_k], \qquad k = 1, \cdots, q$$

$$w_{\beta,j}^* = \frac{1}{\sqrt{nh_n}} \sum_{i=1}^n K\left(\frac{x_i - x_0}{h_n}\right)\left(\frac{x_i - x_0}{h_n}\right)^j \eta_{\beta,i}, \qquad j = 0, \cdots, p$$

where $\eta_{\beta,i} = -2\beta_0\sigma\epsilon_i + \sum_{k=1}^q \beta_k[1(\epsilon_i \le q_{\tau_k}) - \tau_k]$.

Let

$$V_\beta = 4\beta_0^2\sigma^2 - 4\beta_0\sum_{k=1}^q \beta_k\sigma\tau_{0,k} + \sum_{k,k'=1}^q \beta_k\beta_{k'}\tau_{k,k'},$$

and we define

$$\Sigma(\beta) = \begin{pmatrix} \Sigma_{11}(\beta) & \Sigma_{12}(\beta) \\ \Sigma_{21}(\beta) & \Sigma_{22}(\beta) \end{pmatrix}$$

where $\Sigma_{11}(\beta)$ is a $q \times q$ matrix with (k, k')-entry $\beta_k\beta_{k'}\nu_0\tau_{k,k'}$, for $k, k' = 1, \cdots, q$, $\Sigma_{22}(\beta)$ is a $(p+1)\times(p+1)$ matrix with (j, j')th element $V_\beta\nu_{(j+j'-2)}$, for $j, j' = 1, \cdots, p+1$, and $\Sigma_{12}(\beta) = \Sigma_{21}^T(\beta)$ is a $q\times(p+1)$ matrix with (k, j)-entry $(-2\beta_0\beta_k\sigma\tau_{0,k} + \beta_k\sum_{k'=1}^q \beta_{k'}\tau_{k,k'})\nu_{(j-1)}$, for $k = 1, \cdots, q; j = 1, \cdots, (p+1)$.

It is easy to check

$$\text{Cov}(\mathbf{W}_{\beta,n}^*) \xrightarrow{p} g(x_0)\Sigma(\beta).$$

Therefore, we have

$$\mathbf{W}_{\beta,n}^* \xrightarrow{L} N(\mathbf{0}, g(x_0)\Sigma(\beta))$$

Moreover, we can show that for any nonzero $(1+q+p)\times 1$ vector α, $\text{Var}(\alpha^T(\mathbf{W}_{\beta,n}^* - \mathbf{W}_{\beta,n})) = o_p(1)$, and applying Central limit theorem (CLT) yields

$$\frac{\alpha^T \mathbf{W}_{\beta,n} - E[\alpha^T \mathbf{W}_{\beta,n}]}{\sqrt{\text{Var}[\alpha^T \mathbf{W}^*_{\beta,n}]}} \xrightarrow{L} N(0,1)$$

Therefore, we have

$$S(\beta) A_{h_n} \hat{\mathbf{u}} + \frac{1}{2g(x_0)} E[\mathbf{W}_{\beta,n}] \xrightarrow{L} N\left(\mathbf{0}, \frac{1}{4g(x_0)} \Sigma(\beta)\right) \tag{S.6}$$

This completes the proof of Theorem 3.3.1.

Proof of Corollary 3.3.1:

$$\begin{aligned} E[w_{\beta,0}] = &-2\beta_0 \sqrt{nh_n} E[\frac{1}{h_n} K\left(\frac{x_i - x_0}{h_n}\right) r_{ip}] \\ &+ \sum_{k=1}^{q} \beta_k \sqrt{nh_n} E\left[\frac{1}{h_n} K\left(\frac{x_i - x_0}{h_n}\right) \left\{1\left(\epsilon_i \le \frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}\right) - \tau_k\right\}\right] \\ &:= J_1 + J_2 \end{aligned}$$

$$\begin{aligned} J_1 &= -2\beta_0 \sqrt{nh_n} \int_{-M}^{M} K(t_i) \left(m(x_0 + t_i h_n) - \sum_{j=0}^{p} \frac{m^{(j)}(x_0)}{j!} (t_i h_n)^j \right) g(x_0 + t_i h_n) dt_i \\ &= -2\beta_0 \sqrt{nh_n} \int_{-M}^{M} K(t_i) \left(\frac{m^{(p+1)}(x_0)}{(p+1)!} (t_i h_n)^{p+1} + o((t_i h_n)^{p+1}) \right) g(x_0 + t_i h_n) dt_i \\ &= -\frac{2}{(p+1)!} \beta_0 g(x_0) m^{(p+1)}(x_0) \mu_{p+1} \sqrt{nh_n} h_n^{p+1} + o(\sqrt{nh_n} h_n^{p+1}) \end{aligned}$$

and condition on $\mathbf{x}$

$$\begin{aligned} J_2 &= \sum_{k=1}^{q} \beta_k \sqrt{nh_n} E\left[\frac{1}{h_n} K\left(\frac{x_i - x_0}{h_n}\right) \left\{F(\frac{\sigma q_{\tau_k} - r_{i,p}}{\sigma}) - F(q_{\tau_k})\right\}\right] \\ &= -\sum_{k=1}^{q} \beta_k \sqrt{nh_n} E\left[\frac{1}{h_n} K\left(\frac{x_i - x_0}{h_n}\right) \left\{f(q_{\tau_k}) \frac{r_{i,p}}{\sigma} + o(r_{i,p})\right\}\right] \\ &= -\frac{1}{\sigma(p+1)!} \sum_{k=1}^{q} \beta_k f(q_{\tau_k}) g(x_0) m^{(p+1)}(x_0) \mu_{p+1} \sqrt{nh_n} h_n^{p+1} + o(\sqrt{nh_n} h_n^{p+1}) \end{aligned}$$

Thus,

$$E[w_{\beta,0}] = -\frac{2}{(p+1)!}\left(\beta_0 + \frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\right) g(x_0)m^{(p+1)}(x_0)\mu_{p+1}\sqrt{nh_n}h_n^{p+1} + o(\sqrt{nh_n}h_n^{p+1})$$

Set $\alpha = (\underbrace{0,\cdots,0}_{q},1,0,\cdots,0)^T$. According to Theorem 3.3.1, we have

$$\alpha^T S(\beta)A_{h_n}\hat{\mathbf{u}} + \frac{1}{2g(x_0)}E[\alpha^{\mathbf{T}}\mathbf{W}_{\beta,n}] \xrightarrow{L} N\left(\mathbf{0}, \frac{1}{4g(x_0)}\alpha^T\Sigma(\beta)\alpha\right)$$

If $p = 1$, then we obtain

$$\sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})u_{0k} + \left(\beta_0 + \sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})\right)u_0$$
$$-\left(\beta_0 + \sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})\right)\frac{\sqrt{nh_n}}{2}m^{(2)}(x_0)\mu_2 h_n^2 \xrightarrow{L} N\left(0, \frac{\nu_0}{4g(x_0)}V_\beta\right)$$

$$\sqrt{nh_n}\left\{\sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})(\hat{a}_{0k} - q_{\tau_k}) + \left(\beta_0 + \sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})\right)(u_0 - m(x_0))\right\}$$
$$-\left(\beta_0 + \sum_{k=1}^{q}\frac{1}{2\sigma}\beta_k f(q_{\tau_k})\right)\frac{\sqrt{nh_n}}{2}m^{(2)}(x_0)\mu_2 h_n^2 \xrightarrow{L} N\left(0, \frac{\nu_0}{4g(x_0)}V_\beta\right)$$

Set $\sigma\beta_0 + \sum_{k=1}^{q}\beta_k f(q_{\tau_k})/2 = 1$. Since $\hat{m}_\beta(x_0) = \sum_{k=1}^{q}\beta_k f(q_{\tau_k})\hat{a}_{0k}/2 + \hat{a}_0$, then

$$\sqrt{nh_n}\left(\hat{m}_\beta(x_0) - m(x_0) - \frac{1}{2}m^{(2)}(x_0)\mu_2 h_n^2\right) \xrightarrow{L} N\left(0, \frac{\nu_0\sigma^2}{4g(x_0)}V_\beta\right)$$

and

$$\text{MSE}(\hat{m}_\beta(x_0)) = \frac{1}{4}(m^{(2)}(x_0)\mu_2)^2 h_n^4 + \frac{1}{nh_n}\frac{\nu_0\sigma^2}{4g(x_0)}V_\beta + o_p\left(h_n^4 + \frac{1}{nh_n}\right)$$

This completes the proof of Corollary 3.3.1.

Proof of Corollary 3.3.2:

$$
\begin{aligned}
E[w_{\beta,1}] &= -2\beta_0\sqrt{nh_n}E\left[\frac{1}{h_n}K\left(\frac{x_i-x_0}{h_n}\right)\left(\frac{x_i-x_0}{h_n}\right)r_{ip}\right] \\
&\quad + \sum_{k=1}^{q}\beta_k\sqrt{nh_n}E\left[\frac{1}{h_n}K\left(\frac{x_i-x_0}{h_n}\right)\left(\frac{x_i-x_0}{h_n}\right)\left\{1(\epsilon_i \le \frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma})-\tau_k\right\}\right] \\
&\quad := J_3 + J_4
\end{aligned}
$$

If $p=2$,

$$
\begin{aligned}
J_3 &= -2\beta_0\sqrt{nh_n}\int_{-M}^{M}K(t_i)t_i\left(m(x_0+t_ih_n)-\sum_{j=0}^{2}\frac{m^{(j)}(x_0)}{j!}(t_ih_n)^j\right)g(x_0+t_ih_n)dt_i \\
&= -2\beta_0\sqrt{nh_n}\int_{-M}^{M}K(t_i)t_i\left(\frac{m^{(3)}(x_0)}{6}(t_ih_n)^3+o((t_ih_n)^3)\right)g(x_0+t_ih_n)dt_i \\
&= -\frac{1}{3}\beta_0 g(x_0)m^{(3)}(x_0)\mu_4\sqrt{nh_n}h_n^3+o(\sqrt{nh_n}h_n^3)
\end{aligned}
$$

and condition on $\mathbf{x}$

$$
\begin{aligned}
J_4 &= \sum_{k=1}^{q}\beta_k\sqrt{nh_n}E\left[\frac{1}{h_n}K\left(\frac{x_i-x_0}{h_n}\right)\left(\frac{x_i-x_0}{h_n}\right)\left\{F\left(\frac{\sigma q_{\tau_k}-r_{i,p}}{\sigma}\right)-F(q_{\tau_k})\right\}\right] \\
&= -\sum_{k=1}^{q}\beta_k\sqrt{nh_n}E\left[\frac{1}{h_n}K\left(\frac{x_i-x_0}{h_n}\right)\left(\frac{x_i-x_0}{h_n}\right)\left\{f(q_{\tau_k})\frac{r_{i,p}}{\sigma}+o(r_{i,p})\right\}\right] \\
&= -\frac{1}{6\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})g(x_0)m^{(3)}(x_0)\mu_4\sqrt{nh_n}h_n^3+o(\sqrt{nh_n}h_n^3)
\end{aligned}
$$

Therefore,

$$
E[w_{\beta,1}] = -\frac{1}{3}g(x_0)m^{(3)}(x_0)\mu_4\sqrt{nh_n}h_n^3\left(\beta_0+\frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\right)+o(\sqrt{nh_n}h_n^3)
$$

Set $\alpha = (\underbrace{0,\cdots,0}_{q},0,1,0,\cdots,0)^T$. According to Theorem 3.3.1, we have

$$
\alpha^T S(\beta)A_{h_n}\hat{\mathbf{u}}+\frac{1}{2g(x_0)}E[\alpha^{\mathbf{T}}\mathbf{W}_{\beta,n}] \xrightarrow{L} N\left(\mathbf{0},\frac{1}{4g(x_0)}\alpha^T\Sigma(\beta)\alpha\right)
$$

Then

$$\sqrt{nh_n}\left(\beta_0+\frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\right)\mu_2 h_n(\hat{m}_\beta^{(1)}(x_0)-m^{(1)}(x_0))$$
$$-\frac{1}{6}m^{(3)}(x_0)\mu_4\sqrt{nh_n}h_n^3\left(\beta_0+\frac{1}{2\sigma}\sum_{k=1}^{q}\beta_k f(q_{\tau_k})\right)\xrightarrow{L} N\left(0,\frac{\nu_2}{4g(x_0)}V_\beta\right)$$

Therefore, we have

$$\sqrt{nh_n}\mu_2 h_n\left(\hat{m}_\beta^{(1)}(x_0)-m^{(1)}(x_0)-\frac{m^{(3)}(x_0)\mu_4}{6}h_n^3\right)\xrightarrow{L} N\left(0,\frac{\nu_2\sigma^2}{4g(x_0)}V_\beta\right)$$

i.e.

$$\sqrt{nh_n}\left(\hat{m}_\beta^{(1)}(x_0)-m^{(1)}(x_0)-\frac{m^{(3)}(x_0)\mu_4}{6\mu_2}h_n^2\right)\xrightarrow{L} N\left(0,\frac{\nu_2\sigma^2}{4g(x_0)h_n^2\mu_2^2}V_\beta\right)$$

Consequencely,

$$\text{MSE}(\hat{m}_\beta^{(1)}(x_0))=\left(\frac{m^{(3)}(x_0)}{6}\right)^2\frac{\mu_4^2}{\mu_2^2}h_n^4+\frac{\nu_2\sigma^2}{4g(x_0)\mu_2^2}\frac{V_\beta}{nh_n^3}+o_p\left(h_n^4+\frac{1}{nh_n^3}\right)$$

We can show the similar results for $p=1$. This completes the proof of Corollary 3.3.2.

Proof of Lemma B.1

Since F belongs to the domain of attraction of a stable distribution, then

$$\mu(x)=\int_{-\infty}^{\infty}t^2 I(|t|\le x)dF(t)=E[\epsilon^2 I(|\epsilon|<x)]\sim x^{2-\tau}L(x) \tag{S.7}$$

where $0<\tau\le 2$ and $L(\cdot)$ is a slowly varying function, i.e. $\frac{L(mx)}{L(x)}\to 1$ as $x\to\infty$, for any $m>0$. And also

$$\frac{x^2P(|\epsilon_i|>x)}{\mu(x)}\to\frac{2-\tau}{\tau} \tag{S.8}$$

From Feller (1971, pg 579), we also know

$$\frac{n\mu(a_n)}{a_n^2}=\frac{na_n^{2-\tau}L(a_n)}{a_n^2}\to c \tag{S.9}$$

for some constant c.

Part (a)

Case 1 : $1 < \tau \leq 2$

$1 < \tau \leq 2$ implies $E[|\epsilon_i|] < \infty$. And since the truncated second moment and the tailsum of $|\epsilon_i|$ are the same of that of ϵ_i, by applying **Theorem 3** of Feller (1971, pg 580), we obtain,

$$\frac{\sum_{i=1}^{n} |\epsilon_i| - nE[|\epsilon_i|]}{a_n} \xrightarrow{d} U$$

for some nondegenerated distribution U.

From the fact that $na_n^{-\tau}L(a_n) \to c$, we observe $\frac{\sqrt{n}}{a_n^{(\tau+\delta)/2}} \to 0$, where $\delta = 0$ if $\tau = 2$ and any $\delta > 0$, if $\tau < 2$. Then,

$$\begin{aligned}\frac{\sum_{i=1}^{n} |\epsilon_i|}{a_n\sqrt{n}} &= \frac{\sum_{i=1}^{n} |\epsilon_i| - nE[|\epsilon_i|]}{a_n\sqrt{n}} + \frac{nE[|\epsilon_i|]}{a_n\sqrt{n}} \\ &= \frac{\sum_{i=1}^{n} |\epsilon_i| - nE[|\epsilon_i|]}{a_n\sqrt{n}} + \frac{\sqrt{n}}{a_n^{(\tau+\delta)/2}}\frac{a_n^{(\tau+\delta)/2}}{a_n}E[|\epsilon_i|] \\ &\to 0\end{aligned}$$

Case 2 : $\tau = 1$

As shown by Kaminska (2010),

$$\frac{\sum_{i=1}^{n} |\epsilon_i| - nE[|\epsilon_i|I(|\epsilon_i| \leq a_n)]}{a_n} \xrightarrow{d} U$$

for some stable distribution U.

Since $\tau = 1$, then

$$1 - F(x) + F(-x) = P(|\epsilon_i| > x) \sim x^{-1}L(x)$$

Define $\xi_i = sgn(\epsilon_i)|\epsilon_i|^{\frac{1}{2}+\delta}$, for some $0 < \delta < \frac{1}{2}$, then

$$P(|\xi_i| > x) = P(|\epsilon_i| > x^{\frac{1}{\frac{1}{2}+\delta}}) \sim x^{-\frac{1}{\frac{1}{2}+\delta}}L(x^{\frac{1}{\frac{1}{2}+\delta}})$$

It can be verified that $H_1(x) = L(x^{\frac{1}{\frac{1}{2}+\delta}})$ is a slowing varying function. Since $1 < \frac{1}{\frac{1}{2}+\delta} < 2$, then ξ_i belongs to the domain of attraction of some stable distribution U'.

Hence, we obtain $E[\xi_i^2 I(|\xi_i| < t)] \sim t^{2-\frac{1}{\frac{1}{2}+\delta}} L(t^{\frac{1}{\frac{1}{2}+\delta}})(1+2\delta)$. Then

$$\begin{aligned}
\frac{nE[\xi_i^2 I(|\xi_i| < a_n^{\frac{1}{2}+\delta})]}{a_n^{1+2\delta}} &= \frac{E[\xi_i^2 I(|\xi_i| < a_n^{\frac{1}{2}+\delta})]}{(a_n^{\frac{1}{2}+\delta})^{2-\frac{1}{\frac{1}{2}+\delta}} L((a_n^{\frac{1}{2}+\delta})^{\frac{1}{\frac{1}{2}+\delta}})(1+2\delta)} \frac{n(a_n^{\frac{1}{2}+\delta})^{2-\frac{1}{\frac{1}{2}+\delta}} L((a_n^{\frac{1}{2}+\delta})^{\frac{1}{\frac{1}{2}+\delta}})(1+2\delta)}{a_n^{1+2\delta}} \\
&= na_n^{-1} L(a_n)(1+2\delta) \\
&= (1+2\delta)c
\end{aligned}$$

Since ξ_i is symmetric about 0, by **Theorem 3** of Feller (1971, pg 580), we have

$$\frac{\sum_{i=1}^n \xi_i}{a_n^{\frac{1}{2}+\delta}} \xrightarrow{d} U'$$

.

Then,

$$\begin{aligned}
n\frac{E[|\epsilon_i| I(|\epsilon_i| \le a_n)]}{a_n^{1+2\delta}} &= n\frac{E[|\epsilon_i| I(|\epsilon_i| \le 1)]}{a_n^{1+2\delta}} + n\frac{E[|\epsilon_i| I(1 < |\epsilon_i| \le a_n)]}{a_n^{1+2\delta}} \\
&\le \frac{n}{a_n^{1+2\delta}} + n\frac{E[|\epsilon_i|^{1+2\delta} I(|\epsilon_i|^{\frac{1}{2}+\delta} \le a_n^{\frac{1}{2}+\delta})]}{a_n^{1+2\delta}} \\
&= \frac{n}{a_n^{1+2\delta}} + n\frac{E[\xi_i^2 I(|\xi_i| \le a_n^{\frac{1}{2}+\delta})]}{a_n^{1+2\delta}} \\
&\to (1+2\delta)c
\end{aligned}$$

Pick a small δ, and we obtain

$$\begin{aligned}
\frac{\sum_{i=1}^n |\epsilon_i|}{a_n\sqrt{n}} &= \frac{\sum_{i=1}^n |\epsilon_i| - nE[|\epsilon_i| I(|\epsilon_i| \le a_n)]}{a_n\sqrt{n}} + \frac{nE[|\epsilon_i| I(|\epsilon_i| \le a_n)]}{a_n^{1+2\delta}} \frac{a_n^{1+2\delta}}{a_n\sqrt{n}} \\
&\to 0
\end{aligned}$$

Case 3 : $\tau < 1$

Since $\tau < 1$, applying **Theorem 3** of Feller (1971, pg 580) yields

$$\frac{\sum_{i=1}^{n} |\epsilon_i|}{a_n} \xrightarrow{d} U$$

for some nondegenerated distribution U.

The result is clear.

Part (b)

Case 1 : $\tau = 2$

This implies $E[|\epsilon_i|] > 0$ exists (Gnedenko and Kolmogorov 1968). And because $\mu(x) \to \infty$ as $x \to$, we have $L(x) \to \infty$.

Since $\tau = 2$, then F belongs to the domain of attraction of a normal distribution, that is

$$\frac{\sum_{i=1}^{n} \epsilon_i}{a_n} \xrightarrow{d} N(0,1)$$

Applying **Theorem 1.1** of Gut (2006) yields

$$\frac{\sum_{i=1}^{n} \epsilon_i^2}{a_n^2} \xrightarrow{p} 1$$

Case 2 : $\tau < 2$

S.4 implies $P(|\epsilon_i| > x) \sim x^{-\tau} L(x) \frac{2-\tau}{\tau}$. Since $\tau < 2$, we only need to check the tail behavior of ϵ_i^2.

$$P(\epsilon_i^2 > x) = P(|\epsilon_i| > \sqrt{x}) \sim x^{-\tau/2} L(\sqrt{x}) \frac{2-\tau}{\tau}$$

According to **Corollary 2** of Feller (1971 pg 598), ϵ_i^2 belongs to the domain of attraction of a nondegenerate distribution U. And

$$E[(\epsilon_i^2)^2 I(\epsilon_i^2 < t)] \sim t^{2-\tau/2} L(\sqrt{t})$$

Then

$$\begin{aligned}\frac{nE[(\epsilon_i^2)]I(\epsilon_i^2 < a_n^2)]}{a_n^4} &= \frac{E[(\epsilon_i^2)]I(\epsilon_i^2 < a_n^2)]}{(a_n^2)^{2-\tau/2}L(\sqrt{a_n^2})} \frac{na_n^{4-\tau}L(a_n)}{a_n^4} \\ &\to c\end{aligned}$$

Then applying bfTheorem 3 of Feller (1971, pg 580), we obtain

$$\frac{\sum_{i=1}^n \epsilon_i^2}{a_n^2} \xrightarrow{d} U$$

Part (c)

We show that $\frac{1}{\sqrt{h_n}} K(\frac{X_i - x_0}{h_n})\epsilon_i$ is symmetric about 0. $\forall t > 0$,

$$\begin{aligned}P(\frac{1}{\sqrt{h_n}} K(\frac{X_i - x_0}{h_n})\epsilon_i > t) &= \int_{-\infty}^{\infty} P(\epsilon_i > \frac{t}{\frac{1}{\sqrt{h_n}}K(\frac{x-x_0}{h_n})})dG(x) \\ &= \int_{\infty}^{\infty} P(\epsilon_i < -\frac{t}{\frac{1}{\sqrt{h_n}}K(\frac{X_i-x_0}{h_n})})dG(x) \\ &= P(\frac{1}{\sqrt{h_n}} K(\frac{X_i - x_0}{h_n})\epsilon_i < -t)\end{aligned}$$

which implies $\frac{1}{\sqrt{h_n}} K(\frac{X_i - x_0}{h_n})\epsilon_i$ is symmetric about 0.

For simplicity, we only present the proof for the uniform kernel. The proofs for other kernel functions are in the same fashion, since kernel functions can be considered as properly chosen weights.

Define $m_n = \sum_{i=1}^n I(|X_i - x_0| < h_n)$. We know that, as $n \to \infty$

$$\sqrt{nh_n}(\frac{m_n}{nh_n} - g(x_0)) \to O_p(1)$$

then $m_n = nh_n g(x_0) + \sqrt{nh_n} O_p(1)$ and $m_n \to \infty$ almost surely.

$$\begin{aligned}\frac{\sum_{i=1}^{n} \frac{1}{\sqrt{h_n}} I(\frac{X_i - x_0}{h_n})\epsilon_i}{a_n} &= \frac{\sum_{\{i:|X_i - x_0| < h_n\}} \epsilon_i}{a_{m_n}} \frac{a_{m_n}}{a_n \sqrt{h_n}} \\ &=: H_1 \times H_2\end{aligned}$$

Consider H_1 first. Since $m_n \to \infty$ almost surely, we have

$$\frac{\sum_{\{i:|X_i - x_0| < h_n\}} \epsilon_i}{a_{m_n}} \xrightarrow{d} U, \tag{S.9}$$

as $n \to \infty$.

According to (S.8), we obtain the following $\frac{(\frac{m_n L(a_{m_n})}{c})^{\frac{1}{\tau}}}{a_{m_n}} \to 1$ and $\frac{(\frac{nL(a_n)}{c})^{\frac{1}{\tau}}}{a_n} \to 1$ as $n \to \infty$. Then for H_2,

$$\begin{aligned}\lim_{n\to\infty} H_2 &= \lim_{n\to\infty} \frac{a_{m_n}}{(\frac{m_n L(a_{m_n})}{c})^{\frac{1}{\tau}}} \frac{(\frac{m_n L(a_{m_n})}{c})^{\frac{1}{\tau}}}{(\frac{nL(a_n)}{c})^{\frac{1}{\tau}} \sqrt{h_n}} \frac{(\frac{nL(a_n)}{c})^{\frac{1}{\tau}}}{a_n} \\ &= \lim_{n\to\infty} \frac{[(nh_n g(x_0) + \sqrt{nh_n} O_p(1))L(a_{nh_n g(x_0) + \sqrt{nh_n} O_p(1)})]^{\frac{1}{\tau}}}{(nL(a_n))^{\frac{1}{\tau}} \sqrt{h_n}} \\ &= \lim_{n\to\infty} (g(x_0) + \frac{O_p(1)}{\sqrt{nh_n}})^{\frac{1}{\tau}} (\frac{L(a_{nh_n g(x_0) + \sqrt{nh_n} O_p(1)})}{L(a_n)})^{\frac{1}{\tau}} h_n^{(\frac{1}{\tau} - \frac{1}{2})} \\ &\leq (g(x_0))^{\frac{1}{\tau}} \end{aligned} \tag{S.10}$$

Combine (S.9) and (S.10) together, and we have

$$\frac{\sum_{i=1}^{n} \frac{1}{\sqrt{h_n}} I(\frac{X_i - x_0}{h_n})\epsilon_i}{a_n} \leq O_p(1)$$

as $n \to \infty$.

This completes the proof of the lemma.

Proof of Lemma B.2

Let $\tilde{\zeta}_i$'s be residuals of a $\sqrt{nh_n}$-consistent fit. Then $\tilde{\sigma}^2 = \sum_{i=1}^{n} \frac{\tilde{\epsilon}_i^2}{n}$, and $\tilde{\tau}_{0,k} = \sum_{i=1}^{n} \frac{\tilde{\epsilon}}{\tilde{\sigma}} 1\left(\frac{\tilde{\epsilon}}{\tilde{\sigma}} < \tilde{q}_{\tau_k}\right)$, where $\tilde{q}_{\tau_k}$ is the τ_k's sample quantile of $\frac{\tilde{\epsilon}}{\tilde{\sigma}}$, for $k = 1, \cdots, q$.

Let $\tau_l = \frac{1}{2}$ for some $1 \leq l \leq q$, $\hat{\beta}_{0p} = \frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4f(0)^2+8\tilde{f}(0)\tilde{\tau}_{0,l}+1} 1\left(\frac{1+4\tilde{f}(0)\tilde{\tau}_{0,l}}{4\tilde{f}(0)\tilde{\sigma}^2+8\tilde{f}(0)\tilde{\tau}_{0,l}+1} \geq 0\right)$ and $\hat{\beta}_{lp} = \frac{2(1-\hat{\beta}_{0p})}{\tilde{f}(0)}$. And we have

$$\hat{\beta} = \arg \min_{\beta \geq 0,\ \tilde{\alpha}^T\beta=1} (4\beta_0^2\tilde{\sigma}^2 - 4\beta_0 \sum_{k=1}^{q} \beta_k \tilde{\sigma}\tilde{\tau}_{0,k} + \sum_{k,k'=1}^{q} \beta_k\beta_{k'}\tau_{k,k'})$$

We show that $\hat{\beta}_0 \leq \hat{\beta}_{0p}$ almost surely, when $E[\epsilon_i^2]$ does not exist. Suppose it is not. Consider

$$\begin{aligned}\frac{4\hat{\beta}_0^2\tilde{\sigma}^2 - 4\hat{\beta}_0 \sum_{k=1}^{q} \hat{\beta}_k\tilde{\sigma}\tilde{\tau}_{0,k} + \sum_{k,k'=1}^{q} \hat{\beta}_k\hat{\beta}_{k'}\tau_{k,k'}}{4\hat{\beta}_{0p}^2\tilde{\sigma}^2 - 4\hat{\beta}_{0p}\hat{\beta}_{lp}\tilde{\tau}_{0,l} + \hat{\beta}_{lp}^2\frac{1}{4}} &> \frac{4\hat{\beta}_0^2\tilde{\sigma}^2}{4\hat{\beta}_{0p}^2\tilde{\sigma}^2 - 4\hat{\beta}_{0p}\hat{\beta}_{lp}\tilde{\sigma}\tilde{\tau}_{0,l} + \hat{\beta}_{lp}^2\frac{1}{4}} \\ &= \frac{4\hat{\beta}_0^2}{4\hat{\beta}_{0p}^2 - 4\hat{\beta}_{0p}\hat{\beta}_{lp}\frac{\tilde{\tau}_{0,l}}{\tilde{\sigma}} + \hat{\beta}_{lp}^2\frac{1}{4\tilde{\sigma}^2}}\end{aligned}$$

Since $E[\epsilon_i^2]$ does not exist, then as $n \to \infty$, $\tilde{\sigma}^2 \overset{a.s.}{\to} \infty$ and $\frac{\tilde{\tau}_{0,l}}{\tilde{\sigma}} \overset{a.s.}{\to} 0$. Thus, as $n \to \infty$,

$$\frac{4\hat{\beta}_0^2\tilde{\sigma}^2 - 4\hat{\beta}_0 \sum_{k=1}^{q} \hat{\beta}_k\tilde{\tau}_{0,k} + \sum_{k,k'=1}^{q} \hat{\beta}_k\hat{\beta}_{k'}\tau_{k,k'}}{4\hat{\beta}_{0p}^2\tilde{\sigma}^2 - 4\hat{\beta}_{0p}\hat{\beta}_{lp}\tilde{\tau}_{0,l} + \hat{\beta}_{lp}^2\frac{1}{4}} \overset{a.s}{>} 1$$

However, by the definition of $\hat{\beta}$, we have

$$\frac{4\hat{\beta}_0^2\tilde{\sigma}^2 - 4\hat{\beta}_0 \sum_{k=1}^{q} \hat{\beta}_k\tilde{\sigma}\tilde{\tau}_{0,k} + \sum_{k,k'=1}^{q} \hat{\beta}_k\hat{\beta}_{k'}\tau_{k,k'}}{4\hat{\beta}_{0p}^2\tilde{\sigma}^2 - 4\hat{\beta}_{0p}\hat{\beta}_{lp}\tilde{\sigma}\tilde{\tau}_{0,l} + \hat{\beta}_{lp}^2\frac{1}{4}} \leq 1$$

since $\tilde{\sigma}\hat{\beta}_{0p} + \frac{1}{2}\tilde{f}(0)\hat{\beta}_{lp} = 1$.

Contradiction!

This completes the proof.

Bibliography

[1] Belloni, A. and Chernozhukov, V. (2011), “l_1 penalized quantile regression in high-dimensional sparse models” in *The Annals of Statistics*, **39**, 82-130.

[2] Bickel, P.J. (1973), ”On some analogues to linear combinations of order statistics in the linear model” in *Ann. Statist.*, **1**, 597-616.

[3] Bloomfield, P., and Steiger, W. L. (1983), *Least Absolute Deviation: Theory, Applications and Algorithms*, Boston: Birkhauser.

[4] Bradic, J. Fan, J. and Wang, W. (2011). “Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection,” in *J. R. Statist. Soc. Ser. B*, **73**, Part 2, pp 325-349.

[5] Candes, E. and Tao, T. (2007), “The Dantzig selector: Statistical estimation when p is much larger than n” in *The Annals of Statistics*, **35**, 2313-2351.

[6] Chernozhukov, V. (2005), ”Extremal Quantile Regression,” in *Ann. Statist.*, **33**, 806-839.

[7] Fan, J. and Gijbels, I. (1992), ”Variable bandwidth and local linear regression smoothers” in *Ann. Statist.*, **20**, 2008-2036.

[8] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

[9] Fan, J., Hu, T.C. and Truong, Y.K. (1994), ”Robust non-parametric function estimation” in *Scand. J. Statist*, **21**, 433-446.

[10] Fan, J. and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” in *Journal of the American Statistical Association,* , **96**, 1348-1360.

[11] Fan, J. and Peng, H. (2004), “Nonconcave penalized likelihood with a diverging number of parameters,” in *The Annals of Statistics*, **32**, 928-961

[12] He, X. and Shao, Q. (2000) “On Parameters of Increasing Dimensions” in *Journal of Multivariate Analysis*, **73**, 120-135.

[13] Huang, J., Horowitz, J.L. and Ma, S. (2008a) “Asymptotic properties of bridge estimators in sparse high-dimensional regression models” in *The Annals of Statistics*, **36**, 587-613

[14] Huang, J., Ma, S. and Zhang, C. (2008b) “Adaptive lasso for sparse high-dimensional regression models ” in *Statistica Sinica*, **18**, 1603-1618.

[15] Kai, B., Li, R. and Zou, H.(2010) ”Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression,” in *J. R. Statist. Soc., B,* **71**, 49-69.

[16] Knight, K. (1998), Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.*, **26**, 755-770.

[17] Koenker, R. (1984), ”A note on L-estimates for linear models,” in *Statist. Probab. Lett.*, **2**, 323-325.

[18] Koenker, R. (2005), ”Regression Quantiles,” Cambridge Univ. Press, Cambridge.

[19] Koenker, R., and Bassett, G. (1978), ”Regression Quantiles,” in *Econometrica*, **46**, 33-50.

[20] Koenker, R. and Portnoy, S. (1987), ”L-Estimation for Linear Models”, in *J. Am. Statis. Ass.*, **82**, 851-857.

[21] Portnoy, S. and Koenker, R. (1989), “Adaptive L-Estimation for Linear Models”, in *Ann. Statist.*, **17**, 362-381.

[22] Portnoy, S. and Koenker, R. (1997), “The Gaussian Hare and The Laplacian Tortoise: Computability of square-error versus absolute-error estimators”, in *Statist. Sci.*, **12**, 279-300.

[23] Ruppert, D., Sheather, S. and Wand, M.P. (1995) ”An effective bandwidth selector for local least squares regression,” in *J. Am. Statis. Ass.*, **90**, 1257-1270.

[24] Sun, J., Gai, Y., and Lin, L. (2013) “Weighted local linear composite quantile estimation for the case of general error distributions”. in *J. Statist. Plann. Infer.* (In Press)

[25] Tibshirani, R. (1996). “Regression shrinkage and selection via the Lasso,? in *Journal of the Royal Statistical Society, Ser. B,* **58**, 267-288.

[26] Wahba, G. (1990), ”Spline models for observational data” in *SIAM*, Philadelphia.

[27] Wang, H., and Leng, C. (2007) “Unified lasso estimation via least square approximation” in *Journal of American Statistical Association,* **102**, 1039-1048

[28] Wang, H., Li, G. and Jiang, G. (2007) “Robust regression shrinkage and consistent variable selection via the LAD-Lasso” in *Journal of Business and Economic Statistics* **25**, 347-355.

[29] Wang, H., Li, B. and Leng, C. (2009) “Shrinkage tuning parameter selection with a diverging number of parameters,” in *Journal of the Royal Statistical Society, Ser. B,* **71**, 671-683.

[30] Wang, K., Wu, Y. and Li, R.(2012) “Quantile regression: applications and current research areas,” in *Journal of the Royal Statistical Society, Ser. D,* **52**, 331-350.

[31] Watson, G.S. (1964), ”Smooth regression analysis” in *Sankh. Ser. A*, **26**, 359-372.

[32] Welsh, A.H. (1996), ”Robust estimation of smooth regression and spread functions and their derivatives” in *Statist. Sin.*, **6**, 347-366.

[33] Yu, K. and Jones, M.C. (1998), ”Local linear quantile regression” in *J. Am. Statist. Ass.*, **93**, 228-237.

[34] Yu, K., Liu, Z. and Stander, J.(2003) ”Quantile regression: applications and current research areas,” in *J. R. Statist. Soc., D,* **52**, 331-350.

[35] Zou, H. (2006), “Quantile Regression for Analyzing Heterogeneity in Ultra-High Dimension,” in *Journal of American Statistical Association*, **107**, 214-222.

[36] Zou, H. and Yuan, M. (2008), ”Composite Quantile Regression and The Oracle Model Selection Theory” in *Ann. Statist.*, **36**, 1108-1126.