The Pennsylvania State University

The Graduate School

Eberly College of Science

ROBUST NONPARAMETRIC AND SEMIPARAMETRIC MODELING

A Dissertation in

Statistics

by

Bo Kai

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2009

The dissertation of Bo Kai was reviewed and approved* by the following:

Runze Li
Professor of Statistics
Dissertation Co-Adviser
Chair of Committee

David R. Hunter
Associate Professor of Statistics
Dissertation Co-Adviser

Damla Sentürk
Assistant Professor of Statistics

Vernon M. Chinchilli
Distinguished Professor of Public Health Sciences
Professor of Statistics
Chair of Public Health Sciences

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

In this dissertation, several new statistical procedures in nonparametric and semiparametric models are proposed. The concerns of the research are efficiency, robustness and sparsity.

In Chapter 3, we propose complete composite quantile regression (CQR) procedures for estimating both the regression function and its derivatives in fully nonparametric regression models by using local smoothing techniques. The CQR estimator was recently proposed by Zou and Yuan (2008) for estimating the regression coefficients in the classical linear regression model. The asymptotic theory of the proposed estimator was established. We show that, compared with the classical local linear least squares estimator, the new method can significantly improve the estimation efficiency of the local linear least squares estimator for commonly used non-normal error distributions, and at the same time, the loss in efficiency is at most 8.01% in the worst case scenario.

In Chapter 4, we further consider semiparametric models. The complexity of semiparametric models poses new challenges to parametric inferences and model selection that frequently arise from real applications. We propose new robust inference procedures for the semiparametric varying-coefficient partially linear model. We first study a quantile regression estimate for the nonparametric varying-coefficient functions and the parametric regression coefficients. To improve efficiency, we further develop a composite quantile regression procedure for both parametric and nonparametric components. To achieve sparsity, we develop a variable selection procedure for this model to select significant variables. We study the sampling properties of the resulting quantile regression estimate and composite quantile regression estimate. With proper choices of penalty functions

and regularization parameters, we show the proposed variable selection procedure possesses the oracle property in the terminology of Fan and Li (2001).

In Chapter 5, we propose a novel estimation procedure for varying coefficient models based on local ranks. By allowing the regression coefficients to change with certain covariates, the class of varying coefficient models offers a flexible semiparametric approach to modeling nonlinearity and interactions between covariates. Varying coefficient models are useful nonparametric regression models and have been well studied in the literature. However, the performance of existing procedures can be adversely influenced by outliers. The new procedure provides a highly efficient and robust alternative to the local linear least squares method and can be conveniently implemented using existing R software packages. We study the sample properties of the proposed procedure and establish the asymptotic normality of the resulting estimate. We also derive the asymptotic relative efficiency of the proposed local rank estimate to the local linear estimate for the varying coefficient model. The gain of the local rank regression estimate over the local linear regression estimate can be substantial. We further develop nonparametric inferences for the rank-based method. Monte Carlo simulations are conducted to access the finite sample performance of the proposed estimation procedure. The simulation results are promising and consistent with our theoretical findings.

All the proposed procedures are supported by intensive finite sample simulation studies and most are illustrated with real data examples.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First of all, I would like to express my sincere gratitude to my dissertation advisor, Dr. Runze Li, who has irreplaceable influence on my academic career. Without his insight, patient and inspiring guidance, it would be impossible for me to continue the research work and finish this dissertation. His broad knowledge, brilliant ideas and nice personality make him a great mentor and always encourage me.

I am very grateful to my co-advisor, Dr. David R. Hunter, for his invaluable instruction and discussion, and his time and patience for helping me revise the dissertation. I also would like to thank the committee members, Dr. Damla Sentürk and Dr. Vernon M. Chinchilli, for their helpful commentary on my dissertation work.

My special thanks must go to two collaborators, Dr. Hui Zou and Dr. Lan Wang. I really learned a lot through the discussions we had during the exciting research processes.

Finally, I would like to acknowledge my parents, Mr. Guoxing Kai and Ms. Lihua Zhang, for their extreme supports through my life. And in particular, I should thank my wife, Ms. Jingwen Zhang. Without her truly love, support and understanding, I can not finish this dissertation.

This dissertation research has been supported by National Institute on Drug Abuse grants R21 DA024260 and P50 DA10075, and National Science Foundation grants DMS 0348869.

Chapter 1

Introduction

With the advent of modern technology, computing facilities have been improved dramatically over the last several decades. Researchers have realized that, in many real data applications, parametric models are not good enough to capture the relationship between the response variable and its covariates. Various estimation and inference procedures for nonparametric and semiparametric models have been proposed and studied in the literature.

Most existing procedures are built on either least-squares-type or likelihood-type methods. There are two major disadvantages of these methods. The first one is robustness. It is well known that the least squares or likelihood methods are not resistant to outliers. In the presence of outliers or contamination, these methods result in biased estimates and may lead to misleading conclusions. The other one is efficiency. Although the estimates remain asymptotically normal for a large class of random error distributions, their efficiency can deteriorate dramatically when the true error distribution departs from normality. An extreme case is that the least squares estimate fails to be consistent in the presence of infinite variance errors such as a Cauchy error distribution. These considerations motivate us to develop novel procedures. Therefore, this dissertation aims to develop novel statistical methodology and inference procedures for nonparametric and semiparametric models that are highly efficient, robust and computationally simple.

This dissertation consists of three manuscripts. The work in Chapter 3 is based on Kai, Li, and Zou (2009a), in which we propose the local composite quantile regression (CQR) procedures for estimating regression function and its derivatives in the fully nonparametric regression model. The work in Chapter 4

is based on Kai, Li, and Zou (2009b), in which we propose new estimation and variable selection procedures for the semiparametric varying-coefficient partially linear model. The work in Chapter 5 is based on Wang, Kai, and Li (2009), in which we propose new robust estimation and inference procedures for varying coefficient models based on local rank regression.

1.1 Local composite quantile regression

The composite quantile regression (CQR) estimator was recently proposed by Zou and Yuan (2008) for estimating the regression coefficients in the classical linear regression model. The idea of the CQR is to combine the strength across multiple quantile regressions by forcing a single parameter for "slope" to further improve the efficiency. Zou and Yuan (2008) show that the relative efficiency of the CQR estimator compared to the least squares estimator is greater than 70% regardless the error distribution. Furthermore, the CQR estimator could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator. These nice theoretical properties of CQR in linear regression motivate us to construct the local CQR smoothers as nonparametric estimates of the regression function and its derivatives.

Contributions

We consider the general nonparametric regression model. Our interest is to estimate the conditional regression mean function and its derivatives. For an introduction to nonparametric techniques, see Hastie and Tibshirani (1990), Green and Silverman (1994), Wand and Jones (1995) or Fan and Gijbels (1996), among others.

We make several contributions in this work. We first propose the local linear CQR estimator for estimating the regression function and establish the asymptotic

theory of the proposed estimator. We show that, compared with the classical local linear least squares estimator, the new method can significantly improve the estimation efficiency of the local linear least squares estimator for commonly used non-normal error distributions. We further propose the local quadratic CQR estimator for estimating the derivative of the regression function. The asymptotic theory shows that the local quadratic CQR estimator can often drastically improve the estimation efficiency of its local least squares counterpart if the error distribution is non-normal, and at the same time, the loss in efficiency is at most 8.01% in the worst case scenario. For implementation, we adopt the MM algorithm proposed by Hunter and Lange (2000), which works much faster than linear programming solvers for large data sets. In the end, we establish the general asymptotic theory of the local polynomial CQR estimator. Our theory does not require that the error distribution has a finite variance. Therefore, local CQR estimators can work well even when local polynomial regression fails due to the infinite variance of the noise.

1.2 New robust statistical procedures for semiparametric regression models

As researchers are able to collect massive amounts of data without too much cost, high-dimensional modeling has become one of the most important research topics (Donoho 2000; Fan and Li 2006b). Analysis of high-dimensional data is very challenging, and many efforts have been made to develop modeling procedures for high-dimensional data. In many situations, high dimensional data may contain outliers or violate the normality assumption on the errors. In the presence of outliers or contamination, the ordinary least-squares-based methods or likelihood-based methods may lead to a misleading conclusion. If the error distribution departs from the normal or other assumed distribution, these estimates may not be efficient any more.

For analysis of high-dimensional data, fully nonparametric models are not feasible for implementation and parametric models may be somewhat rigid. Semiparametric regression models can be viewed as a compromise between parametric and fully nonparametric models and retain the advantages of both models. See Härdle, Liang, and Gao (2000), Ruppert, Wand, and Carroll (2003) and Yatchew (2003) for various semiparametric models with estimation and inference procedures. For semiparametric regression models, we can impose variable selection techniques to select features and reduce model complexity.

Variable selection is fundamental to select important features in high dimensional data analysis. Traditional variable selection procedures, such as forward stepwise, backward elimination, and best subset selection procedures, are difficult to implement for high-dimensional data due to the heavy computational burden. Fortunately, there are some modern variable selection procedures developed in the recent literature. Frank and Friedman (1993) proposed the bridge regression via the L_q penalty functions and Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO) via the L_1 penalty to select significant variables. Fan and Li (2001) proposed a unified variable selection framework via nonconcave penalized likelihood. All these methods are distinguished from the traditional variable selection procedures in that the methods select significant variables and estimate their coefficients simultaneously. Numerical algorithms, such as the MM algorithm (Hunter and Li 2005) and the one-step local linear approximation (LLA) (Zou and Li 2008), can be used to select significant features. Thus, the computational cost can be dramatically reduced. This makes feature selection for high-dimensional data feasible.

Contributions

We consider the semiparametric varying-coefficient partially linear model. The complexity of semiparametric models poses new challenges to parametric inferences and model selection that frequently arise from real applications. We propose

new inference procedures for this semiparametric varying-coefficient partially linear model. We make several contributions in this work. We first study quantile regression estimates for the unknown varying-coefficient functions and the unknown regression coefficients. To improve efficiency of quantile regression estimates, we further develop a composite quantile regression procedure for regression functions and regression coefficients. To achieve sparsity, we further develop a variable selection procedure for this model to select significant covariates in the linear part. We study the sampling properties of the resulting quantile regression estimates and composite quantile regression estimates. We derive the asymptotic bias and variance of the resulting estimate, and further establish their asymptotic normality. With proper choices of penalty functions and regularization parameters, we show the proposed variable selection procedure possesses the oracle property in the terminology of Fan and Li (2001). Again, our theory does not require that the error distribution has a finite variance. We further address the computation issues of the proposed variable selection procedures. Extensive Monte Carlo simulation studies are conducted to examine the finite sample performance of the proposed procedures. The results are promising and consistent with our theoretical findings.

1.3 Local rank inference for varying coefficient models

As introduced in Cleveland, Grosse, and Shyu (1992) and Hastie and Tibshirani (1993), the varying coefficient model provides a natural and useful extension of the classical linear regression model by allowing the regression coefficients to depend on certain covariates. Due to its flexibility to explore the dynamic features which may exist in the data and its easy interpretation, the varying coefficient model has been widely applied in many scientific areas. It has also experienced rapid developments in both theory and methodology, such as Fan and Zhang (1999), Kauermann and Tutz (1999), Cai et al. (2000), Brumback and Rice

(1998), Hoover et al. (1998), Wu et al. (1998) and Fan and Zhang (2000), among others. See Fan and Zhang (2008) for a comprehensive survey.

Estimation procedures in the aforementioned papers are built on either local least squares or local likelihood methods, which can be adversely influenced when the true error distribution deviates from normality. Furthermore, these estimators are very sensitive to outliers. Even a few outlying data points may introduce undesirable artificial features in the estimated functions. These considerations motivate us to develop a novel local rank estimation procedure that is highly efficient, robust and computationally simple.

Contributions

We propose new robust estimation and inference procedures for varying coefficient models based on local rank regression. The new procedure provides a highly efficient and robust alternative to the local linear least squares method. Theoretical analysis and numerical simulations both reveal that the gain of the local rank estimator over the local linear least squares estimator, measured by the asymptotic mean squared error or the asymptotic mean integrated squared error, can be substantial. For example, the ARE is 167% for estimating the regression coefficient functions when the random error has a t_3 distribution, is 240% for an exponential random error distribution, and is 493% for a lognormal random error distribution. A striking feature of the local rank procedure is that its pronounced efficiency gain comes with only a little loss when the random error actually has a normal distribution. In the normal error case, the asymptotic relative efficiency for estimating both the coefficient functions and the derivative of the coefficient functions is above 96%. Even in the worst case scenarios, the asymptotic relative efficiency has a lower bound of 88.96% for estimating the coefficient functions, and a lower bound of 89.91% for estimating their derivatives. The new estimator is able to achieve the nonparametric convergence rate even when the local linear

least squares method fails due to infinite random error variance. The new estimator proposed in this dissertation minimizes a convex objective function based on local ranks. The implementation of the minimization can be conveniently carried out using existing functions in the R statistical software package via a simple algorithm. The objective function has the form of a generalized U-statistic whose kernel varies with the sample size. We establish the large sample theory of the proposed procedure by utilizing results from generalized U-statistics, whose kernel function may depend on the sample size. We also extend a resampling approach, which perturbs the objective function repeatedly, to the generalized U-statistics setting; and demonstrate that it can accurately estimate the asymptotic covariance matrix.

1.4 Organization of the dissertation

This dissertation is organized as follows. In Chapter 2, we provide the literature review for this dissertation research. Chapter 3 focuses on the local CQR smoothing techniques in a nonparametric model setting. In Chapter 4, we study the robust estimation and variable selection procedures for the semiparametric varying-coefficient partially linear models. Local rank inference procedures are presented in Chapter 5. Finally, concluding remarks and future research directions are discussed in Chapter 6.

Chapter 2

Literature Review

This chapter provides a brief literature review of the dissertation research. This dissertation uses research findings from three topics: first, nonparametric smoothing techniques; second, robust regression; and third, traditional and modern variable selection methods for regression models. All of the three areas are classical but active topics in statistics.

2.1 Nonparametric smoothing techniques

Nonparametric regression is a form of regression analysis that relaxes the structures we assume on the form of a regression function and uses a flexible one instead. Nonparametric regression provides us a powerful tool to explore the data with unknown structure. There are many specific methods of nonparametric smoothing, such as kernel smoothing, local polynomial smoothing, spline smoothing, wavelets based methods, etc. Most of them assume certain smoothness of the regression function. In this section, we will briefly review the kernel smoothing and local polynomial smoothing techniques, which are frequently used in our research.

2.1.1 Kernel smoothing

Kernel smoothing provides a simple way of finding structure in data without imposing a parametric model. Suppose the bivariate sample $\{(x_i, y_i),\ i = 1, \cdots, n\}$ is collected from the model:

$$y = m(x) + \epsilon, \tag{2.1}$$

where ϵ is random error with $E(\epsilon|X) = 0$ and $Var(\epsilon|X = x) = \sigma^2(x)$. The mean function $m(\cdot)$ is the object to be estimated in a nonparametric regression problem. The shape of $m(\cdot)$ describes the underlying relationship between the response variable Y and the predictor variable X. Usually a point closer to x has more information about the value of $m(x)$, so a natural idea for estimating $m(x)$ is to use the running local average.

The Nadaraya-Watson (NW) (Nadaraya 1964; Watson 1964) kernel regression estimator is defined by

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K[(X_i - x)/h]Y_i}{\sum_{i=1}^n K[(X_i - x)/h]}, \tag{2.2}$$

where $K(\cdot)$ is a function usually satisfying $\int K(x)dx = 1$, which is called the kernel function, and h is a positive number, which is called the bandwidth or window width or smoothing parameter. We may introduce a rescaling notation $K_h(u) = K(u/h)/h$. Then, the NW kernel regression estimator can be rewritten as

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n K_h(X_i - x)Y_i}{\sum_{i=1}^n K_h(X_i - x)}. \tag{2.3}$$

By taking the kernel function to be the uniform kernel

$$K(u) = I(|u| < 1/2), \tag{2.4}$$

the NW estimator becomes the running local average, which is similar to the K-nearest neighbor (KNN) estimator.

If we treat the kernel function in (2.3) as a kind of weight function $w_i(x)$, then $\hat{m}_h(x)$ can be viewed as a weighted average, that is,

$$\hat{m}_h(x) = \frac{\sum_{i=1}^n w_i(x)Y_i}{\sum_{i=1}^n w_i(x)}. \tag{2.5}$$

It is the solution of minimizing the locally weighted least squares, that is,

$$\hat{m}_h(x) = \arg\min_{m_h(x)} \sum_{i=1}^{n} \{Y_i - m_h(x)\}^2 w_i(x). \tag{2.6}$$

We can obtain different kernel estimators by adjusting the weight function $w_i(x)$.

Because the denominator in (2.3) is a random variable, it is not convenient to derive its asymptotic properties. So Gasser and Müller (1984) introduced another kernel regression estimator, which is called the Gasser-Müller kernel estimator and given by

$$\hat{m}_h(x) = \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} K_h(X_i - x) du Y_i, \tag{2.7}$$

where $s_i = (X_{(i)} + X_{(i+1)})/2$, $X_{(i)}$ is the i^{th} order statistics of X, $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$. Note that the sum of the weights in (2.7) is one and hence there is no denominator. See Müller (1988) for a detailed discussion of the GM estimator. A basic comparison of asymptotic properties, including the GM estimator, will be given later.

In kernel smoothing, the kernel function K is usually chosen to be a unimodal probability density function that is symmetric about zero. Sometimes we also use kernels that are not densities. It is interesting that the choice of the shape of the kernel function is not a particularly important issue (Marron and Nolan 1988). However, the choice of the smoothing parameter (bandwidth) is critical. It will directly influence the performance of the estimator. If h is chosen to be too small, then we will overfit the data and get an undersmoothed estimate. This estimate pays too much attention to the data in the local neighborhood. On the contrary, if h is chosen to be too large, then we will underfit the data and get an oversmoothed estimate. This estimate will miss some fine features of the data.

The optimal choice of the bandwidth h should be guided by some criteria for performance of the estimator. One choice is the widely used mean squared error (MSE) criterion. It measures the "distance" between a parameter θ and its

estimator $\hat{\theta}$ by

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2. \tag{2.8}$$

The advantage of using MSE is that it can simply decomposed into the summation of variance and the squared bias

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2. \tag{2.9}$$

A good choice of bandwidth h should give the estimator an optimal balance between variance and bias.

Applying the MSE criterion to the kernel regression estimator, we will notice that $MSE\{\hat{m}_h(x)\}$ depends on x. This means that it can only measure the performance of $\hat{m}_h(x)$ at a fixed point x. Therefore, it will be more appropriate to analyze the integration of $MSE\{\hat{m}_h(x)\}$ over the entire real line. Such a criterion is called the mean integrated squared error (MISE):

$$MISE\{\hat{m}_h(\cdot)\} = \int MSE\{\hat{m}_h(x)\}dx. \tag{2.10}$$

The optimal choice of bandwidth h is the one that minimizes the MISE.

Because the exact MSE or MISE depends on the bandwidth in a complicated way, we may calculate the asymptotic MSE or MISE instead and get the asymptotic MISE-optimal bandwidth h.

2.1.2 Local polynomial smoothing

In the previous section we have indicated that both the Nadaraya-Watson and the Gasser-Müller estimators are local constant fits, which means that we locally approximate the mean function $m(\cdot)$ by a constant θ. This idea can be naturally extended to local polynomial fit, in which we approximate the mean function $m(\cdot)$ by a polynomial rather than a constant. This method was first

proposed by Stone (1977) and Cleveland (1979). And then Stone (1980, 1982) , Fan (1992, 1993) and Ruppert and Wand (1994) studied it systematically. Fan and Gijbels (1996) is a very useful reference book on local polynomial regression.

Suppose that $\{(X_i, Y_i),\ i = 1, \cdots, n\}$ is a random sample from

$$Y = m(X) + \epsilon, \tag{2.11}$$

where ϵ is random error with $E(\epsilon|X) = 0$ and $Var(\epsilon|X = x) = \sigma^2(x)$.

Assume that the mean function $m(x)$ is smooth and its $(p+1)^{th}$ derivative at point x_0 exists. Then we can locally approximate $m(x)$ by the Taylor expansion at point x_0 as

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \cdots + \frac{m^{(p)}(X_0)}{p!}(x - x_0)^p. \tag{2.12}$$

This suggests that we consider a locally weighted polynomial regression problem, that is, minimizing

$$\sum_{i=1}^{n}\{Y_i - \sum_{j=0}^{p}\beta_j(X_i - x_0)^j)\}^2 K_h(X_i - x_0), \tag{2.13}$$

where $K_h(\cdot)$ is the rescaled kernel function defined in the last section and h is a bandwidth. Denote by $\hat{\beta}_j\ (j = 0, \cdots, p)$ the solution to the weighted least squares problem (2.13). It is easy to see from (2.12) that

$$\hat{m}_\nu(x_0) = \nu!\hat{\beta}_\nu, \tag{2.14}$$

which is an estimator for $m^{(\nu)}(x_0)$, the ν^{th} derivative of $m(x)$ evaluated at point x_0.

Let us use matrix notation for convenience. Denote $\mathbf{X}$ as the design matrix in the problem (2.13),

$$\mathbf{X}_{n\times(p+1)} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

and denote

$$\mathbf{y}_{n\times 1} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\beta}_{(p+1)\times 1} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Further, let $\mathbf{W}$ be the $n \times n$ diagonal weight matrix

$$\mathbf{W} = \text{diag}\{K_h(X_i - x_0)\}.$$

Then the weighted least squares problem (2.13) could be rewritten as

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{2.15}$$

The solution to (2.15) is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}, \tag{2.16}$$

which is the desired local polynomial estimator. When the order $p = 1$, we usually call it the local linear estimator.

By using local polynomial regression, we can estimate not only the mean function $m(\cdot)$, but also the first p^{th} derivatives of $m(\cdot)$. This is an advantage of local polynomial regression compared to the NW and GM kernel estimators. Furthermore, local polynomial estimators have better asymptotic properties than NW and GM estimators. The basic asymptotic properties of these three nonparametric

estimators have been summarized in Table (2.1) (adopted from Fan and Gijbels (1996), page 17)

Table 2.1. Pointwise asymptotic bias and variance comparison among nonparametric regression estimators

Method	Bias	Variance
NW estimator	$\{m''(x) + \frac{2m'(x)f'(x)}{f(x)}\}b_n$	V_n
GM estimator	$m''(x)b_n$	$1.5V_n$
Local linear estimator	$m''(x)b_n$	V_n

Here, $b_n = \frac{1}{2}\int_{-\infty}^{+\infty} u^2 K(u)du h^2$ and $V_n = \frac{\sigma^2(x)}{f(x)nh}\int_{-\infty}^{+\infty} K^2(u)du$.

2.2 Robust regression

It is well known that ordinary least squares estimation (OLS) for regression is sensitive to outliers or deviation from model assumptions. Instead of OLS, we should consider robust regression if there are strong suspicions of heteroscedasticity or presence of outliers in the data. Outliers can be generated by simple operational mistakes or including a small portion of sample from a different population. The presence of outliers may have a serious effect on statistical inference.

A popular alternative estimating method in a regression model that is less sensitive to outliers is to use least absolute deviation (LAD) regression. The LAD estimator is defined by minimizing the sum of the absolute values of the residuals.

The primary purpose of robust analysis is to provide methods that are competitive with classical methods but are not seriously affected by outliers or other small departures from model assumptions. As described in Huber (1981,

page 5), a good robust statistical procedure should possess the following desirable features:

1. It should have a reasonably good (optimal or nearly optimal) efficiency at the assumed model.

2. It should be robust in the sense that small deviation from the model assumptions should impair the performance only slightly, that is, the latter (described, say, in terms of the asymptotic variance of an estimate, or of the level and power of a test) should be close to the nominal value calculated at the model.

3. Somewhat larger deviations from the model should not cause a catastrophe.

Good reference books on robust statistics include those by Huber (1981), Hampel et al. (1986) and Rousseeuw and Leroy (1987).

2.2.1 Measures of robustness

In order to quantify the robustness of a method, it is necessary to define some measures of robust technique performance in a theoretical sense. The most common of these are the relative efficiency, the breakdown point and the influence function, which will be described below.

Relative efficiency can tell us how well a robust procedure performs relative to the least squares one on data from a certain distribution. High relative efficiency is desirable in estimation. A simple example (Huber 1981, page 2) shows that when the underlying distribution is a mixture of $N(\mu, \sigma^2)$ and $N(\mu, 9\sigma^2)$ with proportions $1 - \epsilon$ and ϵ, the mean absolute deviation (MAD) has larger asymptotic relative efficiency (ARE) than the mean square error (MSE) for all ϵ between 0.002 and 0.5. It means that only 2 outliers in a sample of size 1000 suffice to neutralize the advantage of the MSE.

The breakdown point is defined as the minimum fraction of outliers which may produce an infinite bias (Hampel 1975). For example, the sample mean has a breakdown point of 0 because we can make the sample mean arbitrarily large just by changing any single observation. However, the sample median has a breakdown point of 0.5 because moving half of the data to infinity will not change the estimator. The higher the breakdown point of an estimator, the higher the robustness. A breakdown point will never exceed 0.5 because if more than half of the data are contaminated, we cannot distinguish between the underlying distribution and the contaminating distribution. For more details, see Huber (1981) and Maronna et al. (2006).

The influence function was first introduced by Hampel (1968, 1974). It is a popular tool to describe the infinitesimal stability of estimators. For a real-valued statistic $T(F)$ at a fixed distribution F, Hampel considered a mixture of two distributions F and $\triangle_x$ (the probability measure which puts mass 1 at the point x), with the form of $(1-t)F + t\triangle_x$. The influence function of T at F is defined to be

$$IF(x; T, F) = \lim_{t \to 0^+} \frac{T((1-t)F + t\triangle_x) - T(F)}{t}. \tag{2.17}$$

It describes the effect of an infinitesimal contamination at the point x on the estimator T, standardized by the mass t of the contamination.

2.2.2 Huber's M-estimator

Robust regression estimators were first introduced by Huber (1973, 1981) and they are well known as M-type (Maximum likelihood type) estimators. There are three major types of estimators. Besides M-type estimators, the other two are R-type (Rank tests based type) and L-type (Linear combination of order statistics)

estimators. However, M-type estimators are the most popular one because of their generality, high breakdown point, and their efficiency (see Huber 1981).

M-estimators are a kind of generalization of maximum likelihood estimators (MLEs). We know that an MLE maximizes $\prod_{i=1}^{n} f(\theta; x_i)$ or, equivalently, minimizes $\sum_{i=1}^{n} -\log f(\theta; x_i)$. Huber proposed to generalize this to the minimization of $\sum_{i=1}^{n} \rho(\theta; x_i)$, where ρ is a function with certain properties. Thus, MLEs are a special case of M-estimators with $\rho = -\log f$.

In a linear regression context, the M-estimator is defined by

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{2.18}$$

If ρ is differentiable, minimizing $\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$ is equivalent to solving

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = 0, \tag{2.19}$$

where $\psi(x) = \dfrac{d\rho(x)}{dx}$. This can be done based on the following argument. Define the weight matrix $W = \text{diag}(w_i)$ with $w_i = \dfrac{\psi(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}{(y_i - \mathbf{x}_i^T \boldsymbol{\beta})}$, then (2.19) can be written as

$$\sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i = 0. \tag{2.20}$$

The above equations can be combined into the following single matrix equation:

$$X^T W X \boldsymbol{\beta} = X^T W \mathbf{y}. \tag{2.21}$$

Therefore, the estimator is

$$\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W \mathbf{y}. \tag{2.22}$$

In practice, the weight matrix W involves $\boldsymbol{\beta}$ and is unknown. So we should use an iterative algorithm to solve this problem, that is, use the estimator of $\boldsymbol{\beta}$ in the last iteration to calculate W and then use it to obtain the estimator of $\boldsymbol{\beta}$ in the current iteration. The algorithm stops when the estimator converges. This is the so-called iteratively reweighted least-squares (IRLS) algorithm.

In the 1980s, several alternatives to M-estimators were proposed. Rousseeuw (1984) introduced the least median of squares (LMS) and the least trimmed squares (LTS) estimators. These estimators minimize the median and the trimmed mean of the squared residuals, respectively. They are very high-breakdown-point estimators. However, both of these methods are inefficient, producing parameter estimates with high variability. Moreover, computing any of these estimators exactly is impractical except for small data sets. They are based on resampling techniques and their solutions are determined randomly (Rousseeuw and Leroy 1987), and then they can even be inconsistent. Another proposed solution was S-estimation (Rousseeuw 1984). This method finds a line that minimizes a robust estimate of the scale of the residuals, which is highly resistant to leverage points, and is robust to outliers in the response. But unfortunately, this method was also found to be inefficient.

2.2.3 Quantile regression

Quantile regression is a special type of M-type regression. It is well known that ordinary least squares regression (OLS) estimates the conditional mean function. And least absolute deviation regression (LAD) estimates the conditional

median function. In the seminal paper of Koenker and Bassett (1978), they generalized the idea of LAD and proposed quantile regression (QR), which estimates the conditional quantile function of the response. By using quantile regression, one can easily study the whole percentile path of the conditional distribution of a response variable. So over the last three decades, quantile regression has been widely used in many different fields, such as economics (Koenker and Hallock 2001), survival analysis (Koenker and Geling 2001) and others.

The ρ function for quantile regression is the 'check' function, which is given by

$$\rho_{\tau}(r) = \begin{cases} \tau r & \text{if } r > 0, \\ -(1-\tau)r & \text{otherwise,} \end{cases} \tag{2.23}$$

where $0 < \tau < 1$.

Consider the sample $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ of size n from the linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{2.24}$$

where $P(\epsilon_i < 0) = F(0) = \tau$ and ϵ_i are independent. Then quantile regression estimates $\boldsymbol{\beta}$ by solving the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{\tau}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{2.25}$$

It is well known that under mild regularity conditions (Koenker 2005), the quantile regression estimates have asymptotic normality, that is,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathscr{D}} N\left(0, \frac{\tau(1-\tau)}{f^2(0)} \Sigma^{-1}\right), \tag{2.26}$$

where $\Sigma = \lim_{n \to \infty} \frac{1}{n} X^T X$.

2.3 Variable selection for regression models

Variable selection plays very important roles in statistical learning, especially in high-dimensional cases. At the initial stage of statistical modeling, we may include a large number of prediction variables to reduce possible model biases because we do not know which among them will have an effect on the response variable. However, many of them may have little effect on the response. Therefore, a major task is to find a parsimonious model, which is a model with as few predictors as possible while still achieving a good fit. Typically, parsimonious models are desirable because they will significantly improve the prediction accuracy of the fitted model. Even when we are not sure about the complexity of the true underlying model, selecting significant variables can also improve the interpretability of a model and speed up the learning process.

Suppose that our dataset contains n observations $\{(\mathbf{x}_i, y_i),\ i = 1, \cdots n\}$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{id})^T$ are d prediction variables for the i^{th} observation. Consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots n, \tag{2.27}$$

where ϵ_i are independent and identically distributed (i.i.d.) random errors with mean zero.

If we denote $\mathbf{y} = (y_1, \cdots, y_n)^T$, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$, then the model above can be expressed in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.28}$$

Here, $\mathbf{X}$ is usually called the design matrix for the regression problem.

2.3.1 Classical variable selection criteria

There are various variable selection criteria in the literature. Detailed reviews can be found in Breiman (1996), Shao (1997) and Miller (2002).

A selection criterion is a statistic calculated from the fitted model. In least squares settings, most of them are built on the residual sum of squares (RSS), which is defined by

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2, \tag{2.29}$$

where $\hat{y}_i$ is the predicted value for the i^{th} observation and $\hat{\mathbf{y}} = (\hat{y_1}, \cdots, \hat{y_n})^T$. Denote RSS_p to be the residual sum of squares when there are p $(0 \leq p \leq d)$ predictors in the model.

Based on different statistical perspectives, these selection criteria could be broadly divided into three classes, namely,

- Prediction criteria
- Information (or likelihood) criteria
- Bayesian criteria (maximizing Bayesian posterior probabilities)

We will mainly focus on the first two classes now.

Prediction sum of squares (PRESS) is a prediction-based criterion proposed by Allen (1974). For a given subset of p predictors, each observation is predicted in turn from the model fitted by the other $n-1$ observations. Let $\hat{y}_{ip}$ be the predicted value for y_i. Then the PRESS statistic for a particular subset of p predictors is defined as

$$PRESS_p = \sum_{i=1}^{n} (y_i - \hat{y}_{ip})^2. \tag{2.30}$$

In calculating (2.30), a different set of regression coefficients is calculated for each case with the same subset of p predictors. So this procedure involves a large amount of computation.

In theory, it can be shown that when n is much larger than p, the PRESS statistic has an asymptotic approximation

$$PRESS_p \approx RSS_p \frac{n^2}{(n-p)^2}. \tag{2.31}$$

The PRESS statistic is closely related to the cross-validation (CV) approach. The idea of cross-validation is that we set a small part of set-aside data and then use the model fitted from the remainder to predict the set-aside data. This is done repeatedly by setting aside a different part of the data until all the observations have been set aside. If we set aside one observation each time, it is called leave-one-out cross-validation, which is exactly PRESS. If we equally divide the whole dataset into K parts and leave out one part at a time, it is called K-fold CV. Usually K is chosen to be 5 or 10. In these cases, the computation costs are much cheaper compared to leave-one-out cross-validation, especially when the sample size n is large. These cross-validation approaches provide us a good way to estimate the prediction error of models, which will be introduced in the next subsection.

Craven and Wahba (1979) proposed the generalized cross-validation statistic, which is defined by

$$GCV = \frac{\frac{1}{n}\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{1 - (df/n)^2}, \tag{2.32}$$

where $\hat{\mathbf{y}}$ is a linear estimator in terms of $\mathbf{y}$, that is, there exists a matrix A such that $\hat{\mathbf{y}} = A\mathbf{y}$. The df in (2.32) is defined to be trace(A). When $\hat{\mathbf{y}}$ is the least squares estimator with p predictors, it is easy to see that $df = p$. Now nGCV is equal to $\dfrac{RSS_p}{1 - (p/n)^2}$, which is asymptotically equal to the PRESS statistic.

There is another well-known prediction-based criterion named Mallows' C_p (Mallows 1973). It is defined as

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p). \tag{2.33}$$

In practice, we use the unbiased estimate

$$\hat{\sigma}^2 = \frac{RSS_d}{n-d}, \tag{2.34}$$

for the full model to substitute for σ^2 in (2.33).

Among the information criteria, the most famous two could be the Akaike's information criterion (AIC, Akaike 1973, 1974) and the Bayesian information criterion (BIC, Schwarz 1978).

The AIC was developed by considering the Kullback-Leibler distance of a model from the true likelihood function. It is defined to be

$$AIC = RSS_p + 2p\sigma^2. \tag{2.35}$$

Thus, the AIC is equivalent to Mallows's C_p in least squares settings.

The BIC is defined as

$$BIC = RSS_p + \log(n)p\sigma^2. \tag{2.36}$$

It has been shown that the BIC is a consistent criterion in the sense that if the true model exists and contains only finitely many parameters, the BIC can determine the true model as the sample size goes to infinity. On the contrary, the AIC tends to overfit the model.

Many other classical variable selection criteria are of the form:

$$RSS_p + cp\sigma^2. \tag{2.37}$$

where c is a regularization parameter. For example, the ψ-criterion (Hannan and Quinn 1979) is

$$\psi_p = RSS_p + c\log(\log(n))p\sigma^2$$

for some constant c, and the risk inflation criterion (RIC, Foster and George 1994) is

$$RIC_p = RSS_p + 2\log(d)p\sigma^2.$$

2.3.2 Prediction and model error

In regression analysis, prediction accuracy usually serves as the "gold standard", namely, the model with higher prediction accuracy is better. Prediction and model error are such kinds of measures of prediction accuracy for models.

The prediction error is defined as the average error in predicting y from $\mathbf{x}$ for future cases, which are not used in fitting the regression equation. We know that the design matrix $\mathbf{X}$ could either be random or controlled. In the $\mathbf{X}$-controlled situation, the design matrix $\{\mathbf{x}_i\}$ are selected by the experimenter and only y is random. In the $\mathbf{X}$-random situation, both y and $\mathbf{X}$ are randomly selected. The definitions of prediction error are a little different for these two situations.

In the $\mathbf{X}$-controlled situation, future data are assumed gathered using the same $\{\mathbf{x}_i,\ i = 1, \cdots n\}$ as in the sample data in hand. So they have the form $\{(y_i^{new}, \mathbf{x}_i),\ i = 1, \cdots n\}$. Let $\hat{\mu}(\mathbf{x})$ be the fitted regression equation. Then the prediction error is defined as

$$\mathrm{PE}(\hat{\mu}) = E\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{new} - \hat{\mu}(\mathbf{x}_i))^2\right]. \tag{2.38}$$

Note that $y_i = \mu(\mathbf{x}_i) + \epsilon_i$ and $\{\epsilon_i\}$ are iid with mean zero and variance σ^2, so

$$\mathrm{PE}(\hat{\mu}) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}(\mu(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i))^2. \tag{2.39}$$

The first component of (2.39) is due to the noise. The second component is due to lack of fit to an underlying model, which is called model error and denoted by

$\mathrm{ME}(\hat{\mu})$. If $\mu(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$, then

$$\mathrm{ME}(\hat{\mu}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\frac{1}{n}\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{2.40}$$

In the **X**-random situation, it is assumed that the sample is from a parent distribution $V = (y, \mathbf{x})$. If $\hat{\mu}(\mathbf{x})$ is the fitted regression equation, the prediction error is defined as

$$\mathrm{PE}(\hat{\mu}) = E\left[y - \hat{\mu}(\mathbf{x})\right]^2. \tag{2.41}$$

We can further decompose PE as

$$\begin{aligned} \mathrm{PE}(\hat{\mu}) &= E\left[y - E(y|\mathbf{x})\right]^2 + E\left[E(y|\mathbf{x}) - \hat{\mu}(\mathbf{x})\right]^2 \\ &= \sigma^2 + E[\mu(x) - \hat{\mu}(\mathbf{x})]^2. \end{aligned} \tag{2.42}$$

The first component of (2.42) is due to the noise. The second component is due to lack of fit to an underlying model, which is the model error in the **X**-random situation and is also denoted by $\mathrm{ME}(\hat{\mu})$. In the linear regression model, $\mu(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$,

$$\mathrm{ME}(\hat{\mu}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\mathbf{x}\mathbf{x}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{2.43}$$

In this dissertation, we will focus on the **X**-random situation only. All the results could be extended to the **X**-controlled situation without essential difficulties.

2.3.3 Variable selection via penalized likelihood

Classical stepwise subset selection methods are widely used in practice, but actually they suffer from several drawbacks. First, their theoretical properties are hard to understand because they ignore stochastic errors in the variable selection process. Second, best subset selection may become infeasible for high-dimensional data due to the expensive computational cost. Third, as analyzed in Breiman

(1996), subset selection methods lack stability in the sense that a small change in the data could lead to a large change in the fitted equation. To overcome these deficiencies, modern penalized likelihood estimation methods were introduced gradually beginning in the 1990s. By adding a continuous penalty term to the likelihood and then maximizing the penalized likelihood, we can select variables and obtain estimates simultaneously. This enables us to study theoretical properties and make statistical inferences on the model.

The penalized least squares method is a special case of penalized likelihood, in which our aim is to minimize the least squares with some penalty. The well-known ridge regression (RR) method (Hoerl and Kennard 1970) is just a solution of penalized least squares. The penalty term used in ridge regression is the L_2 penalty, namely $p_\lambda(|\theta|) = \frac{\lambda}{2}|\theta|^2$. Ridge regression shrinks coefficients but does not select variables because it does not force coefficients to zero. So actually it is not a proper method for variable selection. Frank and Friedman (1993) proposed bridge regression via the L_q penalty functions, namely, $p_\lambda(|\theta|) = \frac{\lambda}{q}|\theta|^q$. And Tibshirani (1996) proposed the Least Absolute Shrinkage and Selection Operator (LASSO) via the L_1 penalty to select significant variables. More recently, Fan and Li (2001) proposed a unified approach via nonconcave penalized likelihood and first introduced the oracle property. They showed that the nonconcave penalized likelihood estimators may perform as well as the oracle estimator in variable selection, that is, they work as well as if we knew the true underlying submodel in advance. About the choice of penalty functions, they pointed out that a good penalty function should result in an estimator with three nice properties:

1. Unbiasedness: The penalized estimator should be unbiased or nearly unbiased when the true parameter is large.

2. Sparsity: The penalized estimator should be a thresholding rule and set small estimates to zero.

3. Continuity: The penalized estimator should be a continuous function in the data, that is, a small change in the data will not result in a large change in the estimates.

And they introduced a family of penalty functions that satisfy all three properties above. The smoothly clipped absolute deviation (SCAD) penalty function is a representative among them with a simple form but good performance.

Fan and Li (2001) also mentioned that there are close connections between classical stepwise subset selection and penalized least squares methods. The classical stepwise selection methods may be viewed as special cases of penalized least squares with the so-called L_0 penalty, which is zero at point 0 and a positive constant everywhere else. Furthermore, when the design matrix is orthonormal, the penalized least squares estimators with the hard thresholding penalty function (defined in 2.47) and a proper tuning parameter λ are equivalent to ones obtained by best subsets selection.

The penalized least squares function is defined to be

$$\frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_{j=1}^{d}p_{\lambda j}(|\beta_j|). \tag{2.44}$$

Note that the penalty functions $p_{\lambda j}(\cdot)$ in (2.44) are not necessarily the same for all j. For the sake of simplicity, we assume that the penalty functions are the same for all coefficients and denote it by $p_{\lambda}(|\cdot|)$.

To see clearly the variable selection effect for penalized least squares, we first assume the columns in design matrix $\mathbf{X}/\sqrt{n}$ to be orthonormal, i.e. $\mathbf{X}^T\mathbf{X}=nI_{p\times p}$.

Then we have

$$\begin{aligned} & \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_{j=1}^{d}p_\lambda(|\beta_j|) \\ = & \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2+\frac{1}{2}\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|^2+\sum_{j=1}^{d}p_\lambda(|\beta_j|) \\ = & \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}\|^2+\sum_{j=1}^{d}\left[\frac{1}{2}(\hat{\beta}_j-\beta_j)^2+p_\lambda(|\beta_j|)\right]. \end{aligned} \tag{2.45}$$

The first term in (2.45) does not involve $\boldsymbol{\beta}$. So the minimization problem (2.45) is equivalent to minimizing the second term componentwise. Thus we only need to consider the following equivalent minimization problem

$$\min_\theta\left\{\frac{1}{2}(z-\theta)^2+p_\lambda(|\theta|)\right\}. \tag{2.46}$$

Fan and Li (2001) thoroughly studied the conditions for penalties satisfying the three properties in the orthonormal case. Figure 2.2 clearly displays the solution to (2.46) for four different choices of penalty functions.

The choice of penalty function will directly influence the properties and performance of the resulting estimates. Let's examine in detail the typical and newly proposed choices of penalty functions and their individual properties.

- L_2 penalty:

$$p_\lambda(|\theta|)=\frac{\lambda}{2}|\theta|^2.$$

  The L_2 penalty leads to ridge regression directly (Frank and Friedman 1993; Fu 1998). The solution to (2.46) with the L_2 penalty is $\hat{\theta}=z/(\lambda+1)$. Ridge regression shrinks the coefficient but does not select variables (sparsity does not hold). However, ridge regression is still important in statistical history

because it brought about the idea of shrinkage. Shrinkage via ridge regression provides regularization and stabilization and a favorable bias-variance tradeoff for the estimates (Breiman 1996).

- L_0 penalty:

$$p_\lambda(|\theta|) = \frac{1}{2}\lambda^2 I(|\theta| \neq 0).$$

The L_0 penalty is special because it is not even continuous. It is easy to see that if we adjust the value of λ, the L_0 penalty will lead to a best subset selection with AIC, BIC, etc. So we can see that traditional subset selection methods are actually a special type of penalized least squares with the L_0 penalty. The L_0 penalty is also called the entropy penalty in the literature.

- Hard thresholding penalty:

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda). \tag{2.47}$$

The hard thresholding penalty was introduced in the discussion of Fan (1997) and Antoniadis (1997). The solution to (2.46) with this penalty is a hard thresholding rule

$$\hat{\theta} = zI(|z| > \lambda), \tag{2.48}$$

which coincides with best subset selection for orthonormal designs, but this penalty is a smoother penalty than the entropy penalty, which also results in (2.48).

- L_q penalty:

$$p_\lambda(|\theta|) = \frac{\lambda}{q}|\theta|^q.$$

The L_q penalty will lead to the bridge regression introduced by Frank and Friedman (1993). The solution to (2.46) is sparse only when $q \leq 1$ and is continuous only when $q \geq 1$. So the L_1 penalty is the only one with both

sparsity and continuity in this family. For detailed discussion about the L_q penalty, readers are referred to Fu (1998) and Fan and Li (2006b).

- L_1 penalty:

$$p_\lambda(|\theta|) = \lambda|\theta|.$$

The L_1 penalty function is famous in the family of L_q penalties because of LASSO, which was proposed by Tibshirani (1996). The solution to (2.46) with the L_1 penalty yields a soft thresholding rule

$$\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+. \tag{2.49}$$

The LASSO is very popular in the literature because it possesses both sparsity and continuity. Another important reason is that it can be solved efficiently by the least angle regression (LAR) algorithm (Efron et al. 2004). But the problem of LASSO is that the solution is biased and results of variable selection may be inconsistent, which was first conjectured by Fan and Li (2001) and recently showed by Leng et al. (2006) and Zou (2006).

- SCAD penalty:

$$p'_\lambda(|\theta|) = \lambda \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\}, \tag{2.50}$$

where a is a constant that is greater than 2. The SCAD penalty was proposed by Fan and Li (2001). It combines the merits of the Hard thresholding penalty and the L_1 penalty and has all three nice properties: unbiasedness,

sparsity, and continuity. The solution to (2.46) with SCAD is given by

$$\hat{\theta} = \begin{cases} 0 & \text{if} \quad |z| \leq \lambda \\ (|z| - \lambda)\text{sgn}(z) & \text{if} \quad \lambda < |z| \leq 2\lambda \\ \frac{1}{a-2}[(a-1)z - \text{sgn}(z)a\lambda] & \text{if} \quad 2\lambda < |z| \leq a\lambda \\ z & \text{if} \quad |z| > a\lambda \end{cases}, \tag{2.51}$$

which can be seen clearly from Figure 2.2. Another important fact argued by Fan and Li (2001) is that the SCAD enjoys the oracle property, that is, it works as well as if the true underlying model is known in an asymptotic sense. Actually, the SCAD is only a representative among a large family of penalties with all the three properties above. For detailed discussion, readers are referred to Fan and Li (2001).

- Adaptive LASSO penalty: The adaptive LASSO is a new penalized likelihood method newly proposed by Zou (2006). It starts from the weighted LASSO

$$\arg\min_{\boldsymbol{\beta}} \|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} w_j |\beta_j|, \tag{2.52}$$

and defines the adaptive LASSO as

$$\arg\min_{\boldsymbol{\beta}} \|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} \frac{1}{|\hat{\beta}_j|^\gamma} |\beta_j|, \tag{2.53}$$

where $\gamma > 0$ and $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \cdots, \hat{\beta}_d)^T$ is a root-n-consistent estimator of $\boldsymbol{\beta}_0$. A possible choice for $\hat{\boldsymbol{\beta}}$ is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. The author showed that the adaptive LASSO estimator also has the oracle property, which is a major improvement for LASSO. Meanwhile, the adaptive LASSO estimator can be solved by the same efficient algorithm (LARS) used for solving LASSO. So it could become a favorable alternative for LASSO. Zou (2006) also showed

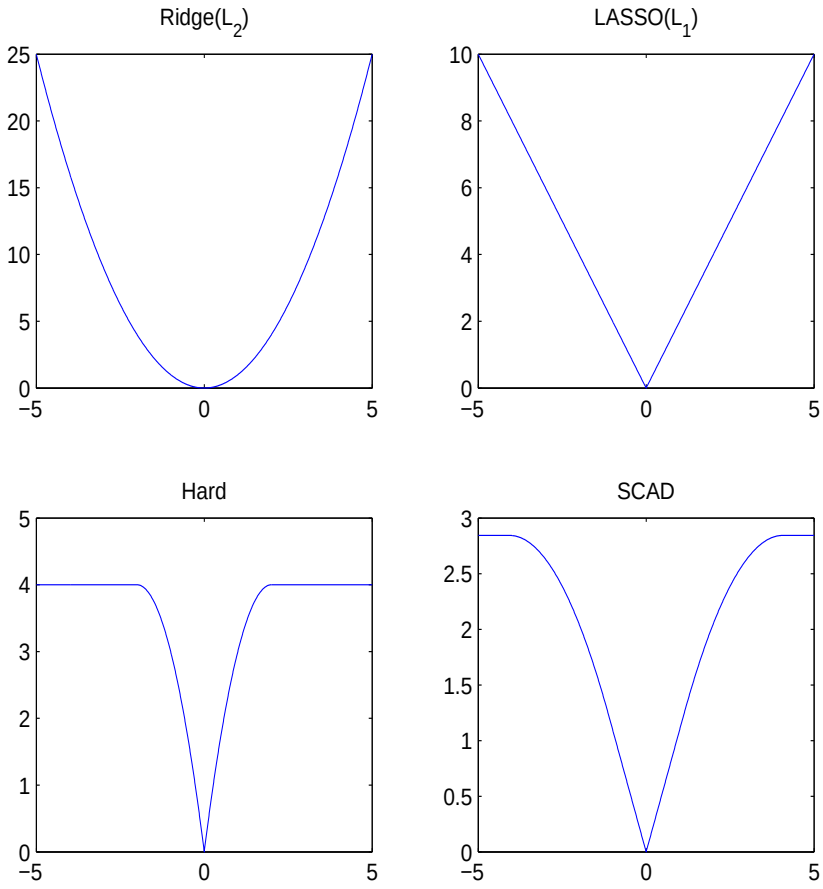


Fig. 2.1. Plots of four penalty functions $p_\lambda(\theta)$.

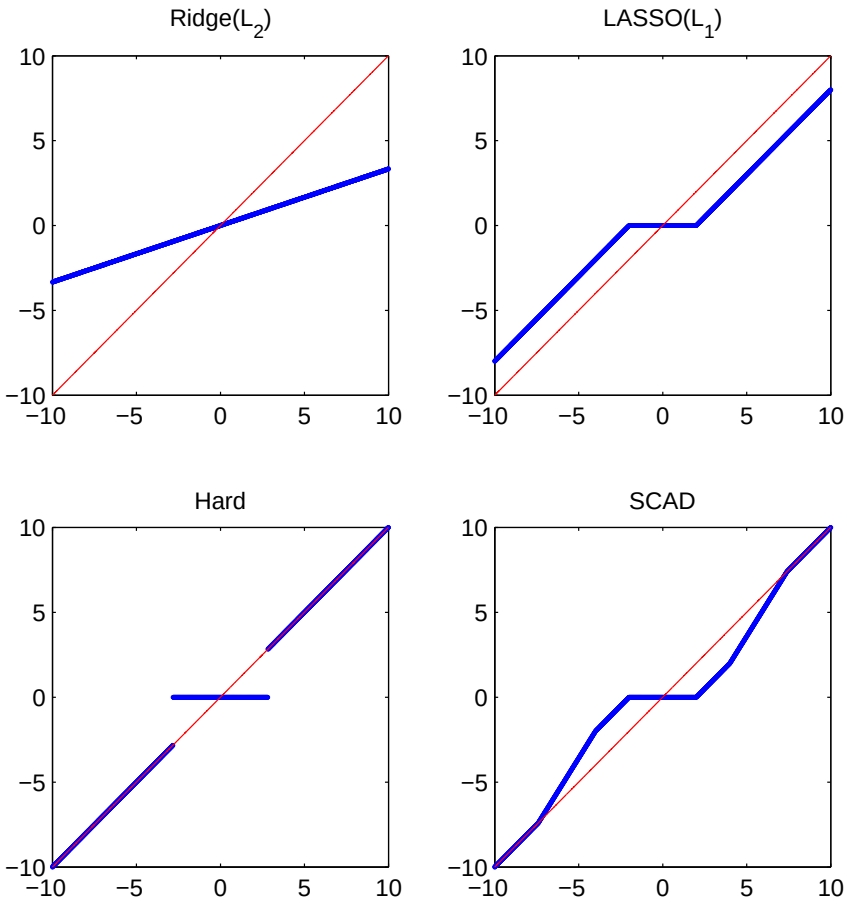


Fig. 2.2. Plots of thresholding functions for four penalties with $\lambda = 2$.

that the nonnegative garotte (Breiman 1995) is closely related to a special case of adaptive LASSO, and hence the nonnegative garotte is also consistent for variable selection.

Here is a summary on these penalty functions. The L_2 penalty (ridge regression) shrinks the coefficients but is not a thresholding rule and not appropriate for variable selection. The L_0 penalty (best subset selection) and hard-thresholding estimators (Antoniadis 1997) are unbiased and sparse but discontinuous. The LASSO gives continuous and sparse estimate but introduces bias. The SCAD and adaptive LASSO give continuous, sparse and unbiased models.

2.3.4 Algorithms for penalized likelihood optimization problems

Because the penalty functions may be nonsmooth, nonconcave and high-dimensional, optimizing penalized least squares or penalized likelihood is a challenging problem.

Tibshirani (1996) proposed LASSO, which can be solved efficiently by the least angle regression (LAR) algorithm (Efron et al. 2004). Fan and Li (2001) proposed a unified algorithm, LQA (local quadratic approximation), for optimizing penalized likelihood. LQA locally approximates $p_\lambda(|\theta|)$ by a quadratic function

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_{j0}|)}{|\beta_{j0}|}(\beta_j^2 - \beta_{j0}^2), \tag{2.54}$$

for $\beta_j \approx \beta_{j0}$. And then the optimization problem can be reduced to a quadratic minimization problem, which can be solved by the Newton-Raphson algorithm.

Hunter and Li (2005) proposed a new class of algorithms for finding a maximizer of the penalized likelihood for a broad class of penalty functions. These algorithms are named MM, which stands for majorize and minimize in minimization context and minorize and maximize in maximization context. The well-known EM algorithm is a special case of this more general class of MM algorithms. Wu

and Liu (2009) proposed to use the difference convex algorithm (DCA) to solve the SCAD penalized likelihood, which is an instance of the MM algorithm. Recently, Zou and Li (2008) proposed a new unified algorithm based on the local linear approximation (LLA) for maximizing the penalized likelihood for a broad class of penalty functions. A distinguished feature of the LLA algorithm is that at each LLA step, the LLA estimator can naturally adopt a sparse representation. So the authors suggest using the one-step LLA estimator with good initial estimators from the LLA algorithm as the final estimates. They also demonstrate that the LLA is the best convex MM algorithm.

Chapter 3

Local CQR Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression

3.1 Introduction

Consider the general nonparametric regression model

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

where Y is the response variable, T is a covariate, $m(T) = E(Y|T)$, which is assumed to be a smooth nonparametric function, and $\sigma(T)$ is a positive function representing the standard deviation. We assume ϵ has mean 0 and variance 1. Local polynomial regression is a popular and successful method for nonparametric regression, and it has been well studied in the literature (Fan and Gijbels 1996). By locally fitting a linear (or polynomial) regression model via adaptively weighted least squares, local polynomial regression is able to explore the fine features of the regression function and its derivatives. Although the least squares method is a popular and convenient choice in local polynomial fitting, we may consider using different local fitting methods. For example, in the presence of outliers, one may consider local least absolute deviation (LAD) polynomial regression (Fan et al. 1994; Welsh 1996). When the error follows a Laplacian distribution, the local LAD polynomial regression is more efficient than the local least squares polynomial regression. Of course, the local LAD polynomial regression can do much worse than the local least squares polynomial regression in other different settings. The aim of this chapter is to develop a new local estimation procedure that can significantly

improve upon the classical local polynomial regression for a wide class of error distributions, and has comparable efficiency in the worst case scenario.

Our proposal is built upon the composite-quantile-regression (CQR) estimator recently proposed by Zou and Yuan (2008) for estimating the regression coefficients in the classical linear regression model. Zou and Yuan (2008) show that the relative efficiency of the CQR estimator compared to the least squares estimator is greater than 70% regardless the error distribution. Furthermore, the CQR estimator could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator. These nice theoretical properties of CQR in linear regression motivate us to construct the local CQR smoothers as nonparametric estimates of the regression function and its derivatives.

We make several contributions in this chapter.

- We propose the local linear CQR estimator for estimating the nonparametric regression function. We establish the asymptotic theory of the local linear CQR estimator and show that, compared with the classical local linear least squares estimator, the new method can significantly improve the estimation efficiency of the local linear least squares estimator for commonly used non-normal error distributions.

- We propose the local quadratic CQR estimator for estimating the derivative of the regression function. The asymptotic theory shows that the local quadratic CQR estimator can often drastically improve the estimation efficiency of its local least squares counterpart if the error distribution is non-normal, and at the same time, the loss in efficiency is at most 8.01% in the worst case scenario.

- The general asymptotic theory of the local p-polynomial CQR estimator is established. Our theory does not require the error distribution to have a

finite variance. Therefore, local CQR estimators can work well even when local polynomial regression fails due to the infinite variance in the noise.

It is a well-known fact that the local linear (polynomial) regression is the best linear smoother in terms of efficiency (Fan and Gijbels 1996). There is no contradiction between this fact and our results, because the proposed local CQR estimator is a *nonlinear* smoother.

The rest of this chapter is organized as follows. In Section 2, we introduce the local linear CQR for the nonparametric regression and study its asymptotic properties. In Section 3, we propose the local quadratic CQR for estimating the derivative of the nonparametric regression, which can further reduce the estimation bias by the local linear CQR. Monte Carlo studies and a real data example are presented in Section 4. In Section 5 we present the general theory of the local p-polynomial CQR and technical proofs. The results for the infinite variance case are presented in Section 6. In Section 7, we also address the boundary behavior of the new estimators. Discussions are included in Section 8.

3.2 Estimation of the regression function

Suppose that (t_i, y_i), $i = 1, \cdots, n$, is an independent and identically distributed random sample. Consider estimating the value of $m(T)$ at t_0. In local linear regression we first approximate $m(t)$ locally by a linear function $m(t) \approx m(t_0) + m'(t_0)(t - t_0)$ and then fit a linear model locally in a neighborhood of t_0. Let $K(\cdot)$ be a smooth kernel function. Then the local linear regression estimator of $m(t_0)$ is $\hat{a}$, where

$$\{\hat{a}, \hat{b}\} = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - a - b(t_i - t_0)\}^2 K\left(\frac{t_i - t_0}{h}\right), \tag{3.2}$$

where h is the smoothing parameter. Local linear regression enjoys many good theoretical properties, such as its design adaptation property and high minimax

efficiency (Fan and Gijbels 1992). However, local least squares regression breaks down when the error distribution does not have finite second moment, for the estimator is no longer consistent. Local least absolute deviation (LAD) polynomial regression (Fan et al. 1994; Welsh 1996) replaces the least squares loss in (3.2) with the L_1 loss. By doing so, the local LAD estimator can deal with the infinite variance case, but for finite variance cases its relative efficiency compared to the local least squares estimator can be arbitrarily small.

We propose the local linear CQR estimator as an efficient alternative to the local linear regression estimator. Let $\rho_{\tau_k}(r) = \tau_k r - rI(r < 0), k = 1, 2, \ldots, q$, be q check loss functions at q quantile positions: $\tau_k = \frac{k}{q+1}$. In the linear regression model the CQR loss is defined as (Zou and Yuan 2008)

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\big(y_i - a_k - bt_i\big).$$

The CQR combines strength across multiple quantile regressions by forcing a single parameter for the "slope". Since the nonparametric function is approximated by a linear model locally, we consider minimizing the locally weighted CQR loss

$$\sum_{k=1}^{q}\left[\sum_{i=1}^{n}\rho_{\tau_k}\big\{y_i - a_k - b(t_i - t_0)\big\}K\left(\frac{t_i - t_0}{h}\right)\right]. \tag{3.3}$$

Denote the minimizer of (3.3) by $(\hat{a}_1, \cdots, \hat{a}_q, \hat{b})$. Then we let

$$\hat{m}(t_0) = \frac{1}{q}\sum_{k=1}^{q}\hat{a}_k \quad \text{and} \quad \tilde{m}'(t_0) = \hat{b}. \tag{3.4}$$

We refer to $\hat{m}(t_0)$ as the local linear CQR estimator of $m(t_0)$. As an estimator of $m'(t_0)$, $\tilde{m}'(t_0)$ can be further improved by using the local quadratic CQR estimator which is discussed in the next section.

Remark 1. It is worth mentioning here that although the check loss function is typically used to estimate the conditional quantile function of y given T (see Koenker (2005) and references therein), we simultaneously employ several check functions to estimate the regression (mean) function. So the local CQR smoother is conceptually different from nonparametric quantile regression by local fitting which has been studied in Yu and Jones (1998) and chapter 5 of Fan and Gijbels (1996).

Remark 2. In a short note Koenker (1984) studied the Hogg estimator as the minimizer of the weighted sum of check functions in the framework of parametric linear models. The focus there is to argue that the Hogg estimator is an alternative to L-estimators. The CQR loss can be regarded as a weighted sum of check functions with uniform weights and uniform quantiles ($\tau_k = \frac{k}{q+1}, k = 1, 2, \ldots, q$). When q is large, such a choice leads to nice oracle-like estimators in the oracle model selection theoretic framework (Zou and Yuan 2008). Koenker (1984) did not discuss relative efficiency of the Hogg estimator relative to the least squares estimator. In this work we consider minimizing the locally weighted CQR loss and show that the local CQR smoothers have very interesting asymptotic efficiency properties. To our best knowledge, none of these has been studied in the literature.

Remark 3. Zou and Yuan (2008) assume parallel quantile lines, whereas our method allows for heteroscedasticity.

3.2.1 Asymptotic properties

To see why local linear CQR is an efficient alternative to local linear regression, we establish the asymptotic properties of the local linear CQR estimator. Some notation is necessary for the discussion. Let $F(\cdot)$ and $f(\cdot)$ denote the density function and cumulative distribution function of the error distribution, respectively. Denote by $f_T(\cdot)$ the marginal density function of the covariate T. We

choose the kernel $K(\cdot)$ as a symmetric density function and let

$$\mu_j = \int u^j K(u)du \quad \text{and} \quad \nu_j = \int u^j K^2(u)du, \quad j = 0, 1, 2, \ldots, .$$

Define

$$R_1(q) = \frac{1}{q^2}\sum_{k=1}^{q}\sum_{k'=1}^{q}\frac{\tau_{kk'}}{f(c_k)f(c_{k'})}, \tag{3.5}$$

where $c_k = F^{-1}(\tau_k)$ and $\tau_{kk'} = \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}$. The quantity $R_1(q)$ plays a fundamental role in the theory. In the following theorem, we present the asymptotic bias, variance and normality of $\hat{m}(t_0)$, whose proof is given in section 5. Let $\mathbf{T}$ be the σ-field generated by $\{T_1, \cdots, T_n\}$.

Theorem 3.1. *Suppose that t_0 is an interior point of the support of $f_T(\cdot)$. Under the regularity conditions (A)—(D) in section 5, if $h \to 0$ and $nh \to \infty$, then the asymptotic conditional bias and variance of the local linear CQR estimator $\hat{m}(t_0)$ are given by*

$$Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{2}m''(t_0)\mu_2 h^2 + o_p(h^2), \tag{3.6}$$

$$Var(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{nh}\frac{\nu_0\sigma^2(t_0)}{f_T(t_0)}R_1(q) + o_p(\frac{1}{nh}). \tag{3.7}$$

Furthermore, conditioning on $\mathbf{T}$, we have

$$\sqrt{nh}\{\hat{m}(t_0) - m(t_0) - \frac{1}{2}m''(t_0)\mu_2 h^2\} \xrightarrow{\mathcal{L}} N\left(0, \frac{\nu_0\sigma^2(t_0)}{f_T(t_0)}R_1(q)\right). \tag{3.8}$$

where $\xrightarrow{\mathcal{L}}$ stands for convergence in distribution.

We see from Theorem 3.1 that the leading term of the asymptotic bias for the local linear CQR estimator is the same as that for the local linear least squares estimator, while their asymptotic variances are different. The mean squared error

of $\hat{m}(t_0)$ is

$$\mathrm{MSE}\{\hat{m}(t_0)\} = \left\{\frac{1}{2}m''(t_0)\mu_2\right\}^2 h^4 + \frac{1}{nh}\frac{\nu_0\sigma^2(t_0)}{f_T(t_0)}R_1(q) + o_p\left(h^4 + \frac{1}{nh}\right).$$

By straightforward calculations we can see that the optimal variable bandwidth minimizing the asymptotic mean squared error of $\hat{m}(t_0)$ is

$$h^{\mathrm{opt}}(t_0) = \left[\frac{\nu_0\sigma^2(t_0)R_1(q)}{f_T(t_0)\{m''(t_0)\mu_2\}^2}\right]^{1/5} n^{-1/5}.$$

In practice, one may select a constant bandwidth by minimizing the mean integrated squared error

$$\mathrm{MISE}(\hat{m}) = \int \mathrm{MSE}\{\hat{m}(t_0)\}w(t)\,dt$$

for a weight function $w(t)$. Similarly, the optimal bandwidth minimizing the asymptotic $\mathrm{MISE}(\hat{m})$ is

$$h^{\mathrm{opt}} = \left[\frac{\nu_0 R_1(q)\int \sigma^2(t)f_T^{-1}(t)w(t)dt}{\mu_2^2\int\{m''(t)\}^2 w(t)dt}\right]^{1/5} n^{-1/5}.$$

The above calculations indicate that the local linear CQR estimator still enjoys the optimal rate of convergence $n^{2/5}$.

3.2.2 Asymptotic relative efficiency

In this section, we study the asymptotic relative efficiency of the local linear CQR estimator with respect to the local linear least squares estimator by comparing their mean squared errors. The role of R_1 becomes clear in the relative efficiency study.

The local linear least squares estimator for $m(t_0)$ has the mean squared error

$$\mathrm{MSE}\{\hat{m}_{\mathrm{LS}}(t_0)\} = \left\{\frac{1}{2}m''(t_0)\mu_2\right\}^2 h^4 + \frac{1}{nh}\frac{\nu_0}{f_T(t_0)}\sigma^2(t_0) + o_p\left(h^4 + \frac{1}{nh}\right),$$

and the optimal variable bandwidth minimizing the asymptotic mean squared error is $h_{\mathrm{LS}}^{\mathrm{opt}}(t_0) = \left[\frac{\nu_0\sigma^2(t_0)}{f_T(t_0)\{m''(t_0)\mu_2\}^2}\right]^{1/5} n^{-1/5}$. Similarly, if considering the mean integrated squared error, the optimal bandwidth is $h_{\mathrm{LS}}^{\mathrm{opt}} = \left[\frac{\nu_0 \int \sigma^2(t) f_T^{-1}(t) w(t) dt}{\mu_2^2 \int \{m''(t)\}^2 w(t) dt}\right]^{1/5} n^{-1/5}$ with a weight function $w(t)$. Therefore, we have

$$h^{\mathrm{opt}}(t_0) = R_1(q)^{1/5} h_{\mathrm{LS}}^{\mathrm{opt}}(t_0), \quad h^{\mathrm{opt}} = R_1(q)^{1/5} h_{\mathrm{LS}}^{\mathrm{opt}}. \tag{3.9}$$

We use $\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}(t_0)\}$ and $\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}_{\mathrm{LS}}(t_0)\}$ to denote the mean squared errors of $\hat{m}(t_0)$ and $\hat{m}_{\mathrm{LS}}(t_0)$ evaluated at their own optimal bandwidth. Then it is easy to see that

$$\frac{\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}_{\mathrm{LS}}(t_0)\}}{\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}(t_0)\}} \longrightarrow R_1(q)^{-4/5}.$$

It is interesting to note that the above ratio does not depend on the location t_0. Similarly, if we compare the mean integrated squared errors with the optimal bandwidths, we also have

$$\frac{\mathrm{MISE}_{\mathrm{opt}}\{\hat{m}_{\mathrm{LS}}(t_0)\}}{\mathrm{MISE}_{\mathrm{opt}}\{\hat{m}(t_0)\}} \longrightarrow R_1(q)^{-4/5}.$$

Thus, the asymptotic relative efficiency (ARE) of the local linear CQR estimator with respect to the local linear least squares estimator is by definition

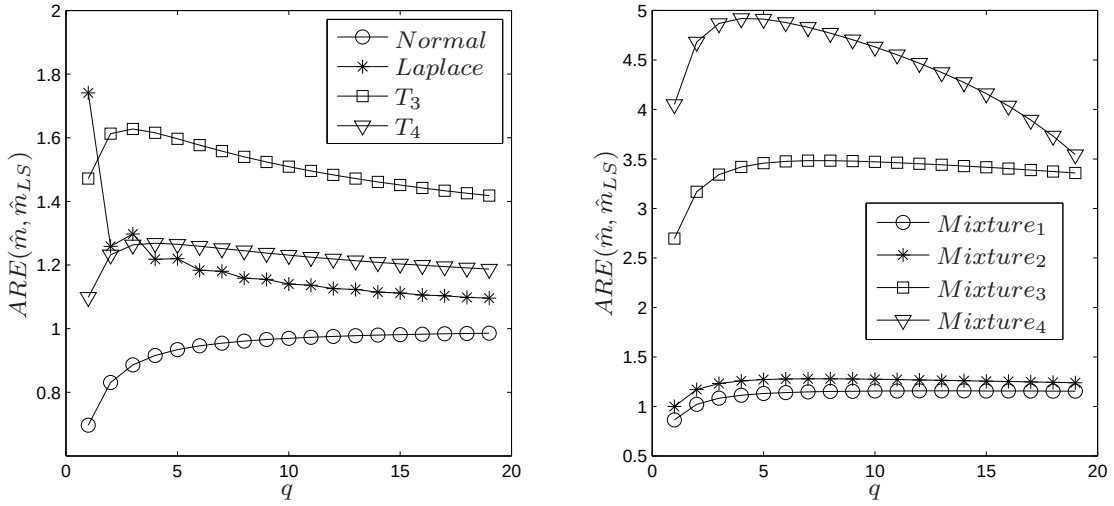
$$\mathrm{ARE}(\hat{m}, \hat{m}_{\mathrm{LS}}) = R_1(q)^{-4/5}. \tag{3.10}$$

The ARE depends on the error distribution in a rather complex way. However, for many commonly seen error distributions, we can obtain the ARE by straightforward calculations. Table 3.1 displays the ARE$(\hat{m}, \hat{m}_{\mathrm{LS}})$ for some commonly seen error distributions. The trends of the ARE$(\hat{m})$ over q are displayed in Figure 3.1 (a).

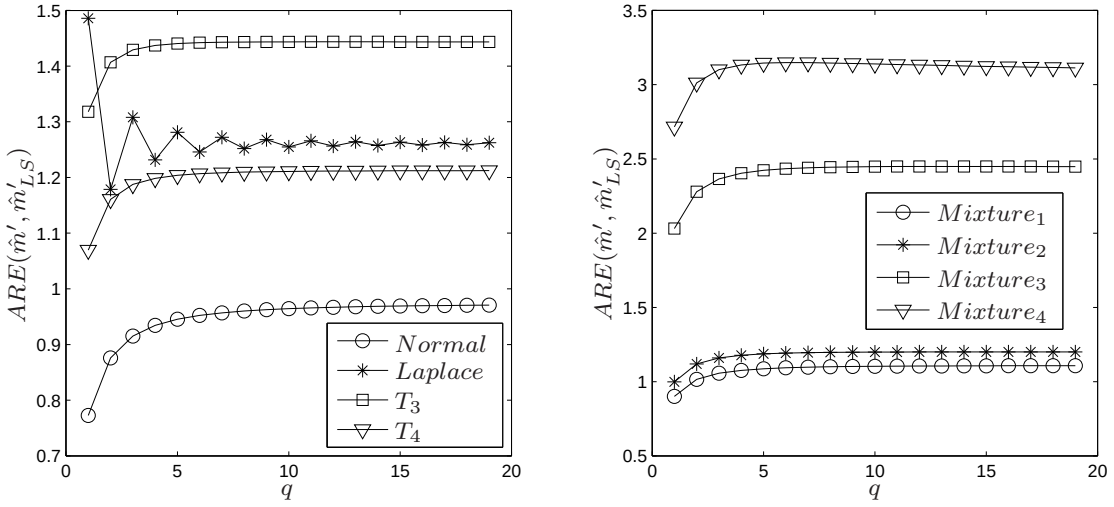
Table 3.1. Comparisons of ARE$(\hat{m}, \hat{m}_{\mathrm{LS}})$

Error Distribution	**ARE$(\hat{m}, \hat{m}_{\mathrm{LS}})$**				
	$q = 1$	$q = 5$	$q = 9$	$q = 19$	$q = 99$
$N(0,1)$	0.6968	0.9339	0.9659	0.9858	0.9980
Laplace	1.7411	1.2199	1.1548	1.0960	1.0296
t-distribution with $df = 3$	1.4718	1.5967	1.5241	1.4181	1.2323
t-distribution with $df = 4$	1.0988	1.2652	1.2377	1.1872	1.0929
$.95N(0,1) + .05N(0,3^2)$	0.8639	1.1300	1.1536	1.1540	1.0804
$.90N(0,1) + .10N(0,3^2)$	0.9986	1.2712	1.2768	1.2393	1.0506
$.95N(0,1) + .05N(0,10^2)$	2.6960	3.4577	3.4783	3.3591	1.3498
$.90N(0,1) + .10N(0,10^2)$	4.0505	4.9128	4.7049	3.5444	1.1379

Several interesting observations can be made from Table 3.1. First, when the error distribution is $N(0,1)$, for which the local linear least squares estimator is expected to have the best performance, the ARE$(\hat{m}, \hat{m}_{\mathrm{LS}})$ is very close to 1 as long as $q > 2$ in the local linear CQR estimator. When $q = 5$ the the local linear CQR only loses at most 7% efficiency, while it performs as well as the local linear least squares estimator when $q = 99$. Secondly, for all the other non-normal distributions listed in Table 3.1, the local linear CQR estimator can have higher efficiencies than the local linear least squares estimator when a small q is used. The mixture of two normals is often used to model the so-called contaminated data. For such distributions, the ARE$(\hat{m}, \hat{m}_{\mathrm{LS}})$ can be as large as 4.9 and even more. Table 3.1 also indicates that, except for the Laplace error, the local CQR



(a)



(b)

Fig. 3.1. Graphs showing $\mathrm{ARE}(\hat{m}, \hat{m}_{\mathrm{LS}})$ and $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ as a function of q for some commonly seen error distributions. (a) for $\mathrm{ARE}(\hat{m}, \hat{m}_{\mathrm{LS}})$, and (b) $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$. Mixture$_v$, $v = 1, 2, 3$ and 4 stand for $0.95N(0,1) + 0.05N(0, 3^2)$, $0.90N(0,1) + 0.10N(0, 3^2)$, $0.95N(0,1) + 0.05N(0, 10^2)$ and $0.95N(0,1) + 0.05N(0, 10^2)$, respectively.

with $q = 5$ or $q = 9$ is significantly better than the one with $q = 1$, which becomes the local LAD for these distributions. Finally, we observe that the ARE values for a variety of distributions are very close to 1 when q is large ($q = 99$). It turns out that this phenomenon is true in general, as demonstrated in the following theorem.

Theorem 3.2. $\lim_{q\to\infty} R_1(q) = 1$, *and thus* $\lim_{q\to\infty} ARE(\hat{m}, \hat{m}_{LS}) = 1$.

Theorem 3.2 provides us insights into the asymptotic behavior of the local linear CQR estimator and implies that the local linear CQR estimator is a safe competitor against the local linear least squares estimator, for it will not lose efficiency when using a large q. On the other hand, substaintial gain in efficiency could be achieved by using a relatively small q such as $q = 9$, as shown in Table 3.1.

3.3 Estimation of derivative

In many situations we are interested in estimating the derivative of $m(t)$. The local linear CQR also provides an estimator $\tilde{m}'(t_0)$ to the derivative of $m(t)$. The asymptotic bias and variance of the estimate $\tilde{m}'(t_0)$ in (3.4) are given in (3.31) and (3.32) in section 5. The local linear CQR estimator and the local linear regression estimator have the same leading bias term which depends on the intrinsic part $m'''(t_0)$ and the extra part $m''(t_0)f_T'(t_0)/f_T(t_0)$. In Chu and Marron (1991) and Fan (1992), the authors already argued that the bias of Nadaraya-Watson estimator (also involves similar term) could be very large in many situations. So $\tilde{m}'(t_0)$ may not be an ideal estimator because of the relatively large bias. The local quadratic regression is often preferred for estimating the derivative function, since it reduces the estimation bias without increasing the estimation variance (Fan and Gijbels 1992). We show here that the same phenomenon is true in local CQR smoothing.

We consider the local quadratic approximation of $m(t)$ in the neighborhood of t_0: $m(t) \approx m(t_0) + m'(t_0)(t - t_0) + \frac{1}{2}m''(t_0)(t - t_0)^2$. Let $\mathbf{a} = (a_1, \cdots, a_q)$ and

$\mathbf{b} = (b_1, b_2)$. We solve

$$(\hat{\mathbf{a}}, \hat{\mathbf{b}}) = \underset{\mathbf{a},\mathbf{b}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\sum_{k=1}^{q} \rho_{\tau_k} \left(y_i - a_k - b_1(t_i - t_0) - \frac{1}{2} b_2 (t_i - t_0)^2 \right) K \left(\frac{t_i - t_0}{h} \right) \right]. \tag{3.11}$$

Then the local quadratic CQR estimator for $m'(t_0)$ is given by

$$\hat{m}'(t_0) = \hat{b}_1. \tag{3.12}$$

3.3.1 Asymptotic properties

Denote

$$R_2(q) = \left(\sum_{k=1}^{q} \sum_{k'=1}^{q} \tau_{kk'} \right) \Big/ \left(\sum_{k=1}^{q} f(c_k) \right)^2. \tag{3.13}$$

The asymptotic bias, variance and normality are given in the following theorem.

Theorem 3.3. *Suppose that t_0 is an interior point of the support of $f_T(\cdot)$. Under the regularity conditions (A)—(D) in section 5, if $h \to 0$ and $nh^3 \to \infty$, then the asymptotic conditional bias and variance of $\hat{m}'(t_0)$, defined in (3.12), is given by*

$$Bias(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{6} m'''(t_0) \frac{\mu_4}{\mu_2} h^2 + o_p(h^2), \tag{3.14}$$

$$Var(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{nh^3} \frac{\nu_2 \sigma^2(t_0)}{\mu_2^2 f_T(t_0)} R_2(q) + o_p\left(\frac{1}{nh^3} \right). \tag{3.15}$$

Furthermore, conditioning on $\mathbf{T}$, we have the following asymptotic normal distribution

$$\sqrt{nh^3} \left(\hat{m}'(t_0) - m'(t_0) - \frac{1}{6} m'''(t_0) \frac{\mu_4}{\mu_2} h^2 \right) \xrightarrow{\mathcal{L}} N\left(0, \frac{\nu_2 \sigma^2(t_0)}{\mu_2^2 f_T(t_0)} R_2(q) \right). \tag{3.16}$$

Comparing (3.31) and (3.14), we see that the extra part $m''(t_0) f_T'(t_0)/f_T(t_0)$ is removed in the local quadratic CQR estimator. Comparing the local quadratic

CQR and the local quadratic least squares estimators for $m'(t_0)$, we see that they have the same leading bias term, while their asymptotic variances are different.

From Theorem 3.3, the mean squared error of local quadratic CQR estimator $\hat{m}'(t_0)$ is given by

$$\mathrm{MSE}\{\hat{m}'(t_0)\} = \left(\frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}\right)^2 h^4 + \frac{1}{nh^3}\frac{\nu_2\sigma^2(t_0)}{\mu_2^2 f_T(t_0)}R_2(q) + o_p\left(h^4 + \frac{1}{nh^3}\right).$$

Thus, the optimal variable bandwidth minimizing $\mathrm{MSE}\{\hat{m}'(t_0)\}$ is

$$h^{\mathrm{opt}}(t_0) = \{R_2(q)\}^{1/7}\left(\frac{27\nu_2\sigma^2(t_0)}{f_T(t_0)\{m'''(t_0)\mu_4\}^2}\right)^{1/7} n^{-1/7}.$$

Furthermore, we consider the mean integrated squared error

$$\mathrm{MISE}(\hat{m}') = \int \mathrm{MSE}\{\hat{m}'(t)\}w(t)\,dt$$

with a weight function $w(t)$. The optimal constant bandwidth minimizing the mean integrated squared error is given by

$$h^{\mathrm{opt}} = \{R_2(q)\}^{1/7}\left(\frac{27\nu_2\int\sigma^2(t)f_T^{-1}(t)w(t)\,dt}{\int\{m'''(t)\}^2 w(t)\,dt\,\mu_4^2}\right)^{1/7} n^{-1/7}.$$

The above calculations indicate that the local quadratic CQR estimator enjoys the optimal rate of convergence $n^{2/7}$.

3.3.2 Asymptotic relative efficiency

In what follows we study the asymptotic relative efficiency of the local quadratic CQR estimator with respect to the local quadratic least squares estimator.

Note that the mean squared error of the local quadratic least squares estimator $\hat{m}'_{\mathrm{LS}}(t_0)$ is given by (Fan and Gijbels 1996)

$$\mathrm{MSE}\{\hat{m}'_{\mathrm{LS}}(t_0)\} = \left(\frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}\right)^2 h^4 + \frac{1}{nh^3}\frac{\nu_2\sigma^2(t_0)}{\mu_2^2 f_T(t_0)} + o_p\left(h^4 + \frac{1}{nh^3}\right),$$

and the mean integrated squared error is

$$\mathrm{MISE}(\hat{m}'_{\mathrm{LS}}) = \int \mathrm{MSE}\{\hat{m}'_{\mathrm{LS}}(t)\}w(t)\,dt$$

with a weight function $w(t)$. Thus, by straightforward calculations, we notice that

$$h^{\mathrm{opt}}(t_0) = h^{\mathrm{opt}}_{\mathrm{LS}}(t_0)R_2(q)^{1/7}, \quad h^{\mathrm{opt}} = h^{\mathrm{opt}}_{\mathrm{LS}}R_2(q)^{1/7}, \tag{3.17}$$

where $h^{\mathrm{opt}}_{\mathrm{LS}}(t_0)$ and $h^{\mathrm{opt}}_{\mathrm{LS}}$ are the corresponding optimal bandwidths of local quadratic least squares estimator for the derivative of the regression function. With the optimal bandwidths, we have

$$\frac{\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}'_{\mathrm{LS}}(t_0)\}}{\mathrm{MSE}_{\mathrm{opt}}\{\hat{m}'(t_0)\}} \longrightarrow R_2(q)^{-4/7}$$

at each t_0 and

$$\frac{\mathrm{MISE}_{\mathrm{opt}}(\hat{m}'_{\mathrm{LS}})}{\mathrm{MISE}_{\mathrm{opt}}(\hat{m}')} \longrightarrow R_2(q)^{-4/7}.$$

Therefore, the asymptotic relative efficiency (ARE) of the local quadratic CQR estimator $(\hat{m}')$ with respect to the local quadratic least squares estimator $(\hat{m}'_{LS})$ is by definition

$$\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}}) = R_2(q)^{-4/7}. \tag{3.18}$$

The ARE only depends on the error distribution and it is scale invariant.

Table 3.2. Comparisons of ARE$(\hat{m}', \hat{m}'_{\text{LS}})$

Error Distribution	**ARE$(\hat{m}', \hat{m}'_{\text{LS}})$**					
	$q = 1$	$q = 5$	$q = 9$	$q = 19$	$q = 99$	$q = \infty$
$N(0,1)$	0.7726	0.9453	0.9625	0.9708	0.9738	0.9740
Laplace	1.4860	1.2812	1.2680	1.2625	1.2608	1.2607
t-distribution with $df = 3$	1.3179	1.4405	1.4435	1.4435	1.4430	1.4431
t-distribution with $df = 4$	1.0696	1.2038	1.2104	1.2123	1.2125	1.2125
$.95N(0,1) + .05N(0,3^2)$	0.9008	1.0867	1.1019	1.1073	1.1077	1.1077
$.90N(0,1) + .10N(0,3^2)$	0.9990	1.1869	1.1982	1.1999	1.1987	1.1987
$.95N(0,1) + .05N(0,10^2)$	2.0308	2.4229	2.4466	2.4482	2.4415	2.4415
$.90N(0,1) + .10N(0,10^2)$	2.7160	3.1453	3.1430	3.1135	3.1094	3.1093

To gain insights into the asymptotic relative efficiency, we consider the limit when q is large. Zou and Yuan (2008) showed that

$$\lim_{q\to\infty} R_2(q)^{-1} > \frac{6}{e\pi} = 0.7026.$$

Immediately, we know that if using a large q, the ARE is bounded below by $0.7026^{4/7} = 0.8173$. Having a universal lower bound is very useful because it prohibits severe loss in efficiency when replacing the local quadratic least squares estimator with the local quadratic CQR estimator. One of our contributions in this work is to provide an improved sharper lower bound, as shown in the following theorem.

Theorem 3.4. *Let $\mathcal{F}$ denote the class of error distributions with mean 0 and variance* 1*. Then we have*

$$\inf_{f\in\mathcal{F}} \lim_{q\to\infty} R_2(q)^{-1} = 0.864. \tag{3.19}$$

The lower bound is reached if and only if the error follows the rescaled Beta(2,2) distribution with mean zero and variance one. Thus

$$\lim_{q\to\infty} ARE(\hat{m}', \hat{m}'_{LS}) \geq 0.9199. \tag{3.20}$$

It is interesting to note that Theorem 3.4 provides us the *exact* lower bound of $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ as $q \to \infty$. Theorem 3.4 indicates that if q is large, even in the worst scenario the potential efficiency loss for the local CQR estimator is only 8.01%.

Theorem 3.4 implies that the local quadratic CQR estimator is a safe alternative to the local quadratic least squares estimator. It concerns the worst case scenario. There are many optimistic scenarios as well in which the ARE can be much bigger than 1. We examine the $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ for the error distributions considered in Table 3.1. The trends of the $\mathrm{ARE}(\hat{m}')$ over q are displayed in Figure 3.1 (b). We also list the results in Table 3.2, where the column labeled $q = \infty$ shows the theoretical limit of the $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$. Obviously, these limits are all larger than the lower bound 0.9199. The local quadratic CQR estimator only loses less than 4% efficiency when the error distribution is normal and $q = 9$. It is interesting to see that for the other non-normal distributions the $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ is larger than 1 and its value is insensitive to the choice of q. For example, with $q = 9$, the AREs are already very close to their theoretical limits.

It is worth emphasizing here that the local LAD estimator does not enjoy such a nice property, for its relative efficiency with respect to the local linear least squares estimator can be arbitrarily small.

3.4 Numerical comparisons and examples

In this section, we first use Monte Carlo simulation studies to assess the finite sample performance of the proposed estimation procedures and then demonstrate the application of the proposed method by using a real data example. Throughout this section we use the Epanechnikov kernel, i.e., $K(z) = \frac{3}{4}(1 - z^2)_+$. We adopt the MM algorithm proposed by Hunter and Lange (2000) for solving the local CQR smoothing estimator. All the numerical results are computed using our MATLAB code, which is available upon request.

3.4.1 Bandwidth selection in practical implementation

Bandwidth selection is an important issue in local smoothing. Here we briefly discuss the bandwidth selection in the local CQR smoothing estimator by using an existing bandwidth selector for the ordinary local polynomial regression. Here we consider two kinds of bandwidth selectors.

1. The "pilot" selector. The idea is to use a pilot bandwidth in local cubic CQR (defined in section 5) to estimate $m''(t)$ and $m'''(t)$. The fitted residuals can be used to estimate $R_1(q)$ and $R_2(q)$. Thus, we can estimate the optimal bandwidth and then refit the data.

2. A short-cut strategy. In our numerical studies, we compare the local CQR and local least squares estimators. Note that in (3.9) and (3.17) we obtain very neat relationships between the optimal bandwidths for the local CQR and local least squares estimators. The optimal bandwidth for the local least squares estimators can be selected by existing bandwidth selectors (see Chapter 4 of Fan and Gijbels (1996)). In addition, we are able to infer the factors $R_1(q)$ and $R_2(q)$ from the residuals of the local least squares fit. Sometimes, we even know the exact values of the two factors (e.g., in simulations). Therefore, after fitting the local least squares estimator with

the optimal bandwidth, we can estimate the optimal bandwidth for the local CQR estimator.

We used the short-cut strategy in our simulation examples. However, if the error variance is infinite or very large, then the local least squares estimator performs poorly. The "pilot" selector is a better choice than the short-cut strategy.

3.4.2 Simulation examples

In our simulation studies, we compare the performance of the newly proposed method with the local polynomial least squares estimate. The bandwidth is set to the optimal one in which the $h_{\mathrm{LS}}^{\mathrm{opt}}$ is selected by a plug-in bandwidth selector (Ruppert et al. 1995).

The performance of estimator $\hat{m}(\cdot)$ and $\hat{m}'(\cdot)$ is assessed via the average squared error (ASE), defined by

$$\mathrm{ASE}(\hat{g}) = \frac{1}{n_{\mathrm{grid}}} \sum_{k=1}^{n_{\mathrm{grid}}} \{\hat{g}(u_k) - g(u_k)\}^2,$$

with g equal to either $m(\cdot)$ or $m'(\cdot)$, where $\{u_k,\, k = 1, \ldots, n_{\mathrm{grid}}\}$ are the grid points at which the functions $\{\hat{g}(\cdot)\}$ are evaluated. In our simulation, we set $n_{\mathrm{grid}} = 200$ and grid points are evenly distributed over the interval at which the $m(\cdot)$ and $m'(\cdot)$ are estimated. We summarize our simulation results using the ratio of average squared errors (RASE),

$$\mathrm{RASE}(\hat{g}) = \frac{\mathrm{ASE}(\hat{g}_{\mathrm{LS}})}{\mathrm{ASE}(\hat{g})}, \tag{3.21}$$

for an estimator $\hat{g}$, where $\hat{g}_{\mathrm{LS}}$ is the local polynomial regression estimator under the least squares loss. We considered two simulation examples.

Example 3.4.1. We generated 400 data sets, each consisting of $n = 200$ observations, from

$$Y = \sin(2T) + 2\exp(-16T^2) + 0.5\epsilon, \tag{3.22}$$

where T follows $N(0,1)$. This model is adopted from Fan and Gijbels (1992). In our simulation, we considered five error distributions for ϵ: $N(0,1)$, Laplace, t_3 distribution, a mixture of two normals $(0.95N(0,1)+0.05N(0,\sigma^2)$ with $\sigma = 3, 10)$. For the local polynomial CQR estimator, we consider $q = 5$, 9 and 19, and estimate $m(\cdot)$ and $m'(\cdot)$ over $[-1.5, 1.5]$. The mean and standard deviation of RASE over 400 simulations are summarized in Table 3.3. To see how the proposed estimate behaves at a typical point, Table 3.3 also depicts the biases and standard deviations of $\hat{m}(t)$ and $\hat{m}'(t)$ at $t = 0.75$. In Table 3.3, CQR_5, CQR_9 and CQR_{19} correspond to the local CQR estimate with $q = 5$, 9 and 19, respectively.

Example 3.4.2. It is of interest to investigate the effect of heteroscedastic errors. To this end, we generated 400 simulation data sets, each consisting of $n = 200$ observations, from

$$Y = T\sin(2\pi T) + \sigma(T)\epsilon, \tag{3.23}$$

where T follows $U(0,1)$, $\sigma(t) = \{2 + \cos(2\pi t)\}/10$, and ϵ is the same as that in Example 3.4.1. In this example, we estimate $m(t)$ and $m'(t)$ over [0,1]. The mean and standard deviation of RASE over 400 simulations are summarized in Table 3.4, in which we also show the biases and standard deviations of $\hat{m}(t)$ and $\hat{m}'(t)$ at $t = 0.4$. The notation of Table 3.4 is the same as that in Table 3.3.

Table 3.3 and Table 3.4 show very similar messages, although Table 3.4 indicates that the local CQR has more gains over the local least squares method. When the error follows the normal distribution, the RASEs of the local CQR estimators are slightly less than one. For non-normal distributions, the RASEs of the local CQR estimators can be greater than one, indicating the gain in efficiency. For estimating the regression function, CQR_5 and CQR_9 seem to have better overall

performance than CQR_{19}. For estimating the derivative, all three CQR estimators perform very similarly. These findings are consistent with the theoretical analysis of AREs.

3.4.3 A real data example

As an illustration, we now apply the proposed local CQR methodology to the U.K. Family Expenditure Survey data subset with high net income, which consists of 363 observations. The scatter plot of data is depicted in the left panel of Figure 3.2. The data set was collected in the U.K. Family Expenditure Survey in 1973. Of interest is to study the relationship between the food expenditure and the net income. Thus, we take the response variable Y to be the logarithm of the food expenditure, and the predictor variable T is the net-income.

We first estimated the regression function using the local least squares estimator with the plug-in bandwidth selector (Ruppert et al. 1995). We further employed the kernel density estimate to infer the error density $f(\cdot)$ based on the residuals from the local least squares estimator. Based on the estimated density, we estimated both $R_1(q)$ and $R_2(q)$, which were used to compute the bandwidth selector for the CQR estimator. For this example, the estimated ratios are close to 1, so we basically use the same bandwidths for these two methods. The selected bandwidths are 0.24 for regression estimation and 0.4 for derivative estimation. The CQR estimates with $q = 5, 9$ and 19 with the selected bandwidths are evaluated. The CQR estimates with three different q's are very similar, so we only present the CQR estimate with $q = 9$ in Figure 3.2.

It is interesting to see from Figure 3.2 that the overall patten of the local least squares and the local CQR estimate are the same. The difference between the local least squares estimate and the local CQR estimate of the regression function becomes large when the net income is around 2.8. From the scatter plot, there are

Table 3.3. Simulation results for example 3.4.1

	$\hat{m}$			$\hat{m}'$		
	RASE	$t = 0.75$		**RASE**	$t = 0.75$	
	Mean(SD)	**Bias**	**Std**	**Mean(SD)**	**Bias**	**Std**
Standard Normal						
LS	—	-0.0239	0.1098	—	-0.0539	0.6871
CQR_5	$0.9314_{(0.1190)}$	-0.0224	0.1161	$0.9518_{(0.1087)}$	-0.0508	0.7257
CQR_9	$0.9588_{(0.0888)}$	-0.0236	0.1133	$0.9614_{(0.1019)}$	-0.0530	0.7165
CQR_{19}	$0.9802_{(0.0592)}$	-0.0228	0.1117	$0.9646_{(0.0998)}$	-0.0513	0.7178
Laplace						
LS	—	-0.0146	0.1215	—	-0.1108	0.6988
CQR_5	$1.1088_{(0.1985)}$	-0.0171	0.1155	$1.1014_{(0.1679)}$	-0.0774	0.6916
CQR_9	$1.0717_{(0.1351)}$	-0.0154	0.1195	$1.1025_{(0.1565)}$	-0.0834	0.6678
CQR_{19}	$1.0346_{(0.0856)}$	-0.0141	0.1214	$1.1005_{(0.1500)}$	-0.0934	0.6529
t-distribution with df $= 3$						
LS	—	-0.0214	0.1266	—	-0.0701	0.7254
CQR_5	$1.2752_{(0.5020)}$	-0.0182	0.1103	$1.2104_{(0.4584)}$	-0.0559	0.6635
CQR_9	$1.1712_{(0.3356)}$	-0.0158	0.1137	$1.2133_{(0.4526)}$	-0.0520	0.6537
CQR_{19}	$1.0710_{(0.2086)}$	-0.0186	0.1222	$1.2182_{(0.4403)}$	-0.0540	0.6431
$.95N(0,1) + .05N(0,9)$						
LS	—	-0.0007	0.1256	—	-0.0382	0.8540
CQR_5	$1.0685_{(0.2275)}$	-0.0060	0.1202	$1.0479_{(0.1773)}$	-0.0182	0.8098
CQR_9	$1.0621_{(0.1740)}$	-0.0049	0.1219	$1.0531_{(0.1727)}$	-0.0154	0.8085
CQR_{19}	$1.0280_{(0.1125)}$	-0.0018	0.1251	$1.0532_{(0.1687)}$	-0.0198	0.8062
$.95N(0,1) + .05N(0,100)$						
LS	—	0.0034	0.1283	—	-0.0456	0.8667
CQR_5	$2.1548_{(1.5318)}$	0.0002	0.0888	$1.7671_{(0.7607)}$	0.0022	0.5953
CQR_9	$1.5240_{(0.8360)}$	-0.0009	0.1181	$1.7527_{(0.7535)}$	0.0024	0.6030
CQR_{19}	$1.1600_{(0.8776)}$	0.0069	0.1365	$1.7560_{(0.7382)}$	0.0044	0.5927

Table 3.4. Simulation results for example 3.4.2.

	$\hat{m}$			$\hat{m}'$		
	RASE	$t = 0.4$		**RASE**	$t = 0.4$	
	Mean(SD)	**Bias**	**Std**	**Mean(SD)**	**Bias**	**Std**
Standard Normal						
LS	—	-0.0177	0.0263	—	0.0329	0.2753
CQR_5	$0.9574_{(0.1699)}$	-0.0166	0.0271	$0.9376_{(0.3587)}$	0.0289	0.3019
CQR_9	$0.9783_{(0.1286)}$	-0.0165	0.0266	$0.9458_{(0.3092)}$	0.0283	0.3013
CQR_{19}	$0.9838_{(0.0815)}$	-0.0168	0.0266	$0.9491_{(0.2952)}$	0.0278	0.2962
Laplace						
LS	—	-0.0175	0.0249	—	0.0236	0.2718
CQR_5	$1.1938_{(0.3279)}$	-0.0145	0.0237	$1.2063_{(0.6794)}$	0.0106	0.2701
CQR_9	$1.1405_{(0.2523)}$	-0.0150	0.0243	$1.2046_{(0.6413)}$	0.0079	0.2719
CQR_{19}	$1.0857_{(0.1584)}$	-0.0157	0.0248	$1.2019_{(0.6035)}$	0.0098	0.2693
t-distribution with df $= 3$						
LS	—	-0.0167	0.0261	—	0.0025	0.3068
CQR_5	$1.5974_{(1.0324)}$	-0.0120	0.0229	$1.6099_{(1.7558)}$	0.0004	0.2503
CQR_9	$1.4247_{(0.8170)}$	-0.0132	0.0228	$1.5975_{(1.8047)}$	-0.0002	0.2560
CQR_{19}	$1.2111_{(0.4330)}$	-0.0140	0.0242	$1.5948_{(1.8291)}$	0.0006	0.2567
$.95N(0,1) + .05N(0,9)$						
LS	—	-0.0175	0.0247	—	-0.0130	0.2916
CQR_5	$1.1788_{(0.6248)}$	-0.0157	0.0228	$1.2268_{(2.0608)}$	-0.0050	0.2778
CQR_9	$1.1507_{(0.4715)}$	-0.0157	0.0230	$1.2132_{(1.8791)}$	-0.0048	0.2754
CQR_{19}	$1.0835_{(0.2603)}$	-0.0159	0.0234	$1.2104_{(1.8546)}$	-0.0066	0.2742
$.95N(0,1) + .05N(0,100)$						
LS	—	-0.0162	0.0260	—	0.0335	0.3728
CQR_5	$3.1661_{(2.4820)}$	-0.0077	0.0173	$3.0593_{(5.6699)}$	0.0245	0.2420
CQR_9	$2.4179_{(1.7012)}$	-0.0080	0.0171	$3.0287_{(5.3433)}$	0.0209	0.2533
CQR_{19}	$1.3469_{(0.5075)}$	-0.0085	0.0241	$3.0146_{(5.2728)}$	0.0234	0.2452

two possible outlier observations: $(2.7902, -2.5207)$ and $(2.8063, -2.6105)$ (circled in the plot). To understand the impact of these two possible outliers, we re-evaluated the local CQR and the local least squares estimates after excluding these two possible outliers. The resulting estimates are depicted in the top panel of Figure 3.3, from which we can see that the local CQR estimate remains almost the same, while the local least squares estimate changes a lot. We also note that after removing these two possible outliers, the local least squares estimator becomes very close to the local CQR estimator. Furthermore, as a more extreme demonstration, we kept these two possible outliers in the data set and moved them to more extreme cases, i.e, we moved $(2.7902, -2.5207)$ and $(2.8063, -2.6105)$ to $(2.7902, -6.5207)$ and $(2.8063, -6.6105)$, respectively. After perturbing (distorting) the two observations, we re-calculated the local CQR and the local least squares estimate. The resulting estimates are depicted in the bottom panel of Figure 3.3, which clearly demonstrates that the local least squares estimate changes dramatically. In contrast, the local CQR estimate is nearly un-affected by the artificial data distortion.

3.5 Local p-polynomial CQR smoothing and proofs

In this section we establish asymptotic theory of the local p-polynomial CQR estimators. We then treat Theorems 3.1 and 3.3 as two special cases of the general theory. As a generalization of the local linear and local quadratic CQR estimators, the local p-polynomial CQR estimator is constructed by minimizing

$$\sum_{k=1}^{q}\left[\sum_{i=1}^{n}\rho_{\tau_k}\Big\{y_i - a_k - \sum_{j=1}^{p} b_j(t_i - t_0)^j\Big\}K\left(\frac{t_i - t_0}{h}\right)\right], \tag{3.24}$$

and the local p-polynomial CQR estimators of $m(t_0)$ and $m^{(r)}(t_0)$ are given by

$$\hat{m}(t_0) = \frac{1}{q}\sum_{k=1}^{q}\hat{a}_k, \quad \text{and} \quad \hat{m}^{(r)}(t_0) = r!\hat{b}_r, \quad r = 1, \cdots, p. \tag{3.25}$$

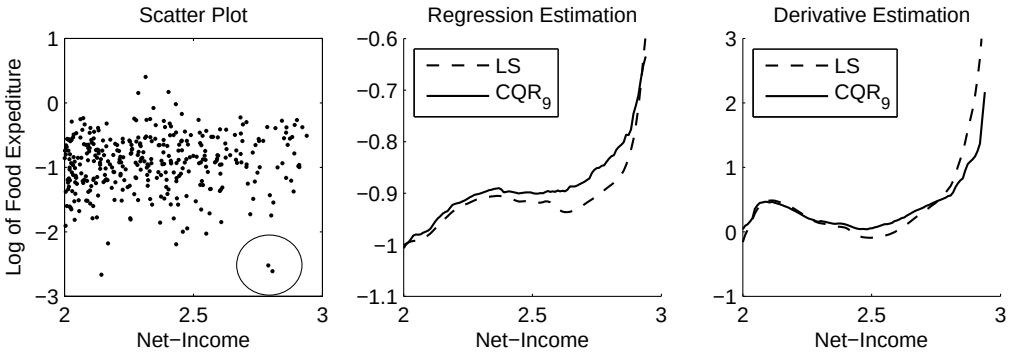


Fig. 3.2. The left panel is the scatter plot of data, the middle panel is the estimated regression function, and the right panel is the estimated derivative function.

For the asymptotic analysis, we need the following regularity conditions:

(A) The regression function $m(t)$ has a continuous $(p+2)^{th}$ derivative in the neighborhood of t_0.

(B) The marginal density function $f_T(\cdot)$ of T is differentiable and positive in the neighborhood of t_0.

(C) The conditional variance $\sigma^2(t)$ is continuous in the neighborhood of t_0.

(D) Assume that the error has a symmetric distribution with density $f(\cdot)$, and $f(\cdot)$ is positive in the neighborhoods of $\{c_k\}$.

We choose the kernel function K such that K is a symmetric density function with finite support $[-M, M]$. The following notation is needed to present the asymptotic properties of the local p-polynomial CQR estimator. Let S_{11} be a $q \times q$ diagonal matrix with diagonal elements $f(c_k)$, $k = 1, \cdots, q$; S_{12} a $q \times p$ matrix

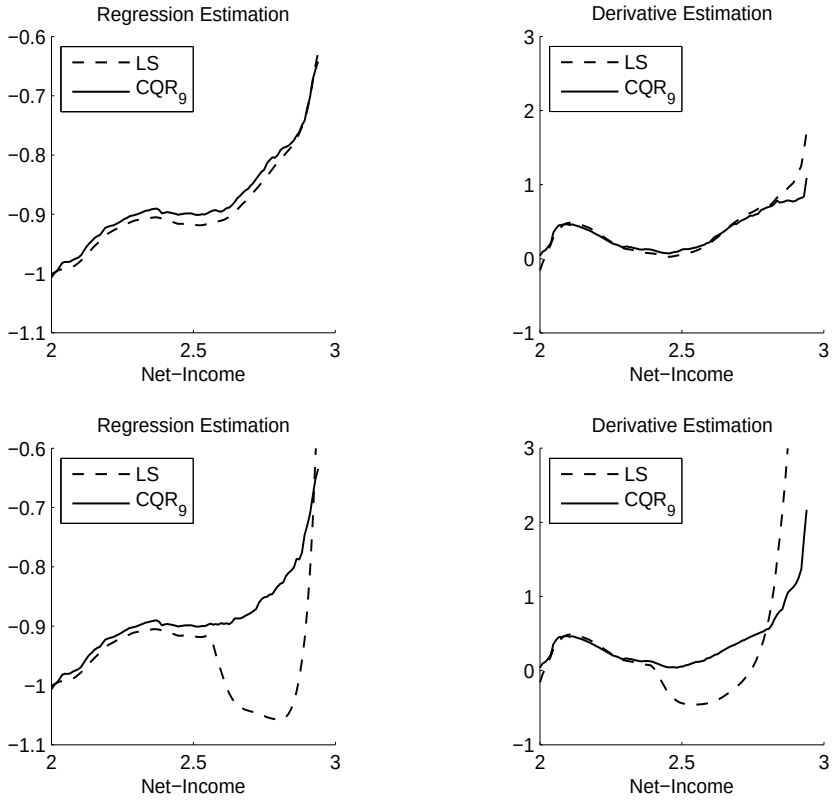


Fig. 3.3. Plot of estimated regression function and its derivative. The top panel is for the estimate removing the two possible outliers, and bottom panel is for the estimate moving the two possible outliers to more extreme cases. The left panel is for the estimated regression function, and the right panel is the estimated derivative function.

with (k,j)-element $f(c_k)\mu_j$, $k=1,\cdots,q$ and $j=1,\cdots,p$; $S_{21}=S_{12}^T$; and S_{22} a $p\times p$ matrix with (j,j')-element $\sum_{k=1}^{q} f(c_k)\mu_{j+j'}$, for $j,j'=1,\cdots,p$. Similarly, Let Σ_{11} be a $q\times q$ matrix with (k,k')-element $\nu_0\tau_{kk'}$, $k,k'=1,\cdots,q$; Σ_{12} a $q\times p$ matrix with (k,j)-element $\nu_j\sum_{k'=1}^{q}\tau_{kk'}$, $k=1,\cdots,q$ and $j=1,\cdots,p$; $\Sigma_{21}=\Sigma_{12}^T$; and Σ_{22} a $p\times p$ matrix with (j,j')-element $(\sum_{k,k'=1}^{q}\tau_{kk'})\nu_{j+j'}$, for $j,j'=1,\cdots,p$. Define

$$S=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix}, \quad \text{and} \quad \Sigma=\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}.$$

Partition S^{-1} into four submatrices as follows

$$S^{-1}=\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22}\end{pmatrix}^{-1}=\begin{pmatrix} (S^{-1})_{11} & (S^{-1})_{12} \\ (S^{-1})_{21} & (S^{-1})_{22}\end{pmatrix},$$

where hereafter, we use $(\cdot)_{11}$ to denote the left-top $q\times q$ submatrix and use $(\cdot)_{22}$ to denote the right-bottom $p\times p$ submatrix.

Furthermore, let $u_k=\sqrt{nh}\{a_k-m(t_0)-\sigma(t_0)c_k\}$ and $v_j=h^j\sqrt{nh}\{j!b_j-m^{(j)}(t_0)\}/j!$. Let $x_i=(t_i-t_0)/h$, $K_i=K(x_i)$ and $\Delta_{i,k}=\dfrac{u_k}{\sqrt{nh}}+\sum_{j=1}^{p}\dfrac{v_j x_i^j}{\sqrt{nh}}$. Write $d_{i,k}=c_k[\sigma(t_i)-\sigma(t_0)]+r_{i,p}$ with $r_{i,p}=m(t_i)-\sum_{j=0}^{p}m^{(j)}(t_0)(t_i-t_0)^j/j!$. Define $\eta_{i,k}^*$ to be $I(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)})-\tau_k$. let $W_n^*=(w_{11}^*,\cdots,w_{1q}^*,w_{21}^*,\cdots,w_{2p}^*)^T$ with $w_{1k}^*=\frac{1}{\sqrt{nh}}\sum_{i=1}^{n}K_i\eta_{i,k}^*$ and $w_{2j}^*=\frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n}K_i x_i^j\eta_{i,k}^*$.

The asymptotic properties of the local p-polynomial CQR estimator are based on the following theorem.

Theorem 3.5. *Let $\hat{\theta}_n=(\hat{u}_1,\cdots,\hat{u}_q,\hat{v}_1,\cdots,\hat{v}_p)$ be the minimizer of (3.24). Then under the regularity conditions (A)—(C), we have*

$$\hat{\theta}_n+\frac{\sigma(t_0)}{f_T(t_0)}S^{-1}E(W_n^*|\mathbf{T})\xrightarrow{\mathcal{L}}MVN\left(\mathbf{0},\frac{\sigma^2(t_0)}{f_T(t_0)}S^{-1}\Sigma S^{-1}\right).$$

To prove theorem 3.5, we first establish Lemmas 3.6—3.7.

Lemma 3.6. *Minimizing (3.24) is equivalent to minimizing*

$$\sum_{k=1}^{q} u_k \left(\sum_{i=1}^{n} \frac{K_i \eta_{i,k}^*}{\sqrt{nh}} \right) + \sum_{j=1}^{p} v_j \left(\sum_{k=1}^{q} \sum_{i=1}^{n} \frac{K_i x_i^j \eta_{i,k}^*}{\sqrt{nh}} \right) + \sum_{k=1}^{q} B_{n,k}(\theta)$$

with respect to $\theta = (u_1, \cdots, u_q, v_1, \cdots, v_p)^T$, *where*

$$B_{n,k}(\theta) = \sum_{i=1}^{n} \left\{ K_i \int_0^{\Delta_{i,k}} \left[I\left(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)} + \frac{z}{\sigma(t_i)} \right) - I\left(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)} \right) \right] dz \right\}.$$

Proof. To apply the identity (Knight 1998)

$$\rho_\tau(x - y) - \rho_\tau(x) = y(I(x \le 0) - \tau) + \int_0^y \{I(x \le z) - I(x \le 0)\} dz, \qquad (3.26)$$

we write

$$\begin{aligned} & y_i - a_k - \sum_{j=1}^{p} b_j (t_i - t_0)^j = \sigma(t_i)\epsilon_i + m(t_i) - a_k - \sum_{j=1}^{p} b_j (t_i - t_0)^j \\ = \;& \sigma(t_i)(\epsilon_i - c_k) + \big(\sigma(t_i) - \sigma(t_0)\big) c_k + r_{i,p} - \frac{u_k}{\sqrt{nh}} - \sum_{j=1}^{p} \frac{v_j x_i^j}{\sqrt{nh}} \\ = \;& \sigma(t_i)(\epsilon_i - c_k) + d_{i,k} - \Delta_{i,k} \,. \end{aligned}$$

Minimizing (3.24) is equivalent to minimizing

$$L_n(\theta) = \sum_{i=1}^{n} \left\{ K_i \sum_{k=1}^{q} \left[\rho_{\tau_k}\big(\sigma(t_i)(\epsilon_i - c_k) + d_{i,k} - \Delta_{i,k}\big) - \rho_{\tau_k}\big(\sigma(t_i)(\epsilon_i - c_k) + d_{i,k}\big) \right] \right\}.$$

Using the identity (3.26) and with some straightforward calculations, it follows that

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n}\left\{K_i\sum_{k=1}^{q}\Delta_{i,k}\left[I(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)}) - \tau_k\right]\right\} \\ &\quad + \sum_{i=1}^{n}\left\{K_i\sum_{k=1}^{q}\int_0^{\Delta_{i,k}}\left[I(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)} + \frac{z}{\sigma(t_i)}) - I(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)})\right]dz\right\} \\ &= \sum_{k=1}^{q} u_k\left(\sum_{i=1}^{n}\frac{K_i\eta^*_{i,k}}{\sqrt{nh}}\right) + \sum_{j=1}^{p} v_j\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_i x_i^j\eta^*_{i,k}}{\sqrt{nh}}\right) + \sum_{k=1}^{q} B_{n,k}(\theta). \end{aligned}$$

This completes the proof. □

Let $S_{n,11}$ be a $q\times q$ diagonal matrix with diagonal elements $f(c_k)\sum_{i=1}^{n}\frac{K_i}{nh\sigma(t_i)}$, $k = 1,\cdots,q$; $S_{n,12}$ be a $q\times p$ matrix with (k,j)-element $f(c_k)\sum_{i=1}^{n}\frac{K_i x_i^j}{nh\sigma(t_i)}$, $j = 1,\cdots,p$; $S_{n,22}$ be a $p\times p$ matrix with (j,j') element $\sum_{k=1}^{q} f(c_k)\sum_{i=1}^{n}\frac{K_i x_i^{j+j'}}{nh\sigma(t_i)}$. Denote

$$S_n = \begin{pmatrix} S_{n,11} & S_{n,12} \\ S_{n,12}^T & S_{n,22}\end{pmatrix}.$$

Lemma 3.7. *Under Conditions (A)—(C),* $L_n(\theta) = \frac{1}{2}\theta^T S_n\theta + (W_n^*)^T\theta + o_p(1)$.

Proof. Write $L_n(\theta)$ as

$$\begin{aligned} L_n(\theta) &= \sum_{k=1}^{q} u_k\left(\sum_{i=1}^{n}\frac{K_i\eta^*_{i,k}}{\sqrt{nh}}\right) + \sum_{j=1}^{p} v_j\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_i x_i^j\eta^*_{i,k}}{\sqrt{nh}}\right) \\ &\quad + \sum_{k=1}^{q} E_\epsilon[B_{n,k}(\theta)|\mathbf{T}] + \sum_{k=1}^{q} R_{n,k}(\theta), \end{aligned}$$

where $R_{n,k}(\theta) = B_{n,k}(\theta) - E_\epsilon[B_{n,k}(\theta)|\mathbf{T}]$.

Using $F(c_k + z) - F(c_k) = zf(c_k) + o(z)$, then $\sum_{k=1}^{q} E_\epsilon[B_{n,k}(\theta)|\mathbf{T}]$ equals

$$\begin{aligned}
& \sum_{k=1}^{q}\sum_{i=1}^{n}\left[K_i\int_0^{\Delta_{i,k}}\left\{\frac{z}{\sigma(t_i)}f\left(c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)+o(z)\right\}dz\right] \\
= & \sum_{k=1}^{q}\sum_{i=1}^{n}\left[K_i\Delta_{i,k}^2\frac{f(c_k-\frac{d_{i,k}}{\sigma(t_i)})}{2\sigma(t_i)}\right]+o_p(1) \\
= & \sum_{k=1}^{q}\sum_{i=1}^{n}\left[K_i\Delta_{i,k}^2\frac{f(c_k)}{2\sigma(t_i)}\right]+o_p(1) \\
= & \frac{1}{2}\theta^T S_n\theta+o_p(1).
\end{aligned}$$

We now prove $R_{n,k}(\theta) = o_p(1)$. It is sufficient to show $Var_\epsilon[B_{n,k}(\theta)|\mathbf{T}] = o_p(1)$.

$$\begin{aligned}
& Var_\epsilon[B_{n,k}(\theta)|\mathbf{T}] \\
= & Var_\epsilon\left[\sum_{i=1}^{n}\left\{K_i\int_0^{\Delta_{i,k}}\left[I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{z}{\sigma(t_i)}\right)-I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right]dz\right\}|\mathbf{T}\right] \\
= & \sum_{i=1}^{n}Var_\epsilon\left[\left\{K_i\int_0^{\Delta_{i,k}}\left[I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{z}{\sigma(t_i)}\right)-I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right]dz\right\}|\mathbf{T}\right] \\
\le & \sum_{i=1}^{n}E_\epsilon\left[\left\{K_i\int_0^{\Delta_{i,k}}\left[I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{z}{\sigma(t_i)}\right)-I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right]dz\right\}^2|\mathbf{T}\right] \\
= & \sum_{i=1}^{n}K_i^2\int_0^{\Delta_{i,k}}\int_0^{\Delta_{i,k}}E_\epsilon\left[\left\{I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{z_1}{\sigma(t_i)}\right)-I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right\}\right. \\
& \left.\left\{I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{z_2}{\sigma(t_i)}\right)-I\left(\epsilon_i\le c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right\}|\mathbf{T}\right]dz_1dz_2 \\
\le & \sum_{i=1}^{n}K_i^2\int_0^{|\Delta_{i,k}|}\int_0^{|\Delta_{i,k}|}\left[F\left(c_k-\frac{d_{i,k}}{\sigma(t_i)}+\frac{|\Delta_{i,k}|}{\sigma(t_i)}\right)-F\left(c_k-\frac{d_{i,k}}{\sigma(t_i)}\right)\right]dz_1dz_2 \\
= & o\left(\sum_{i=1}^{n}K_i^2\Delta_{i,k}^2\right)=o_p(1)
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.5. Similar to Parzen (1962), we have $\frac{1}{nh}\sum_{i=1}^{n} K_i x_i^j \xrightarrow{P} f_T(t_0)\mu_j$, where $\xrightarrow{P}$ stands for convergence in probability. Thus,

$$S_n \xrightarrow{P} \frac{f_T(t_0)}{\sigma(t_0)} S = \frac{f_T(t_0)}{\sigma(t_0)} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

This, together with Lemmas 3.6, 3.7, leads to

$$L_n(\theta) = \frac{1}{2}\frac{f_T(t_0)}{\sigma(t_0)}\theta^T S\theta + (W_n^*)^T\theta + o_p(1).$$

Since the convex function $L_n(\theta) - (W_n^*)^T\theta$ converges in probability to the convex function $\frac{1}{2}\frac{f_T(t_0)}{\sigma(t_0)}\theta^T S\theta$, it follows from the convexity lemma (Pollard 1991) that for any compact set Θ, the quadratic approximation to $L_n(\theta)$ holds uniformly for θ in Θ, which leads to

$$\hat{\theta}_n = -\frac{\sigma(t_0)}{f_T(t_0)} S^{-1} W_n^* + o_p(1).$$

Denote $\eta_{i,k} = I(\epsilon_i \le c_k) - \tau_k$ and $W_n = (w_{11}, \cdots, w_{1q}, w_{21}, \cdots, w_{2p})^T$ with $w_{1k} = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_i \eta_{i,k}$ and $w_{2j} = \frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n} K_i x_i^j \eta_{i,k}$. By the Cramér-Wald theorem, it is easy to see that the CLT for $W_n|\mathbf{T}$ holds:

$$\frac{W_n|\mathbf{T} - E[W_n|\mathbf{T}]}{\sqrt{Var[W_n|\mathbf{T}]}} \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, I_{(p+q)\times(p+q)}). \tag{3.27}$$

Note that

$$Cov(\eta_{i,k}, \eta_{i,k'}) = \tau_{kk'}, \qquad Cov(\eta_{i,k}, \eta_{j,k'}) = 0, \quad if \quad i \neq j.$$

And similar to Parzen (1962), we have $\frac{1}{nh}\sum_{i=1}^{n} K_i^2 x_i^j \xrightarrow{P} f_T(t_0)\nu_j$, Therefore, $Var[W_n|\mathbf{T}] \xrightarrow{P} f_T(t_0)\Sigma$. Combined with (3.27), we have

$$W_n|\mathbf{T} \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, f_T(t_0)\Sigma).$$

Moreover, we have

$$Var(w^*_{1k} - w_{1k}|\mathbf{T}) = \frac{1}{nh}\sum_{i=1}^{n} K_i^2 Var(\eta^*_{i,k} - \eta_{i,k}) \leq \frac{1}{nh}\sum_{i=1}^{n} K_i^2 \{F(c_k + \frac{|d_{i,k}|}{\sigma(t_i)}) - F(c_k)\} = o_p(1)$$

and also

$$\begin{aligned} Var(w^*_{2j} - w_{2j}|\mathbf{T}) &= \frac{1}{nh}\sum_{i=1}^{n} K_i^2 x_i^j Var(\sum_{k=1}^{q} \eta^*_{i,k} - \eta_{i,k}) \\ &\leq \frac{q^2}{nh}\sum_{i=1}^{n} K_i^2 x_i^j \max_k \{F(c_k + \frac{|d_{i,k}|}{\sigma(t_i)}) - F(c_k)\} = o_p(1). \end{aligned}$$

Thus

$$Var(W^*_n - W_n|\mathbf{T}) = o_p(1).$$

So by Slutsky's theorem, conditioning on $\mathbf{T}$, we have

$$W^*_n|\mathbf{T} - E(W^*_n|\mathbf{T}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, f_T(t_0)\Sigma).$$

Therefore,

$$\hat{\theta}_n + \frac{\sigma(t_0)}{f_T(t_0)} S^{-1} E(W^*_n|\mathbf{T}) \xrightarrow{\mathcal{L}} MVN\left(\mathbf{0}, \frac{\sigma^2(t_0)}{f_T(t_0)} S^{-1}\Sigma S^{-1}\right). \tag{3.28}$$

This completes the proof. □

Proof of Theorem 3.1. The asymptotic normality follows Theorem 3.5 with $p = 1$. Let us calculate the conditional bias and variance, respectively. Denote by $e_{q\times 1}$ the vector that contains q 1's. When $p = 1$, S is a diagonal matrix with diagonal elements $f(c_1), \cdots, f(c_q), \mu_2 \sum_{k=1}^{q} f(c_k)$. So the asymptotic conditional bias of

$\hat{m}(t_0) = \frac{1}{q}\sum_{k=1}^{q} \hat{a}_k$ is

$$
\begin{aligned}
Bias(\hat{m}(t_0)|\mathbf{T}) &= \frac{1}{q}\sigma(t_0)\sum_{k=1}^{q} c_k - \frac{1}{q\cdot\sqrt{nh}}\frac{\sigma(t_0)}{f_T(t_0)} e^T_{q\times 1}(S^{-1})_{11} E(W^*_{1n}|\mathbf{T}) \\
&= \frac{1}{q}\sigma(t_0)\sum_{k=1}^{q} c_k - \frac{1}{q\cdot nh}\frac{\sigma(t_0)}{f_T(t_0)}\sum_{i=1}^{n} K_i \sum_{k=1}^{q}\frac{1}{f(c_k)}\left\{F\left(c_k - \frac{d_{i,k}}{\sigma(t_i)}\right) - F(c_k)\right\}.
\end{aligned}
$$

Note that the error is symmetric, thus $\sum_{k=1}^{q} c_k = 0$, and furthermore, it is easy to check that $\frac{1}{q}\sum_{k=1}^{q}\frac{1}{f(c_k)}\{F(c_k - \frac{d_{i,k}}{\sigma(t_i)}) - F(c_k)\} = -\frac{r_{i,p}}{\sigma(t_i)}\{1 + o_p(1)\}$. Therefore,

$$
Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{nh}\frac{\sigma(t_0)}{f_T(t_0)}\sum_{i=1}^{n} K_i \frac{r_{i,p}}{\sigma(t_i)}\{1 + o_p(1)\}.
$$

By using the fact that

$$
\frac{1}{nh}\sum_{i=1}^{n} K_i \frac{r_{i,p}}{\sigma(t_i)} = \frac{f_T(t_0)m''(t_0)}{2\sigma(t_0)}\mu_2 h^2\{1 + o_p(1)\},
$$

we obtain

$$
Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{2}m''(t_0)\mu_2 h^2 + o_p(h^2). \tag{3.29}
$$

Furthermore, the conditional variance of $\hat{m}(t_0)$ is

$$
\begin{aligned}
Var(\hat{m}(t_0)|\mathbf{T}) &= \frac{1}{nh}\frac{\sigma^2(t_0)}{f_T(t_0)}\frac{1}{q^2} e^T_{q\times 1}(S^{-1}\Sigma S^{-1})_{11} e_{q\times 1} + o_p\left(\frac{1}{nh}\right) \\
&= \frac{1}{nh}\frac{\nu_0\sigma^2(t_0)}{f_T(t_0)} R_1(q) + o_p\left(\frac{1}{nh}\right), \qquad (3.30)
\end{aligned}
$$

which completes the proof. □

By using Theorem 3.5, we can further derive the asymptotic bias and variance of $\tilde{m}'(t_0)$ given in (3.4):

$$
Bias(\tilde{m}'(t_0)|\mathbf{T}) = \frac{1}{6}\left(m'''(t_0) + 3m''(t_0)\frac{f'_T(t_0)}{f_T(t_0)}\right)\frac{\mu_4}{\mu_2}h^2 + o_p(h^2), \tag{3.31}
$$

$$Var(\tilde{m}'(t_0)|\mathbf{T}) = \frac{1}{nh^3}\frac{\nu_2\sigma^2(t_0)}{\mu_2^2 f_T(t_0)}R_2(q) + o_p\left(\frac{1}{nh^3}\right). \tag{3.32}$$

Proof of Theorem 3.2. Note that

$$\begin{aligned}\lim_{q\to\infty} R_1(q) &= \int_0^1\int_0^1 \frac{s_1 \wedge s_2 - s_1 s_2}{f(F^{-1}(s_1))f(F^{-1}(s_2))}ds_1 ds_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(F(z_1)\wedge F(z_2) - F(z_1)F(z_2)\Big)dz_1 dz_2. \end{aligned}\tag{3.33}$$

by change of variables. Define functions

$$G(s) = \int_{-\infty}^{s} F(t)dt, \quad H(s) = \int_{-\infty}^{s} G(t)dt.$$

We have

$$\begin{aligned} G(s) &= \int_{-\infty}^{s}\Big(\int_{-\infty}^{t} f(x)dx\Big)dt = \int_{-\infty}^{s}\Big(\int_{x}^{s} f(x)dt\Big)dx \\ &= \int_{-\infty}^{s}(s-x)f(x)dx = sF(s) - k_1(s), \end{aligned}\tag{3.34}$$

where $k_1(s) = \int_{-\infty}^{s} xf(x)dx$. Similarly, we obtain

$$\begin{aligned} 2H(s) &= 2\int_{-\infty}^{s}\Big(\int_{-\infty}^{t}(t-x)f(x)dx\Big)dt = \int_{-\infty}^{s}\Big(\int_{x}^{s}2(t-x)f(x)dt\Big)dx \\ &= \int_{-\infty}^{s}(s-x)^2 f(x)dx = s^2F(s) - 2sk_1(s) + k_2(s), \end{aligned}\tag{3.35}$$

where $k_2(s) = \int_{-\infty}^{s} x^2 f(x)dx$. For the integral in (3.33) we have

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(F(z_1)\wedge F(z_2) - F(z_1)F(z_2)\Big)dz_1dz_2 \\
= \;& 2\int_{-\infty}^{\infty}\Big(1-F(z_1)\Big)\Big(\int_{-\infty}^{z_1}F(z_2)dz_2\Big)dz_1 \\
= \;& 2\int_{-\infty}^{\infty}\Big(\int_{z_1}^{\infty}f(t)dt\Big)G(z_1)dz_1 \\
= \;& 2\int_{-\infty}^{\infty}f(t)\Big(\int_{-\infty}^{t}G(z_1)dz_1\Big)dt \\
= \;& \int_{-\infty}^{\infty}2f(t)H(t)dt. \qquad (3.36)
\end{aligned}$$

By the definition of G and H, we know $\frac{d(2H(t)F(t)-G^2(t))}{dt} = 2H(t)f(t)$; and combining (3.34) and (3.35) yields $2H(t)F(t)-G^2(t) = k_2(t)F(t)-k_1^2(t)$. Now it is easy to see that the integral in (3.36) equals 1, by the facts that $\int_{-\infty}^{\infty}x^2f(x)dx = E_F[\epsilon^2] = 1$ and $\int_{-\infty}^{\infty}xf(x)dx = E_F[\epsilon] = 0$. □

Proof of Theorem 3.3. We apply Theorem 3.5 to get the asymptotic normality. Denote by e_r the p-vector $(0,0,\cdots,1,0,\cdots,0)^T$ with 1 in the r^{th} position. When $p=2$, S_{12} and S_{22} have the following forms:

$$S_{12} = \begin{pmatrix}\mathbf{0}_{q\times 1} & \mu_2\Big(f(c_k)\Big)_{q\times 1}\end{pmatrix}, S_{22} = \begin{pmatrix}\mu_2\sum_{k=1}^{q}f(c_k) & 0\\ 0 & \mu_4\sum_{k=1}^{q}f(c_k)\end{pmatrix}.$$

Thus,

$$(S^{-1})_{22} = (S_{22}-S_{21}S_{11}^{-1}S_{12})^{-1} = \begin{pmatrix}\dfrac{1}{\mu_2\sum_{k=1}^{q}f(c_k)} & 0\\ 0 & \dfrac{1}{(\mu_4-\mu_2^2)\sum_{k=1}^{q}f(c_k)}\end{pmatrix},$$

$$(S^{-1})_{21} = -(S^{-1})_{22}S_{21}S_{11}^{-1} = \begin{pmatrix}\mathbf{0}_{1\times q}\\ \Big(\dfrac{\mu_2}{(\mu_4-\mu_2^2)\sum_{k=1}^{q}f(c_k)}\Big)_{1\times q}\end{pmatrix},$$

since $S_{11} = \text{diag}\left(f(c_1), \cdots, f(c_q)\right)$. By Theorem 3.5,

$$\begin{aligned} Bias(\hat{m}'(t_0)|\mathbf{T}) &= -\frac{\sigma(t_0)}{hf_T(t_0)}\frac{1}{\sqrt{nh}}e_1^T\left\{(S^{-1})_{21}E(W_{1n}^*|\mathbf{T}) + (S^{-1})_{22}E(W_{2n}^*|\mathbf{T})\right\} \\ &= -\frac{\sigma(t_0)}{hf_T(t_0)}\frac{1}{\mu_2\sum_{k=1}^q f(c_k)}\frac{1}{\sqrt{nh}}E(w_{21}^*|\mathbf{T}). \end{aligned}$$

Note that

$$E(w_{2j}^*|\mathbf{T}) = \frac{1}{\sqrt{nh}}\sum_{i=1}^n K_i x_i^j \sum_{k=1}^q \left\{F\left(c_k - \frac{d_{i,k}}{\sigma(t_i)}\right) - F(c_k)\right\}$$

Similarly, under condition (D), we have $\sum_{k=1}^q\{F(c_k - \frac{d_{i,k}}{\sigma(t_i)}) - F(c_k)\} = -\sum_{k=1}^q f(c_k)\cdot \frac{r_{i,p}}{\sigma(t_i)}\{1 + o_p(1)\}$. Therefore, $Bias(\hat{m}'(t_0)|\mathbf{T})$ is equal to $\frac{1}{nh^2}\frac{\sigma(t_0)}{f_T(t_0)}\sum_{i=1}^n K_i x_i \frac{r_{i,p}}{\sigma(t_i)}\{1 + o_p(1)\}$. Still using the fact that with $p = 2$

$$\frac{1}{nh}\sum_{i=1}^n K_i x_i \frac{r_{i,p}}{\sigma(t_i)} = \frac{f_T(t_0)m'''(t_0)}{6\sigma(t_0)}\frac{\mu_4}{\mu_2}h^3\{1 + o_p(1)\},$$

we obtain

$$Bias(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}h^2 + o_p(h^2). \tag{3.37}$$

Furthermore, the conditional variance of $\hat{m}(t_0)$ is

$$\begin{aligned} Var(\hat{m}'(t_0)|\mathbf{T}) &= \frac{1}{nh^3}\frac{\sigma^2(t_0)}{f_T(t_0)}e_1^T(S^{-1}\Sigma S^{-1})_{22}e_1 + o_p(\frac{1}{nh^3}) \\ &= \frac{1}{nh^3}\frac{\nu_2\sigma^2(t_0)}{\mu_2^2 f_T(t_0)}R_2(q) + o_p(\frac{1}{nh^3}), \end{aligned} \tag{3.38}$$

which completes the proof. □

Proof of Theorem 3.4. From Zou and Yuan (2008), we know that

$$\lim_{q\to\infty} \frac{\Big(\sum_{k=1}^{q} f(c_k)\Big)^2}{\sum_{k=1}^{q}\sum_{k'=1}^{q} \tau_{kk'}} = 12E_F^2[f(\epsilon)] = 12\left(\int f^2(x)dx\right)^2.$$

Thus

$$\lim_{q\to\infty} \frac{1}{R_2(q)} = 12\left(\int f^2(x)dx\right)^2.$$

We notice that $12\left(\int f^2(x)dx\right)^2$ is also the asymptotic Pitman efficiency of the Wilcoxon test relative to the t-test (Hodges and Lehmann 1956). For the rest of the proof, readers are referred to Hodges and Lehmann (1956). □

3.6 Infinite variance case

Suppose that

$$Y = m(T) + \epsilon,$$

where ϵ has a density f with mean 0 and variance infinity.

Suppose that t_0 is an interior point of the support of $f_T(\cdot)$. Note that the local p-polynomial CQR estimator is constructed by minimizing

$$\sum_{k=1}^{q}\left[\sum_{i=1}^{n} \rho_{\tau_k}\Big\{y_i - a_k - \sum_{j=1}^{p} b_j(t_i - t_0)^j\Big\}K(\frac{t_i - t_0}{h})\right], \tag{3.39}$$

and the local p-polynomial CQR estimators of $m(t_0)$ and $m^{(r)}(t_0)$ are given by

$$\hat{m}(t_0) = \frac{1}{q}\sum_{k=1}^{q}\hat{a}_k, \quad \text{and} \quad \hat{m}^{(r)}(t_0) = r!\hat{b}_r, \quad r = 1, \cdots, p. \tag{3.40}$$

Let $u_k = \sqrt{nh}\{a_k - m(t_0) - c_k\}$, $v_j = h^j\sqrt{nh}\{j!b_j - m^{(j)}(t_0)\}/j!$. Let $x_i = (t_i - t_0)/h$, $K_i = K(x_i)$ and $\Delta_{i,k} = \dfrac{u_k}{\sqrt{nh}} + \sum_{j=1}^{p} \dfrac{v_j x_i^j}{\sqrt{nh}}$. Write $r_{i,p} = m(t_i) - \sum_{j=0}^{p} m^{(j)}(t_0)(t_i - t_0)^j/j!$. Define $\eta_{i,k}^*$ to be $I(\epsilon_i \le c_k - r_{i,p}) - \tau_k$. let $W_n^* = (w_{11}^*, \cdots, w_{1q}^*, w_{21}^*, \cdots, w_{2p}^*)^T$ with $w_{1k}^* = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_i\eta_{i,k}^*$ and $w_{2j}^* = \frac{1}{\sqrt{nh}}\sum_{k=1}^{q} \sum_{i=1}^{n} K_i x_i^j \eta_{i,k}^*$. The asymptotic properties of the local p-polynomial CQR estimator are based on the following theorem.

Theorem 3.8. *Let $\hat{\theta}_n = (\hat{u}_1, \cdots, \hat{u}_q, \hat{v}_1, \cdots, \hat{v}_p)$ be the minimizer of (3.39). Assume that $f_T(t_0) > 0$, $f_T(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of t_0 and $f(\cdot)$ is positive in the neighborhoods of $\{\tau_k\}$. Then we have*

$$\hat{\theta}_n + \frac{1}{f_T(t_0)} S^{-1} E(W_n^* | \mathbf{T}) \xrightarrow{\mathcal{L}} MVN\left(\mathbf{0}, \frac{1}{f_T(t_0)} S^{-1}\Sigma S^{-1}\right).$$

Proof. To apply the identity

$$\rho_\tau(x - y) - \rho_\tau(x) = y(I(x \le 0) - \tau) + \int_0^y [I(x \le z) - I(x \le 0)]dz, \qquad (3.41)$$

we write

$$\begin{aligned} & y_i - a_k - \sum_{j=1}^{p} b_j(t_i - t_0)^j = \epsilon_i + m(t_i) - a_k - \sum_{j=1}^{p} b_j(t_i - t_0)^j \\ = \;& (\epsilon_i - c_k) + r_{i,p} - \frac{u_k}{\sqrt{nh}} - \sum_{j=1}^{p} \frac{v_j x_i^j}{\sqrt{nh}} \\ = \;& (\epsilon_i - c_k) + r_{i,p} - \Delta_{i,k}\,, \end{aligned}$$

Minimizing (3.39) is equivalent to minimizing

$$L_n(\theta) = \sum_{i=1}^{n}\left\{K_i \sum_{k=1}^{q}\left[\rho_{\tau_k}\big((\epsilon_i - c_k) + r_{i,p} - \Delta_{i,k}\big) - \rho_{\tau_k}\big((\epsilon_i - c_k) + r_{i,p}\big)\right]\right\}.$$

Using the identity (3.41) and with some straightforward calculations, it follows that

$$\begin{aligned} L_n(\theta) &= \sum_{i=1}^{n}\left\{K_i\sum_{k=1}^{q}\Delta_{i,k}\left[I(\epsilon_i \le c_k - r_{i,p}) - \tau_k\right]\right\} \\ &\quad + \sum_{i=1}^{n}\left\{K_i\sum_{k=1}^{q}\int_0^{\Delta_{i,k}}\left[I(\epsilon_i \le c_k - r_{i,p}+z) - I(\epsilon_i \le c_k - r_{i,p})\right]dz\right\} \\ &= \sum_{k=1}^{q}u_k\left(\sum_{i=1}^{n}\frac{K_i\eta_{i,k}^*}{\sqrt{nh}}\right) + \sum_{j=1}^{p}v_j\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_i x_i^j\eta_{i,k}^*}{\sqrt{nh}}\right) + \sum_{k=1}^{q}B_{n,k}(\theta), \end{aligned}$$

where

$$B_{n,k}(\theta) = \sum_{i=1}^{n}\left\{K_i\int_0^{\Delta_{i,k}}\left[I(\epsilon_i \le c_k - r_{i,p}+z) - I(\epsilon_i \le c_k - r_{i,p})\right]dz\right\}.$$

Let $S_{n,11}$ be a $q\times q$ diagonal matrix with diagonal elements $f(c_k)\sum_{i=1}^{n}K_i/nh$, $k = 1,\cdots,q$; $S_{n,12}$ be a $q\times p$ matrix with (k,j)-element $f(c_k)\sum_{i=1}^{n}K_i x_i^j/nh$, $j = 1,\cdots,p$; $S_{n,22}$ be a $p\times p$ matrix with (j,j') element $\sum_{k=1}^{q}f(c_k)\sum_{i=1}^{n}K_i x_i^{j+j'}/nh$. Denote

$$S_n = \begin{pmatrix} S_{n,11} & S_{n,12} \\ S_{n,12}^T & S_{n,22}\end{pmatrix}.$$

We write $L_n(\theta)$ as

$$\begin{aligned} L_n(\theta) &= \sum_{k=1}^{q}u_k\left(\sum_{i=1}^{n}\frac{K_i\eta_{i,k}^*}{\sqrt{nh}}\right) + \sum_{j=1}^{p}v_j\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{K_i x_i^j\eta_{i,k}^*}{\sqrt{nh}}\right) \\ &\quad + \sum_{k=1}^{q}E_\epsilon[B_{n,k}(\theta)|\mathbf{T}] + \sum_{k=1}^{q}R_{n,k}(\theta), \end{aligned}$$

where $R_{n,k}(\theta) = B_{n,k}(\theta) - E_\epsilon[B_{n,k}(\theta)|\mathbf{T}]$.

By similar arguments, we can show that $\sum_{k=1}^{q} E_{\epsilon}[B_{n,k}(\theta)|\mathbf{T}] = \frac{1}{2}\theta^T S_n \theta + o_p(1)$ and $R_{n,k}(\theta) = o_p(1)$. Together with $\sum_{i=1}^{n} K_i x_i^j / nh \xrightarrow{P} f_T(t_0)\mu_j$ and

$$S_n \xrightarrow{P} f_T(t_0) S = f_T(t_0) \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

we have

$$L_n(\theta) = \frac{f_T(t_0)}{2}\theta^T S\theta + (W_n^*)^T\theta + o_p(1).$$

Since the convex function $L_n(\theta) - (W_n^*)^T\theta$ converges in probability to the convex function $\frac{f_T(t_0)}{2}\theta^T S\theta$, it follows from the convexity lemma that for any compact set Θ, the quadratic approximation to $L_n(\theta)$ holds uniformly for θ in Θ, which leads to

$$\hat{\theta}_n = -\frac{1}{f_T(t_0)} S^{-1} W_n^* + o_p(1).$$

Denote $\eta_{i,k} = I(\epsilon_i \le c_k) - \tau_k$ and $W_n = (w_{11}, \cdots, w_{1q}, w_{21}, \cdots, w_{2p})^T$ with $w_{1k} = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_i \eta_{i,k}$ and $w_{2j} = \frac{1}{\sqrt{nh}} \sum_{k=1}^{q} \sum_{i=1}^{n} K_i x_i^j \eta_{i,k}$. By the Cramér-Wald theorem, it is easy to see that the CLT for $W_n|\mathbf{T}$ holds:

$$\frac{W_n|\mathbf{T} - E[W_n|\mathbf{T}]}{\sqrt{Var[W_n|\mathbf{T}]}} \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, I_{(p+q)\times(p+q)}). \tag{3.42}$$

Note that

$$Cov(\eta_{i,k}, \eta_{i,k'}) = \tau_{kk'}, \qquad Cov(\eta_{i,k}, \eta_{j,k'}) = 0, \quad if \quad i \neq j.$$

and $\sum_{i=1}^{n} K_i^2 x_i^j / nh \xrightarrow{P} f_T(t_0)\nu_j$, Therefore, $Var[W_n|\mathbf{T}] \xrightarrow{P} f_T(t_0)\Sigma$. Combined with (3.42), we have

$$W_n|\mathbf{T} \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, f_T(t_0)\Sigma).$$

Moreover, we have $Var(w^*_{1k} - w_{1k}|\mathbf{T}) = \frac{1}{nh}\sum_{i=1}^n K_i^2 Var(\eta^*_{i,k} - \eta_{i,k}) \leq \frac{1}{nh}\sum_{i=1}^n K_i^2\{F(c_k + |r_{i,p}|) - F(c_k)\} = o_p(1)$ and also $Var(w^*_{2j} - w_{2j}|\mathbf{T}) = \frac{1}{nh}\sum_{i=1}^n K_i^2 x_i^j Var(\sum_{k=1}^q \eta^*_{i,k} - \eta_{i,k}) \leq \frac{q^2}{nh}\sum_{i=1}^n K_i^2 x_i^j \max_k\{F(c_k + |r_{i,p}|) - F(c_k)\} = o_p(1)$, thus

$$Var(W^*_n - W_n|\mathbf{T}) = o_p(1).$$

So by Slutsky's theorem, conditioning on $\mathbf{T}$, we have

$$W^*_n|\mathbf{T} - E(W^*_n|\mathbf{T}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, f_T(t_0)\Sigma).$$

Therefore,

$$\hat{\theta}_n + \frac{1}{f_T(t_0)} S^{-1} E(W^*_n|\mathbf{T}) \xrightarrow{\mathcal{L}} MVN(\mathbf{0}, \frac{1}{f_T(t_0)} S^{-1}\Sigma S^{-1}). \tag{3.43}$$

This completes the proof. □

The asymptotic properties of the local CQR estimators $\hat{m}(t)$ and $\hat{m}'(t)$ are two special cases of the general result.

Theorem 3.9. *Under the regularity conditions in Theorem 3.8, if the error ϵ follows a symmetric distribution and $h \to 0$, $nh \to \infty$ as $n \to \infty$, the asymptotic conditional bias and variance of the local linear CQR estimator $\hat{m}(t_0)$ are given by*

$$Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{2} m''(t_0)\mu_2 h^2 + o_p(h^2), \tag{3.44}$$

$$Var(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{nh}\frac{\nu_0}{f_T(t_0)} R_1(q) + o_p\left(\frac{1}{nh}\right). \tag{3.45}$$

Furthermore, conditioning on $\mathbf{T}$, we have

$$\sqrt{nh}\left\{\hat{m}(t_0) - m(t_0) - \frac{1}{2} m''(t_0)\mu_2 h^2\right\} \xrightarrow{\mathcal{L}} N\left(0, \frac{\nu_0}{f_T(t_0)} R_1(q)\right). \tag{3.46}$$

Remark. The symmetric error condition is only used to eliminate the non-vanishing bias term $\frac{1}{q}\sum_{k=1}^{q} c_k$. If this condition is removed, the asymptotic bias would be

$$Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{q}\sum_{k=1}^{q} c_k + \frac{1}{2}m''(t_0)\mu_2 h^2 + o_p(h^2), \tag{3.47}$$

Proof of Theorem 3.9. The asymptotic normality follows Theorem 3.8 with $p = 1$. Let us calculate the conditional bias and variance, respectively. Denote by $e_{q\times 1}$ the vector that contains q 1's. When $p = 1$, S is a diagonal matrix with diagonal elements $f(c_1), \cdots, f(c_q), \mu_2 \sum_{k=1}^{q} f(c_k)$. So the asymptotic conditional bias of $\hat{m}(t_0) = \frac{1}{q}\sum_{k=1}^{q} \hat{a}_k$ is

$$\begin{aligned} Bias(\hat{m}(t_0)|\mathbf{T}) &= \frac{1}{q}\sum_{k=1}^{q} c_k - \frac{1}{q\cdot\sqrt{nh}}\frac{1}{f_T(t_0)} e_{q\times 1}^T (S^{-1})_{11} E(W_{1n}^*|\mathbf{T}) \\ &= \frac{1}{q}\sum_{k=1}^{q} c_k - \frac{1}{q\cdot nh}\frac{1}{f_T(t_0)}\sum_{i=1}^{n} K_i \sum_{k=1}^{q}\frac{1}{f(c_k)}\{F(c_k - r_{i,p}) - F(c_k)\} \\ &= \frac{1}{nh}\frac{1}{f_T(t_0)}\sum_{i=1}^{n} K_i r_{i,p}\{1 + o_p(1)\}. \end{aligned}$$

By using the fact that

$$\frac{1}{nh}\sum_{i=1}^{n} K_i r_{i,p} = \frac{f_T(t_0)m''(t_0)}{2}\mu_2 h^2\{1 + o_p(1)\},$$

we obtain

$$Bias(\hat{m}(t_0)|\mathbf{T}) = \frac{1}{2}m''(t_0)\mu_2 h^2 + o_p(h^2). \tag{3.48}$$

Furthermore, the conditional variance of $\hat{m}(t_0)$ is

$$\begin{aligned} Var(\hat{m}(t_0)|\mathbf{T}) &= \frac{1}{nh}\frac{1}{f_T(t_0)}\frac{1}{q^2} e_{q\times 1}^T (S^{-1}\Sigma S^{-1})_{11} e_{q\times 1} + o_p(\frac{1}{nh}) \\ &= \frac{1}{nh}\frac{\nu_0}{f_T(t_0)} R_1(q) + o_p(\frac{1}{nh}), \end{aligned} \tag{3.49}$$

which completes the proof. □

Theorem 3.10. *Under the regularity conditions in Theorem 3.8, if* $h \to 0$, $nh^3 \to \infty$ *as* $n \to \infty$, *the asymptotic conditional bias and variance of* $\hat{m}'(t_0)$ *from local quadratic CQR are given by*

$$Bias(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}h^2 + o_p(h^2), \tag{3.50}$$

$$Var(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2 f_T(t_0)}R_2(q) + o_p(\frac{1}{nh^3}). \tag{3.51}$$

Furthermore, conditioning on $\mathbf{T}$, *we have the following asymptotic normal distribution*

$$\sqrt{nh^3}\left(\hat{m}'(t_0) - m'(t_0) - \frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}h^2\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{\nu_2}{\mu_2^2 f_T(t_0)}R_2(q)\right). \tag{3.52}$$

Proof of Theorem 3.10. We apply Theorem 3.8 to get the asymptotic normality. Denote by e_r the p-vector $(0, 0, \cdots, 1, 0, \cdots, 0)^T$ with 1 on the r^{th} position. When $p = 2$, S_{12} and S_{22} have the following forms

$$S_{12} = \begin{pmatrix} \mathbf{0}_{q\times 1} & \mu_2\Big(f(c_k)\Big)_{q\times 1} \end{pmatrix}, S_{22} = \begin{pmatrix} \mu_2\sum_{k=1}^q f(c_k) & 0 \\ 0 & \mu_4\sum_{k=1}^q f(c_k) \end{pmatrix}.$$

Thus,

$$(S^{-1})_{22} = (S_{22} - S_{21}S_{11}^{-1}S_{12})^{-1} = \begin{pmatrix} \dfrac{1}{\mu_2\sum_{k=1}^q f(c_k)} & 0 \\ 0 & \dfrac{1}{(\mu_4 - \mu_2^2)\sum_{k=1}^q f(c_k)} \end{pmatrix},$$

$$(S^{-1})_{21} = -(S^{-1})_{22}S_{21}S_{11}^{-1} = \begin{pmatrix} \mathbf{0}_{1\times q} \\ \Big(\dfrac{\mu_2}{(\mu_4 - \mu_2^2)\sum_{k=1}^q f(c_k)}\Big)_{1\times q} \end{pmatrix},$$

since $S_{11} = \text{diag}\Big(f(c_1), \cdots, f(c_q)\Big)$. By Theorem 3.8

$$\begin{aligned} Bias(\hat{m}'(t_0)|\mathbf{T}) &= -\frac{1}{hf_T(t_0)}\frac{1}{\sqrt{nh}}e_1^T\Big\{(S^{-1})_{21}E(W_{1n}^*|\mathbf{T}) + (S^{-1})_{22}E(W_{2n}^*|\mathbf{T})\Big\} \\ &= -\frac{1}{hf_T(t_0)}\frac{1}{\mu_2\sum_{k=1}^q f(c_k)}\frac{1}{\sqrt{nh}}E(w_{21}^*|\mathbf{T}). \end{aligned}$$

Note that

$$E(w_{2j}^*|\mathbf{T}) = \frac{1}{\sqrt{nh}}\sum_{i=1}^n K_i x_i^j \sum_{k=1}^q \{F(c_k - r_{i,p}) - F(c_k)\}$$

Therefore, $Bias(\hat{m}'(t_0)|\mathbf{T})$ is equal to $\frac{1}{nh^2}\frac{1}{f_T(t_0)}\sum_{i=1}^n K_i x_i r_{i,p}\{1+o_p(1)\}$. Still using the fact that with $p = 2$

$$\frac{1}{nh}\sum_{i=1}^n K_i x_i r_{i,p} = \frac{f_T(t_0)m'''(t_0)}{6}\frac{\mu_4}{\mu_2}h^3\{1 + o_p(1)\},$$

we obtain

$$Bias(\hat{m}'(t_0)|\mathbf{T}) = \frac{1}{6}m'''(t_0)\frac{\mu_4}{\mu_2}h^2 + o_p(h^2). \tag{3.53}$$

Furthermore, the conditional variance of $\hat{m}(t_0)$ is

$$\begin{aligned} Var(\hat{m}'(t_0)|\mathbf{T}) &= \frac{1}{nh^3}\frac{1}{f_T(t_0)}e_1^T(S^{-1}\Sigma S^{-1})_{22}e_1 + o_p(\frac{1}{nh^3}) \\ &= \frac{1}{nh^3}\frac{\nu_2}{\mu_2^2 f_T(t_0)}R_2(q) + o_p(\frac{1}{nh^3}), \end{aligned} \tag{3.54}$$

which completes the proof. □

Now let us use a simulation example to demonstrate the performance of the local CQR estimate when the error follows a Cauchy distribution.

Example 3.6.1 (Infinite error variance)**.** We generated 400 data sets, each consisting of $n = 200$ observations, from

$$Y = \sin(2T) + 2\exp(-16T^2) + 0.5\epsilon, \tag{3.55}$$

where T follows $N(0,1)$. In our simulation, the error ϵ follows the Cauchy distribution. Thus, the error variance is infinite. For the local polynomial CQR estimator, we consider $q = 5$, 9 and 19, and estimate $m(\cdot)$ and $m'(\cdot)$ over $[-1.5, 1.5]$. The mean and standard deviation of RASE over 400 simulations are summarized in Table 3.5. To see how the proposed estimate behaves at a typical point, Table 3.5 also depicts the biases and standard deviations of $\hat{m}(t)$ and $\hat{m}'(t)$ at $t = 0.75$. In Table 3.5, CQR_5, CQR_9 and CQR_{19} correspond to the local CQR estimate with $q = 5$, 9 and 19, respectively. From Table 3.5, we can see that the RASE of the local CQR estimate is much less than that of local LS estimate. This is because the local LS estimator is not a consistent estimator for the regression function, while the local CQR estimator is. This is also evidenced from the standard deviation of the local estimator at $t = 0.75$.

Table 3.5. Simulation results for example 3.6.1

	$\hat{\boldsymbol{m}}$			$\hat{\boldsymbol{m}}'$		
	RASE	$\boldsymbol{t = 0.75}$		**RASE**	$\boldsymbol{t = 0.75}$	
	Mean(SD)	**Bias**	**Std**	**Mean(SD)**	**Bias**	**Std**
Cauchy						
LS	—	-0.0881	7.8740	—	5.1324	87.7494
CQR_5	$10228_{(125981)}$	-0.0241	0.2965	$14386_{(160902)}$	0.0716	1.5997
CQR_9	$4798_{(51545)}$	-0.0713	0.9690	$14243_{(158913)}$	0.0686	1.6133
CQR_{19}	$1120_{(12889)}$	-0.0929	1.2995	$14224_{(159441)}$	0.0727	1.6064

3.7 Boundary behavior of local CQR estimators

Back to the general nonparametric regression model

$$Y = m(T) + \sigma(T)\epsilon, \tag{3.56}$$

now we study the behavior of the estimator at the boundary of the support of T. Without loss of generality, assume $f_T(\cdot)$ has support on $[0, 1]$. We consider the left boundary point $t = ch$, where c is a positive constant. let

$$\mu_j(c) = \int_{-c}^{\infty} u^j K(u)du \quad \text{and} \quad \nu_j(c) = \int_{-c}^{\infty} u^j K^2(u)du, \quad j = 0, 1, 2, \ldots$$

Note that the local p-polynomial CQR estimator at the boundary point t is constructed by minimizing

$$\sum_{k=1}^{q} \left[\sum_{i=1}^{n} \rho_{\tau_k} \left\{ y_i - a_k - \sum_{j=1}^{p} b_j (t_i - t)^j \right\} K\left(\frac{t_i - t}{h} \right) \right], \tag{3.57}$$

and the local p-polynomial CQR estimators of $m(t)$ and $m^{(r)}(t)$ are given by

$$\hat{m}(t) = \frac{1}{q} \sum_{k=1}^{q} \hat{a}_k, \quad \text{and} \quad \hat{m}^{(r)}(t) = r!\hat{b}_r, \quad r = 1, \cdots, p. \tag{3.58}$$

We first establish asymptotic theory of the local p-polynomial CQR estimators at $t = ch$, and then discuss the special case of $p = 1$ and 2.

For the asymptotic analysis, we need the following regularity conditions:

(A) The regression function $m(t)$ has a right continuous $(p + 1)^{th}$ derivative at the point 0.

(B) The marginal density function $f_T(\cdot)$ is right continuous and positive at the point 0.

(C) The conditional variance $\sigma^2(\cdot)$ is right continuous at the point 0.

(D) The error has a symmetric distribution with positive density $f(\cdot)$.

We also need some new notations. Let $S_{11}(c)$ be a $q \times q$ diagonal matrix with diagonal elements $f(c_k)$, $k = 1, \cdots, q$; $S_{12}(c)$ be a $q \times p$ matrix with (k, j)-element $f(c_k)\mu_j(c)$, $k = 1, \cdots, q$ and $j = 1, \cdots, p$; $S_{21}(c) = S_{12}^T(c)$; and $S_{22}(c)$ be a $p \times p$ matrix with (j, j')-element $\sum_{k=1}^{q} f(c_k)\mu_{j+j'}(c)$, for $j, j' = 1, \cdots, p$. Similarly, let $\Sigma_{11}(c)$ be a $q \times q$ matrix with (k, k')-element $\nu_0(c)\tau_{kk'}$, $k, k' = 1, \cdots, q$; $\Sigma_{12}(c)$ be a $q \times p$ matrix with (k, j)-element $\nu_j(c)\sum_{k'=1}^{q}\tau_{kk'}$, $k = 1, \cdots, q$ and $j = 1, \cdots, p$; $\Sigma_{21}(c) = \Sigma_{12}^T(c)$; and $\Sigma_{22}(c)$ be a $p \times p$ matrix with (j, j')-element $(\sum_{k,k'=1}^{q}\tau_{kk'})\nu_{j+j'}(c)$, for $j, j' = 1, \cdots, p$. Define

$$S(c) = \begin{pmatrix} S_{11}(c) & S_{12}(c) \\ S_{21}(c) & S_{22}(c) \end{pmatrix}, \quad \text{and} \quad \Sigma(c) = \begin{pmatrix} \Sigma_{11}(c) & \Sigma_{12}(c) \\ \Sigma_{21}(c) & \Sigma_{22}(c) \end{pmatrix}.$$

Partition $S^{-1}(c)$ into four submatrices as follows

$$S^{-1}(c) = \begin{pmatrix} S_{11}(c) & S_{12}(c) \\ S_{21}(c) & S_{22}(c) \end{pmatrix}^{-1} = \begin{pmatrix} (S^{-1}(c))_{11} & (S^{-1}(c))_{12} \\ (S^{-1}(c))_{21} & (S^{-1}(c))_{22} \end{pmatrix},$$

where we use $(\cdot)_{11}$ to denote the left-top $q \times q$ submatrix and use $(\cdot)_{22}$ to denote the right-bottom $p \times p$ submatrix.

Furthermore, let $u_k = \sqrt{nh}\{a_k - m(t) - \sigma(t)c_k\}$ and $v_j = h^j\sqrt{nh}\{j!b_j - m^{(j)}(t)\}/j!$. Let $x_i = (t_i - t)/h$, $K_i = K(x_i)$ and $\Delta_{i,k} = \dfrac{u_k}{\sqrt{nh}} + \sum_{j=1}^{p}\dfrac{v_j x_i^j}{\sqrt{nh}}$. Write $d_{i,k} = c_k[\sigma(t_i) - \sigma(t)] + r_{i,p}$ with $r_{i,p} = m(t_i) - \sum_{j=0}^{p} m^{(j)}(t)(t_i - t)^j/j!$. Define $\eta_{i,k}^*$ to be $I(\epsilon_i \le c_k - \frac{d_{i,k}}{\sigma(t_i)}) - \tau_k$. let $W_n^* = (w_{11}^*, \cdots, w_{1q}^*, w_{21}^*, \cdots, w_{2p}^*)^T$ with $w_{1k}^* = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K_i\eta_{i,k}^*$ and $w_{2j}^* = \frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n} K_i x_i^j \eta_{i,k}^*$.

Theorem 3.11. *Denote* $\hat{\theta}_n = (\hat{u}_1, \cdots, \hat{u}_q, \hat{v}_1, \cdots, \hat{v}_p)$ *be the minimizer of (3.57). Under the regularity conditions (A)—(C) listed in this section, we have*

$$\hat{\theta}_n + \frac{\sigma(0+)}{f_T(0+)} S^{-1}(c) E(W_n^* | \mathbf{T}) \xrightarrow{\mathcal{L}} MVN\left(\mathbf{0}, \frac{\sigma^2(0+)}{f_T(0+)} S^{-1}(c)\Sigma(c)S^{-1}(c)\right).$$

The proof is quite similar to the one for interior points, so we omit it here. Now let's look at the asymptotic behavior of local CQR estimators $\hat{m}(t)$ and $\hat{m}'(t)$ at the boundary.

Theorem 3.12. *Under the regularity conditions (A)—(D), if* $h \to 0$, $nh \to \infty$ *as* $n \to \infty$, *the asymptotic conditional bias and variance of the local linear CQR estimator* $\hat{m}(t)$ *are given by*

$$Bias(\hat{m}(t)|\mathbf{T}) = \frac{1}{2} a(c) m''(0+) h^2 + o_p(h^2), \tag{3.59}$$

$$Var(\hat{m}(t)|\mathbf{T}) = \frac{1}{nh} \frac{b(c)\sigma^2(0+)}{f_T(0+)} R_1(q) + o_p\left(\frac{1}{nh}\right). \tag{3.60}$$

where $a(c) = \dfrac{\mu_2^2(c) - \mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - \mu_1^2(c)}$ *and* $b(c) = \dfrac{\mu_2^2(c)\nu_0(c) - 2\mu_1(c)\mu_2(c)\nu_1(c) + \mu_1^2(c)\nu_2(c)}{\{\mu_0(c)\mu_2(c) - \mu_1^2(c)\}^2}$.

Furthermore, conditioning on $\mathbf{T}$, *we have*

$$\sqrt{nh}\left\{\hat{m}(t) - m(t) - \frac{1}{2} a(c) m''(0+) h^2\right\} \xrightarrow{\mathcal{L}} N\left(0, \frac{b(c)\sigma^2(0+)}{f_T(0+)} R_1(q)\right). \tag{3.61}$$

Proof of Theorem 3.12. The asymptotic normality follows Theorem 3.11 with $p = 1$. Let us calculate the conditional bias and variance, respectively. Denote by $e_{q\times 1}$ the vector that contains q 1's. The asymptotic conditional bias of $\hat{m}(t) = \frac{1}{q}\sum_{k=1}^{q} \hat{a}_k$

is

$$Bias(\hat{m}(t)|\mathbf{T}) = \frac{1}{q}\sigma(t)\sum_{k=1}^{q} c_k - \frac{1}{q\cdot\sqrt{nh}}\frac{\sigma(0+)}{f_T(0+)}(e^T_{q\times 1}\ 0)S^{-1}(c)E(W^*_n|\mathbf{T})$$

Note that the error is symmetric, thus $\sum_{k=1}^{q} c_k = 0$, and similarly we can show that

$$E(w^*_{1k}|\mathbf{T}) = f(c_k)\frac{f_T(0+)m''(0+)}{2\sigma(0+)}\mu_2(c)h^2\{1+o_p(1)\} \quad k=1,\cdots,q,$$

and

$$E(w^*_{21}|\mathbf{T}) = \{\sum_{k=1}^{q} f(c_k)\}\frac{f_T(0+)m''(0+)}{2\sigma(0+)}\mu_3(c)h^2\{1+o_p(1)\}.$$

Therefore,

$$\begin{aligned} Bias(\hat{m}(t)|\mathbf{T}) &= -\frac{1}{q\cdot\sqrt{nh}}\frac{\sigma(0+)}{f_T(0+)}(e^T_{q\times 1}\ 0)S^{-1}(c)E(W^*_n|\mathbf{T}) \\ &= \frac{1}{2}\frac{\mu^2_2(c)-\mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c)-\mu^2_1(c)}m''(0+)h^2+o_p(h^2) \\ &= \frac{1}{2}a(c)m''(0+)h^2+o_p(h^2). \end{aligned}$$

Furthermore, the conditional variance of $\hat{m}(t_0)$ is

$$\begin{aligned} Var(\hat{m}(t)|\mathbf{T}) &= \frac{1}{nh}\frac{\sigma^2(0+)}{f_T(0+)}\frac{1}{q^2}e^T_{q\times 1}(S^{-1}(c)\Sigma(c)S^{-1}(c))_{11}e_{q\times 1}+o_p\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{\sigma^2(0+)}{f_T(0+)}\frac{\mu^2_2(c)\nu_0(c)-2\mu_1(c)\mu_2(c)\nu_1(c)+\mu^2_1(c)\nu_2(c)}{\{\mu_0(c)\mu_2(c)-\mu^2_1(c)\}^2}R_1(q)+o_p\left(\frac{1}{nh}\right) \\ &= \frac{1}{nh}\frac{b(c)\sigma^2(0+)}{f_T(0+)}R_1(q)+o_p\left(\frac{1}{nh}\right), \end{aligned} \tag{3.62}$$

which completes the proof. □

From Theorem 3.12, it can be seen that the leading team of the asymptotic bias of the local linear CQR estimator is the same as that of the local linear LS estimator. This relationship is the same as that when x is an interior point. Furthermore, the relationship between the asymptotic variances of the local CQR and that of LS estimators at boundary is also the same as that for interior points, i.e., they are different by the factor $R_1(q)$. Thus, Theorem 3.12 clearly indicates that the local CQR estimator shares the property of the automatic boundary correction, a nice property of local linear least squares estimator.

Theorem 3.13. *Under the regularity conditions (A)—(D), if $h \to 0$, $nh^3 \to \infty$ as $n \to \infty$, the asymptotic conditional bias and variance of the local quadratic CQR estimator $\hat{m}'(t)$ are given by*

$$Bias(\hat{m}'(t)|\mathbf{T}) = \frac{1}{2}a^*(c)m''(0+)h^2 + o_p(h^2), \tag{3.63}$$

$$Var(\hat{m}'(t)|\mathbf{T}) = \frac{1}{nh^3}\frac{b^*(c)\sigma^2(0+)}{f_T(0+)}R_2(q) + o_p\left(\frac{1}{nh^3}\right), \tag{3.64}$$

where $a^*(c)$ and $b^*(c)$ are constants that depend only on c and the kernel K.

Furthermore, conditioning on $\mathbf{T}$, we have

$$\sqrt{nh^3}\left\{\hat{m}'(t) - m(t) - \frac{1}{6}a^*(c)m'''(0+)h^2\right\} \xrightarrow{\mathcal{L}} N\left(0, \frac{b^*(c)\sigma^2(0+)}{f_T(0+)}R_2(q)\right). \tag{3.65}$$

Proof of Theorem 3.13. We apply Theorem 3.11 to get the asymptotic normality. Denote by e_r the p-vector $(0, 0, \cdots, 1, 0, \cdots, 0)^T$ with 1 in the r^{th} position. When $p = 2$, we have

$$E(w^*_{1k}|\mathbf{T}) = f(c_k)\frac{f_T(0+)m'''(0+)}{6\sigma(0+)}\mu_2(c)h^3\{1 + o_p(1)\} \quad k = 1, \cdots, q,$$

and

$$E(w^*_{2j}|\mathbf{T}) = \{\sum_{k=1}^{q} f(c_k)\}\frac{f_T(0+)m'''(0+)}{6\sigma(0+)}\mu_{2+j}(c)h^3\{1+o_p(1)\} \quad j=1,2.$$

Therefore,

$$\begin{aligned} Bias(\hat{m}'(t)|\mathbf{T}) &= -\frac{\sigma(0+)}{hf_T(0+)}\frac{1}{\sqrt{nh}}e_1^T\left\{(S^{-1}(c))_{21}E(W^*_{1n}|\mathbf{T}) + (S^{-1}(c))_{22}E(W^*_{2n}|\mathbf{T})\right\} \\ &= \frac{1}{6}a^*(c)m'''(0+)h^2 + o_p(h^2). \end{aligned}$$

Furthermore, the conditional variance of $\hat{m}'(t)$ is

$$\begin{aligned} Var(\hat{m}'(t)|\mathbf{T}) &= \frac{1}{nh^3}\frac{\sigma^2(0+)}{f_T(0+)}e_1^T(S^{-1}(c)\Sigma(c)S^{-1}(c))_{22}e_1 + o_p\left(\frac{1}{nh^3}\right) \\ &= \frac{1}{nh^3}\frac{b^*(c)\sigma^2(0+)}{f_T(0+)}R_2(q) + o_p\left(\frac{1}{nh^3}\right). \end{aligned} \tag{3.66}$$

This completes the proof. □

From Theorem 3.13, it can be seen that the asymptotic bias of the local CQR estimator at the boundary is of order h^2, and its asymptotic variance is of order $1/nh^3$. Thus, the orders of the asymptotic bias and variance are the same as those of local quadratic regression. Thus, the local quadratic CQR estimator possesses the property of automatic boundary correction.

Now let us use a simulation example to compare the boundary behavior of the local CQR estimator and the local least squares estimator.

Example 3.7.1 (Boundary behavior)**.** We generated 400 data sets, each consisting of $n = 200$ observations, from

$$Y = \sin(2T) + 2\exp(-16T^2) + 0.5\epsilon, \tag{3.67}$$

where T follows $N(0,1)$. In our simulation, the error ϵ follows $0.95N(0,1) + 0.05N(0,10^2)$. Figure 3.4 depicts the 400 estimated coefficient functions of CQR_9 for all 400 simulations. Results for CQR_5 and CQR_{19} are similar, so we opt not to present them here. Figure 3.5 depicts the plots of the estimate of the regression function and its derivative based on a typical data set. From Figures 3.4 and 3.5, it can be clearly seen that the local CQR estimator improves over the local least squares estimator for both interior and boundary points.

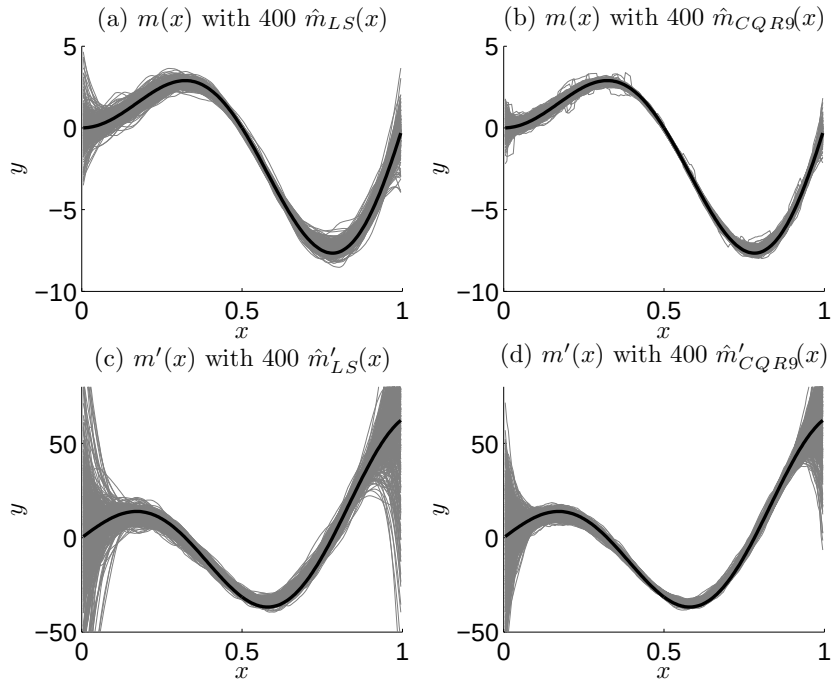


Fig. 3.4. (a) and (c) are plots of 400 local least squares estimators of $m(\cdot)$ and $m^{'}(\cdot)$ over 400 simulation, respectively. (b) and (d) are plots of 400 local CQR estimators of $m(\cdot)$ and $m^{'}(\cdot)$, respectively.

3.8 Discussion

In this Chapter, we have proposed the local linear and quadratic CQR estimators for estimating the nonparametric regression function and its derivative,

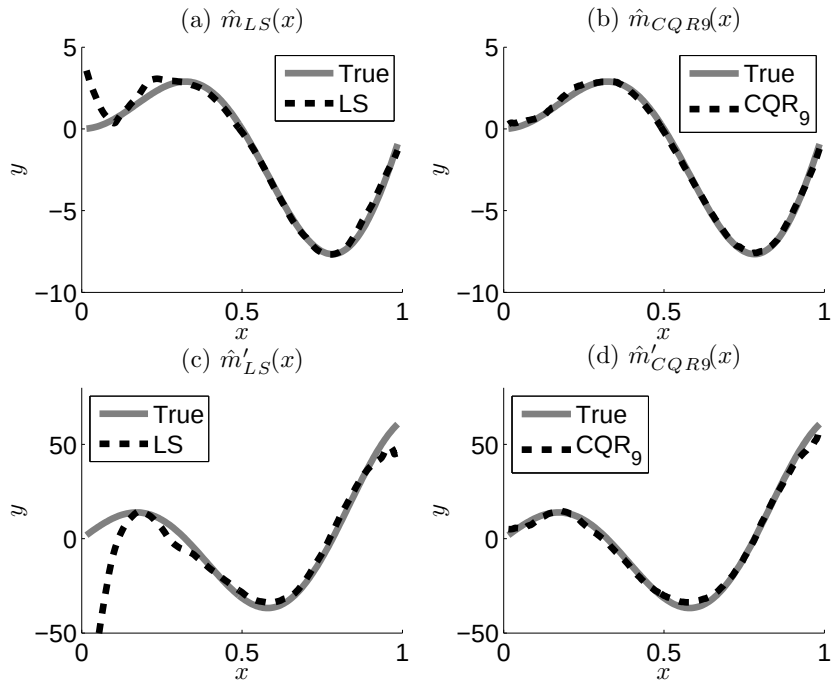


Fig. 3.5. (a) and (c) are plots of a typical local least squares estimators of $m(\cdot)$ and $m^{'}(\cdot)$, respectively. (b) and (d) are plots of a typical local CQR estimators of $m(\cdot)$ and $m^{'}(\cdot)$, respectively.

respectively. We have shown that, compared with the classical local least squares estimators, the new methods enjoy advantages in terms of estimation efficiency measured by MSE or MISE. The theoretical analysis of the two AREs in Theorem 3.2 and Theorem 3.4 provides useful insights into the behavior of the local CQR estimators. Theorem 3.2 indicates that, if using a large number of quantiles, the local linear CQR estimator neither loses nor gains estimation efficiency, compared with the local linear least squares estimator. On the other hand, an interesting phenomenon emerges when a relatively smaller q is used, for the $\mathrm{ARE}(\hat{m}, \hat{m}_{\mathrm{LS}})$ could be much greater than 1 for some non-normal distributions and is almost 1 when the error follows a normal distribution. Theorem 3.4 tells us that, if using a large number of quantiles, $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ can be much greater than 1 for many non-normal error distributions and is 0.97 when the error follows the standard normal distribution. Further study has shown that the value of $\mathrm{ARE}(\hat{m}', \hat{m}'_{\mathrm{LS}})$ for a small q is very close to the theoretical limit. All these results suggest that the local CQR could be a much more efficient alternative to the local least squares regression for estimating both the regression function and its derivative. The theory and numerical results suggest that $q = 9$ can be a good default choice for constructing local CQR smoothers.

Although we have assumed the error has mean zero and variance one for convenience in this work, the validity of the local CQR estimator does not require that the error distribution has a finite variance, unlike the local least squares estimator. This property can be important for real applications, since we have no information on the error distribution in practice. Suppose the error distribution is Cauchy, then the local least squares estimator fails to be consistent, but the local CQR estimator still has consistency and asymptotic normality.

Finally, we would like to point out that the local CQR procedure is efficiently implemented using the MM algorithm. Our experiences show that for $q = 9$ and sample size $n = 7000$, the local CQR fit at a given location can be computed within

0.32 seconds on an AMD 1.9GHz machine. The MM implementation seems to be more efficient than the standard linear programming algorithm.

Chapter 4

New Robust Statistical Procedures for Semiparametric Regression Models

4.1 Introduction

Semiparametric regression modeling has become popular in the recent literature. The partially linear model, the most commonly-used semiparametric regression model, keeps the flexibility of nonparametric models for the baseline function, while maintaining the explanatory power of parametric models. Thus, it has received a lot of attention in the literature. See Härdle, Liang, and Gao (2000), Yatchew (2003) and references therein for theory and application of partially linear models. Various extensions of the partially linear model have been proposed in the literature. See Ruppert, Wand, and Carroll (2003) for applications and theory developments of semiparametric regression models. As an important extension of the partially linear model, the semiparametric varying-coefficient partially linear model is becoming popular in the recent literature.

Let Y be a response variable, and $\{U, \mathbf{X}, \mathbf{Z}\}$ be its covariates. The semiparametric varying-coefficient partially linear model is defined to be

$$Y = \alpha_0(U) + \mathbf{X}^T \boldsymbol{\alpha}(U) + \mathbf{Z}^T \boldsymbol{\beta} + \epsilon, \tag{4.1}$$

where $\alpha_0(U)$ is a baseline function, $\boldsymbol{\alpha}(U) = \{\alpha_1(U), \cdots, \alpha_{d_1}(U)\}^T$ consists of d_1 unknown varying coefficient functions, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{d_2})^T$ is a d_2-dimensional coefficient vector, and ϵ is random error. In this chapter, we will focus on univariate U only, although the proposed procedure is directly applicable for multivariate $\mathbf{U}$. Zhang, Lee, and Song (2002) proposed an estimation procedure for model (4.1)

based on local polynomial regression techniques. Xia, Zhang, and Tong (2004) proposed a semi-local estimation procedure to further reduce the bias of the estimator for $\boldsymbol{\beta}$ suggested in Zhang, Lee, and Song (2002). Fan and Huang (2005) proposed a profile least squares estimator for model (4.1), and demonstrated that their estimator is semiparametrically efficient. Fan and Huang (2005) further developed statistical inference procedures for model (4.1). Li and Liang (2008) proposed model selection procedures for model (4.1) under the framework of generalized linear models. As an extension of Fan and Huang (2005), a profile likelihood estimation procedure was developed in Lam and Fan (2008) under the generalized linear model framework with diverging number of covariates.

Existing estimation procedures for model (4.1) were built on either least squares or likelihood based methods. Thus, the existing estimation procedures are expected to be sensitive to outliers, and their efficiency may be significantly improved for many commonly-used non-normal errors. There is little work on robust estimation and inference procedures for model (4.1), although some robust estimation procedures have been developed for nonparametric regression models and partially linear models in the literature. See, for example, Koenker, Ng, and Portnoy (1994), Fan, Hu, and Truong (1994), He and Shi (1996), Yu and Jones (1998), He, Zhu, and Fung (2002), Lee (2003), among others. It is desirable to develop new robust statistical inference procedures for model (4.1). In this chapter, we propose a quantile regression procedure, a composite quantile regression procedure and a variable selection procedure for model (4.1).

In many situations, quantiles may reveal a more comprehensive view of a distribution than the mean. Quantile regression has appeared as an alternative to least squares in the recent literature. For a complete review, see Koenker (2005) and references therein. Quantile regression has been studied with various nonparametric methods to overcome the limitations of a linear model. Lee (2003)

proposed a $\sqrt{n}$-consistent average quantile regression estimator (AQR) for partially linear regression models. However, the AQR estimator may not be good when outliers exist because the sample mean is not robust. And the ideas of the profile method proposed by Fan and Huang (2005) may not be easily applied here because quantile regression yields non-linear estimates. We propose a new quantile regression procedure for model (4.1). We investigate the sampling properties of the proposed quantile regression estimator and show that the estimators for both the parametric and nonparametric parts achieve the best convergence rate. We also show the asymptotic normality for both estimators. The proposed estimators are less sensitive to data outliers and the choice of bandwidth. The idea of the proposed methodology is quite general and it is easy to implement with efficient computation.

As a special case of quantile regression, least absolute deviation regression provides an estimate of the regression function when the distribution of ϵ is symmetric. Least absolute deviation regression is robust in the presence of outliers, but its estimation efficiency can be dramatically improved by considering other robust loss functions. With the tools of quantile regression in hand, we further propose composite quantile regression for estimation of $\alpha_0(\cdot), \boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}$, the unknown parameters in the regression function of model (4.1).

Composite quantile regression was first proposed for classical linear regression models by Zou and Yuan (2008). They show that the composite quantile regression estimator for the regression coefficients in classical linear regression models could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator. Furthermore, the asymptotic relative efficiency of the composite quantile regression estimator compared to the least squares estimator is greater than 70% regardless the error distribution. In chapter 3, we propose the local polynomial CQR estimator for estimating the nonparametric regression function and its derivative. We establish the asymptotic theory of the local CQR estimator

and show that, compared with the classical local least squares estimator, the new method can significantly improve the estimation efficiency of the local least squares estimator for commonly used non-normal error distributions. At the same time, the loss in efficiency is at most 8.01% in the worst case scenario. By using the same idea of the methodology for quantile regression, we propose the semi-CQR estimators for estimation of means of both the nonparametric and parametric parts in the semiparametric varying-coefficient partially linear model. We show that our estimators achieve the best convergence rates. We also prove the asymptotic normality of the new CQR estimators. The new estimators can dramatically improve the efficiency when errors depart from normal and they only lose a little efficiency for normal errors. The new estimators also work well when the variance of the errors is infinite.

In practice, there are many covariates available in the initial stage of modeling. To reduce model approximation error, it is typical to include many variables in the models. On the other hand, it is always desirable to have a parsimonious model to enhance model predictability and model interpretation by excluding insignificant covariates. Variable selection for model (4.1) is challenging because it involves both nonparametric and parametric parts. Traditional variable selection methods, such as stepwise regression or best subset variable selection, may not work effectively for the semiparametric model because they need to choose smoothing parameters for each sub-model. One aim of this chapter is to develop an effective variable selection procedure to select significant z-variables in model (4.1). We propose a class of variable selection procedures for model (4.1), and demonstrate that the proposed procedures possess the oracle property in the terminology of Fan and Li (2001). Compared to the variable selection procedure based on least squares, our new proposed method is much more robust and consistent for selecting the correct variables. Finite sample simulation studies confirm our findings.

This chapter is organized as follows. In Section 2, we propose a quantile regression procedure for model (4.1), and study the asymptotic properties of the proposed estimator. In Section 3, we propose a composite quantile regression estimation for the unknown coefficient functions and parameters to improve the least absolute deviation regression, and study the asymptotic efficiency of the proposed procedures. In Section 4, we propose a class of variable selection procedures for model (4.1) under both the quantile loss and the composite quantile loss. Simulation studies are presented in Section 5. Regularity conditions and technical proofs are given in Section 6.

4.2 Quantile regression

Define $\rho_\tau(r) = \tau r - rI(r < 0)$ to be the check loss function at $\tau \in (0, 1)$. Quantile regression was first introduced by Koenker and Bassett (1978) to estimate the conditional quantile functions of Y, which are defined to be

$$Q_\tau(u, \mathbf{x}, \mathbf{z}) = \operatorname*{argmin}_a E\big[\rho_\tau(Y - a)|(U, \mathbf{X}, \mathbf{Z}) = (u, \mathbf{x}, \mathbf{z})\big],$$

and the semiparametric varying-coefficient partially linear model assumes that

$$Q_\tau(u, \mathbf{x}, \mathbf{z}) = \alpha_{0,\tau}(u) + \mathbf{x}^T\boldsymbol{\alpha}_\tau(u) + \mathbf{z}^T\boldsymbol{\beta}_\tau.$$

Define

$$\epsilon_\tau = Y - Q_\tau(u, \mathbf{x}, \mathbf{z}) = Y - \alpha_{0,\tau}(U) - \mathbf{X}^T\boldsymbol{\alpha}_\tau(U) - \mathbf{Z}^T\boldsymbol{\beta}_\tau.$$

Then ϵ_τ is random error with conditional τ^{th} quantile zero.

Suppose that $\{U_i, \mathbf{X}_i, \mathbf{Z}_i, Y_i\}$, $i = 1, \cdots, n$ is an independent and identically distributed sample from the model

$$Y = \alpha_{0,\tau}(U) + \mathbf{X}^T\boldsymbol{\alpha}_\tau(U) + \mathbf{Z}^T\boldsymbol{\beta}_\tau + \epsilon_\tau, \tag{4.2}$$

Quantile regression estimates $\alpha_{0,\tau}(\cdot)$, $\boldsymbol{\alpha}_{\tau}(\cdot)$ and $\boldsymbol{\beta}_{\tau}$ by minimizing the quantile loss function

$$\sum_{i=1}^{n} \rho_{\tau}\big\{Y_i - \alpha_0(U_i) - \mathbf{X}_i^T \boldsymbol{\alpha}(U_i) - \mathbf{Z}_i^T \boldsymbol{\beta}\big\}. \tag{4.3}$$

Because (4.3) involves nonparametric functions, we employ local linear regression techniques to estimate $\alpha_{0,\tau}(\cdot)$ and $\boldsymbol{\alpha}_{\tau}(\cdot)$. That is, for U in the neighborhood of u, we locally approximate

$$\alpha_j(U) \approx \alpha_j(u) + \alpha_j'(u)(U-u) \triangleq a_j + b_j(U-u)$$

for $j = 0, \cdots, d_1$. Let $\{\tilde{a}_{0,\tau}, \tilde{b}_{0,\tau}, \tilde{\mathbf{a}}_{\tau}, \tilde{\mathbf{b}}_{\tau}, \tilde{\boldsymbol{\beta}}_{\tau}\}$ be the minimizer of the local weighted quantile loss function

$$\sum_{i=1}^{n} \rho_{\tau}\Big\{Y_i - a_0 - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T \boldsymbol{\beta}\Big\} K_h(U_i - u),$$

where $\mathbf{a} = (a_1, \cdots, a_{d_1})^T$, $\mathbf{b} = (b_1, \cdots, b_{d_1})^T$, $K(\cdot)$ is a given kernel function, and $K_h(\cdot) = K(\cdot/h)/h$ is the rescaling function of K with bandwidth h. Then

$$\tilde{\alpha}_{0,\tau}(u) = \tilde{a}_{0,\tau}, \quad \tilde{\boldsymbol{\alpha}}_{\tau}(u) = \tilde{\mathbf{a}}_{\tau}.$$

Let $F_{\tau}(\cdot|u, \mathbf{x}, \mathbf{z})$ and $f_{\tau}(\cdot|u, \mathbf{x}, \mathbf{z})$ be the density function and cumulative distribution function of the error conditional on $(U, \mathbf{X}, \mathbf{Z}) = (u, \mathbf{x}, \mathbf{z})$, respectively. Denote by $f_U(\cdot)$ the marginal density function of the covariate U. The kernel $K(\cdot)$ is chosen as a symmetric density function and let

$$\mu_j = \int u^j K(u) du \quad \text{and} \quad \nu_j = \int u^j K^2(u) du, \quad j = 0, 1, 2, \ldots$$

We then have the following result.

Theorem 4.1. *Under the regularity conditions (A1) – (A6) given in the Appendix, if* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$*, then*

$$\sqrt{nh}\left[\begin{pmatrix}\tilde{\alpha}_{0,\tau}(u)-\alpha_{0,\tau}(u)\\ \tilde{\boldsymbol{\alpha}}_{\tau}(u)-\boldsymbol{\alpha}_{\tau}(u)\\ \tilde{\boldsymbol{\beta}}_{\tau}-\boldsymbol{\beta}_{\tau}\end{pmatrix}-\frac{\mu_2 h^2}{2}\begin{pmatrix}\alpha''_{0,\tau}(u)\\ \boldsymbol{\alpha}''_{\tau}(u)\\ \mathbf{0}\end{pmatrix}\right] \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \frac{\nu_0\tau(1-\tau)}{f_U(u)}\mathbf{A}_1^{-1}(u)\mathbf{B}_1(u)\mathbf{A}_1^{-1}(u)\right) \quad (4.4)$$

where $\mathbf{A}_1(u) = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})(1,\mathbf{X}^T,\mathbf{Z}^T)^T(1,\mathbf{X}^T,\mathbf{Z}^T)|U=u\big]$ *and* $\mathbf{B}_1(u) = E\big[(1,\mathbf{X}^T,\mathbf{Z}^T)^T(1,\mathbf{X}^T,\mathbf{Z}^T)|U=u\big]$.

Theorem 4.1 implies $\tilde{\boldsymbol{\beta}}_\tau$ are $\sqrt{nh}$-consistent estimators. This is because we use data only in a local neighborhood of u to estimate $\boldsymbol{\beta}_\tau$. Note that $\sqrt{nh}$ is a nonparametric convergent rate. To improve $\tilde{\boldsymbol{\beta}}_\tau$, we propose the following estimation procedure for $\boldsymbol{\beta}_\tau$. Define

$$Y^*_{i,\tau} = Y_i - \tilde{\alpha}_{0,\tau}(U_i) - \mathbf{X}_i^T\tilde{\boldsymbol{\alpha}}_\tau(U_i).$$

A $\sqrt{n}$-consistent quantile regression estimate $\hat{\boldsymbol{\beta}}_\tau$ for $\boldsymbol{\beta}_\tau$ can be obtained by conducting quantile regression $Y^*_{i,\tau}$ over $\mathbf{Z}_i$. That is,

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^n \rho_\tau(Y^*_{i,\tau} - \mathbf{Z}_i^T\boldsymbol{\beta}). \quad (4.5)$$

We now study the asymptotic properties of $\hat{\boldsymbol{\beta}}_\tau$. Denote

$$\boldsymbol{\xi}_\tau(u,\mathbf{x},\mathbf{z}) = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})\mathbf{Z}(1,\mathbf{X}^T,\mathbf{0})|U=u\big]\mathbf{A}_1^{-1}(u)(1,\mathbf{x}^T,\mathbf{z}^T)^T.$$

Theorem 4.2. *Under the regularity conditions (A1) – (A6) given in the Appendix, if* $nh^4 \to 0$ *and* $nh^2/\log(1/h) \to \infty$ *as* $n \to \infty$*, then the asymptotic distribution*

of $\hat{\boldsymbol{\beta}}_\tau$ *is given by*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\right) \xrightarrow{\mathcal{L}} N\left(0, \mathbf{S}_\tau^{-1}\boldsymbol{\Xi}_\tau\mathbf{S}_\tau^{-1}\right), \tag{4.6}$$

where $\mathbf{S}_\tau = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})\mathbf{Z}\mathbf{Z}^T\big]$ *and* $\boldsymbol{\Xi}_\tau = \tau(1-\tau)E\big[\{\mathbf{Z} - \boldsymbol{\xi}_\tau(U,\mathbf{X},\mathbf{Z})\}\{\mathbf{Z} - \boldsymbol{\xi}_\tau(U,\mathbf{X},\mathbf{Z})\}^T\big]$.

Theorem 4.1 suggests the optimal bandwidth $h \sim n^{-1/5}$. It is easy to check that the optimal bandwidth does not satisfy the condition in Theorem 4.2. Hence, in order to obtain the root-n consistency and asymptotic normality for $\hat{\boldsymbol{\beta}}_\tau$, undersmoothing for $\tilde{\alpha}_{0,\tau}(u)$ and $\tilde{\boldsymbol{\alpha}}_\tau(u)$ is necessary. This is a common requirement in semiparametric models, see Carroll et al. (1997) for a detailed discussion.

Both $\tilde{\alpha}_{0,\tau}(u)$ and $\tilde{\boldsymbol{\alpha}}_\tau(u)$ are $\sqrt{nh}$-consistent estimators, but their efficiencies can be further improved. To this end, let $\{\hat{a}_{0,\tau}, \hat{b}_{0,\tau}, \hat{\mathbf{a}}_\tau, \hat{\mathbf{b}}_\tau\}$ be the minimizer of

$$\sum_{i=1}^{n}\rho_\tau\Big\{Y_i - \mathbf{Z}_i^T\hat{\boldsymbol{\beta}}_\tau - a_0 - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\}\Big\}K_h(U_i - u). \tag{4.7}$$

Thus, we have

$$\hat{\alpha}_{0,\tau}(u) = \hat{a}_{0,\tau}, \quad \hat{\boldsymbol{\alpha}}_\tau(u) = \hat{\mathbf{a}}_\tau. \tag{4.8}$$

The theorem below provides the result of asymptotic normality of the refined estimators $\hat{\alpha}_{0,\tau}(u)$ and $\hat{\boldsymbol{\alpha}}_\tau(u)$.

Theorem 4.3. *Under the regularity conditions (A1) – (A6) given in the Appendix, if* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$*, then*

$$\sqrt{nh}\left[\begin{pmatrix}\hat{\alpha}_{0,\tau}(u) - \alpha_{0,\tau}(u)\\ \hat{\boldsymbol{\alpha}}_\tau(u) - \boldsymbol{\alpha}_\tau(u)\end{pmatrix} - \frac{\mu_2 h^2}{2}\begin{pmatrix}\alpha''_{0,\tau}(u)\\ \boldsymbol{\alpha}''_\tau(u)\end{pmatrix}\right] \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \frac{\nu_0\tau(1-\tau)}{f_U(u)}\mathbf{A}_2^{-1}(u)\mathbf{B}_2(u)\mathbf{A}_2^{-1}(u)\right), \tag{4.9}$$

where $\mathbf{A}_2(u) = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})(1,\mathbf{X}^T)^T(1,\mathbf{X}^T)|U=u\big]$ *and* $\mathbf{B}_2(u) = E\big[(1,\mathbf{X}^T)^T(1,\mathbf{X}^T)|U=u\big]$.

Theorem 4.3 shows that $\hat{\alpha}_{0,\tau}(u)$ and $\hat{\boldsymbol{\alpha}}_{\tau}(u)$ have the same conditional asymptotic biases as $\tilde{\alpha}_{0,\tau}(u)$ and $\tilde{\boldsymbol{\alpha}}_{\tau}(u)$, while they have smaller conditional asymptotic variances than $\tilde{\alpha}_{0,\tau}(u)$ and $\tilde{\boldsymbol{\alpha}}_{\tau}(u)$, respectively. Hence, they are asymptotically more efficient than $\tilde{\alpha}_{0,\tau}(u)$ and $\tilde{\boldsymbol{\alpha}}_{\tau}(u)$.

Note that although one can continue as in traditional backfitting algorithms until convergence, it is generally not necessary. Compared with fully iterated backfitting algorithms, the proposed one-step backfitting method is much more computationally efficient and easily implemented.

It is of interest to study the situations in which the random error ϵ is independent of $(U, \mathbf{X}, \mathbf{Z})$. Let us assume that

$$Y = \alpha_0(U) + \mathbf{X}^T\boldsymbol{\alpha}(U) + \mathbf{Z}^T\boldsymbol{\beta} + \epsilon, \tag{4.10}$$

where ϵ follows a distribution F with mean 0. In such situations,

$$Q_{\tau}(u, \mathbf{x}, \mathbf{z}) = \alpha_0(u) + c_{\tau} + \mathbf{x}^T\boldsymbol{\alpha}(u) + \mathbf{z}^T\boldsymbol{\beta},$$

where $c_{\tau} = F^{-1}(\tau)$. Thus, it follows from Theorem 4.3 that

$$\sqrt{nh}\left[\begin{pmatrix}\hat{\alpha}_{0,\tau}(u) - \alpha_0(u) - c_{\tau}\\ \hat{\boldsymbol{\alpha}}_{\tau}(u) - \boldsymbol{\alpha}(u)\end{pmatrix} - \frac{\mu_2 h^2}{2}\begin{pmatrix}\alpha_0''(u)\\ \boldsymbol{\alpha}''(u)\end{pmatrix}\right] \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \frac{\nu_0\tau(1-\tau)}{f_U(u)f^2(c_{\tau})}\mathbf{B}_2^{-1}(u)\right). \tag{4.11}$$

And from Theorem 4.2, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{\tau} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{\tau(1-\tau)}{f^2(c_{\tau})}\mathbf{S}^{-1}\boldsymbol{\Xi}_*\mathbf{S}^{-1}\right), \tag{4.12}$$

where $\mathbf{S} = E(\mathbf{Z}\mathbf{Z}^T)$, $\boldsymbol{\Xi}_* = E\big[(\mathbf{Z} - \boldsymbol{\xi}_*(U, \mathbf{X}, \mathbf{Z}))(\mathbf{Z} - \boldsymbol{\xi}_*(U, \mathbf{X}, \mathbf{Z}))^T\big]$ and

$$\boldsymbol{\xi}_*(u, \mathbf{x}) = E\big[\mathbf{Z}(1, \mathbf{X}^T)|U = u\big]\mathbf{B}_2^{-1}(u)(1, \mathbf{x}^T)^T.$$

Notice that $\mathbf{B}_2^{-1}(u)\mathbf{B}_2(u) = \boldsymbol{I}_{d_1+1}$. It can be easily shown that $E\big[\boldsymbol{\xi}_*(U,\mathbf{X},\mathbf{Z})\mathbf{Z}^T\big] = \mathbf{0}$. Denote $\mathbf{X}^* = (1,\mathbf{X}^T)^T$. Then, it can be shown that

$$\boldsymbol{\Xi}_* = E[E(\mathbf{Z}\mathbf{Z}^T|U)\{E(\mathbf{Z}\mathbf{Z}^T|U) - E(\mathbf{Z}\mathbf{X}^{*T}|U)E(\mathbf{X}^*\mathbf{X}^{*T}|U)^{-1}E(\mathbf{X}^*\mathbf{Z}^T|U)\}^{-1}E(\mathbf{Z}\mathbf{Z}^T|U)]$$

For the partially linear model, $\mathbf{X}^* = 1$. Thus,

$$\boldsymbol{\Xi}_* = E[E(\mathbf{Z}\mathbf{Z}^T|U)\{\mathrm{cov}(\mathbf{Z}|U)\}^{-1}E(\mathbf{Z}\mathbf{Z}^T|U)].$$

From the above analysis, when the random error is independent of covariates, then $\hat{\boldsymbol{\alpha}}_\tau(u)$ and $\hat{\boldsymbol{\beta}}_\tau$ are consistent estimates for $\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}$, respectively, for different τs. This motivates us to improve the efficiency of $\hat{\boldsymbol{\alpha}}(\cdot)$ and $\hat{\boldsymbol{\beta}}$ by using the composite quantile regression method.

4.3 Composite quantile regression

Median regression, as a special case of the quantile regression with $\tau = 1/2$, provides us an estimate for the mean function when the error distribution is symmetric about the origin. When the regression function is our primary interest, the median regression may be significantly improved. Zou and Yuan (2008) proposed composite quantile regression (CQR) to simultaneously improve the robustness of the least squares estimate and estimation efficiency of median regression for the regression coefficients in the linear regression models. Both theoretic and empirical results in Zou and Yuan (2008) encourage us to consider CQR estimator for model (4.1).

Suppose $\{U_i, \mathbf{X}_i, \mathbf{Z}_i, Y_i\}$, $i = 1, \cdots, n$ is an independent and identically distributed sample from model

$$Y = \alpha_0(U) + \mathbf{X}^T\boldsymbol{\alpha}(U) + \mathbf{Z}^T\boldsymbol{\beta} + \epsilon, \tag{4.13}$$

where ϵ is the random error with mean zero. For a given q, let $\tau_k = k/(q+1)$ for $k = 1, 2, \ldots, q$. The CQR procedure estimates $\alpha_0(\cdot)$, $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}_0$ via minimizing the CQR loss function:

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\{Y_i - \alpha_{0k}(U_i) - \mathbf{x}_i^T\boldsymbol{\alpha}(U_i) - \mathbf{z}_i^T\boldsymbol{\beta}\}, \tag{4.14}$$

The estimation procedures proposed in the last section can be adapted for (4.14). Let $\{\tilde{\mathbf{a}}_0, \tilde{b}_0, \tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\beta}}\}$ be the minimizer of the following local CQR loss function

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\Big\{Y_i - a_{0k} - b_0(U_i - u) - \mathbf{X}_i^T\big[\mathbf{a} + \mathbf{b}(U_i - u)\big] - \mathbf{Z}_i^T\boldsymbol{\beta}\Big\}K_h(U_i - u), \tag{4.15}$$

where $\mathbf{a}_0 = (a_{01}, \cdots, a_{0q})^T$, $\mathbf{a} = (a_1, \cdots, a_{d_1})^T$, $\mathbf{b} = (b_1, \cdots, b_{d_1})^T$. Then initial estimates of $\alpha_0(u)$ and $\boldsymbol{\alpha}(u)$ are given by

$$\tilde{\alpha}_0(u) = \frac{1}{q}\sum_{k=1}^{q}\tilde{a}_{0k}, \quad \tilde{\boldsymbol{\alpha}}(u) = \tilde{\mathbf{a}}.$$

To establish asymptotic behaviors of $\tilde{\alpha}_0(u)$, $\tilde{\boldsymbol{\alpha}}(u)$ and $\tilde{\boldsymbol{\beta}}$, let us begin with some new notations. Denote by $F(\cdot)$ and $f(\cdot)$ the density function and cumulative distribution function of the error, respectively. Let $c_k = F^{-1}(\tau_k)$, C be a $q \times q$ diagonal matrix with the j-th diagonal element $f(c_j)$, $\mathbf{c} = C\mathbf{1}$ and $c = \mathbf{1}^T C\mathbf{1}$. We write

$$\mathbf{D}_1(u) = E\left[\begin{pmatrix} C & \mathbf{c}\mathbf{X}^T & \mathbf{c}\mathbf{Z}^T \\ \mathbf{X}\mathbf{c}^T & c\mathbf{X}\mathbf{X}^T & c\mathbf{X}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{c}^T & c\mathbf{Z}\mathbf{X}^T & c\mathbf{Z}\mathbf{Z}^T \end{pmatrix}\middle| U = u\right].$$

Let $\tau_{kk'} = \tau_k \wedge \tau_{k'} - \tau_k\tau_{k'}$, and T be a $q \times q$ matrix with (k, k')-element being $\tau_{kk'}$. $\mathbf{t} = T\mathbf{1}$ and $t = \mathbf{1}^T T\mathbf{1}$.

$$\mathbf{\Sigma}_1(u) = E\left[\left.\begin{pmatrix} T & \mathbf{t}\mathbf{X}^T & \mathbf{t}\mathbf{Z}^T \\ \mathbf{X}\mathbf{t}^T & t\mathbf{X}\mathbf{X}^T & t\mathbf{X}\mathbf{Z}^T \\ \mathbf{Z}\mathbf{t}^T & t\mathbf{Z}\mathbf{X}^T & t\mathbf{Z}\mathbf{Z}^T \end{pmatrix}\right| U = u\right].$$

The following theorem presents the sampling distribution of $\{\tilde{\mathbf{a}}_0, \tilde{b}_0, \tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\beta}}\}$.

Theorem 4.4. *Under the regularity conditions (B1) – (B6) given in the Appendix, if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then*

$$\sqrt{nh}\left[\begin{pmatrix} \tilde{\mathbf{a}}_0 - \boldsymbol{\alpha}_0(u) \\ \tilde{\mathbf{a}} - \boldsymbol{\alpha}(u) \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} - \frac{\mu_2 h^2}{2}\begin{pmatrix} \boldsymbol{\alpha}_0''(u) \\ \boldsymbol{\alpha}''(u) \\ \mathbf{0} \end{pmatrix}\right] \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \frac{\nu_0}{f_U(u)}\mathbf{D}_1^{-1}(u)\mathbf{\Sigma}_1(u)\mathbf{D}_1^{-1}(u)\right), \tag{4.16}$$

where $\boldsymbol{\alpha}_0(u) = \big(\alpha_0(u) + c_1, \cdots, \alpha_0(u) + c_q\big)^T$ and $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$.

As $\boldsymbol{\beta}$ was estimated locally in (4.15), the resulting estimate $\tilde{\boldsymbol{\beta}}$ does not have $\sqrt{n}$-consistent rate. Thus, $\tilde{\boldsymbol{\beta}}$ can be estimated at $\sqrt{n}$-consistent rate by using all data. To this end, define

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{k=1}^{q}\sum_{i=1}^{n} \rho_{\tau_k}\big\{Y_i - \tilde{a}_{0k}(U_i) - \mathbf{X}_i^T\tilde{\mathbf{a}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta}\big\}, \tag{4.17}$$

which is called the *semi-CQR estimator* for $\boldsymbol{\beta}$. We now study the asymptotic properties of $\hat{\boldsymbol{\beta}}$. Let

$$\boldsymbol{\delta}(u, \mathbf{x}, \mathbf{z}) = E\big[\mathbf{Z}(\mathbf{c}^T, c\mathbf{X}^T, \mathbf{0})|U = u\big]\mathbf{D}_1^{-1}(u)(I_q, \mathbf{1}^T\mathbf{x}, \mathbf{1}^T\mathbf{z})^T,$$

which is a $d_2 \times q$ matrix.

Theorem 4.5. *Under the regularity conditions (B1) – (B6) given in the Appendix, if* $nh^4 \to 0$ *and* $nh^2/\log(1/h) \to \infty$ *as* $n \to \infty$*, then the asymptotic distribution of* $\hat{\boldsymbol{\beta}}$ *is given by*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{c^2}\mathbf{S}^{-1}\boldsymbol{\Xi}\mathbf{S}^{-1}\right), \tag{4.18}$$

where $\mathbf{S} = E(\mathbf{Z}\mathbf{Z}^T)$ *and* $\boldsymbol{\Delta} = \sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}E\big[\{\mathbf{Z}-\boldsymbol{\delta}_k(U,\mathbf{X},\mathbf{Z})\}\{\mathbf{Z}-\boldsymbol{\delta}_{k'}(U,\mathbf{X},\mathbf{Z})\}^T\big]$*, and* $\boldsymbol{\delta}_k(U,\mathbf{X},\mathbf{Z})$ *stands for the* k*-th column of the matrix* $\boldsymbol{\delta}(U,\mathbf{X},\mathbf{Z})$*.*

We can further refine the estimates for the nonparametric part. For $\alpha_0(u)$ and $\boldsymbol{\alpha}(u)$, let $\{\hat{\mathbf{a}}_0, \hat{b}_0, \hat{\mathbf{a}}, \hat{\mathbf{b}}\}$ be the minimizer of

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\Big[Y_i - \mathbf{Z}_i^T\hat{\boldsymbol{\beta}} - a_{0k} - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\}\Big]K_h(U_i - u),$$

where $\mathbf{a}_0 = (a_{01}, \cdots, a_{0q})^T$. As a result,

$$\hat{\alpha}_0(u) = \frac{1}{q}\sum_{k=1}^{q}\hat{a}_{0k}, \quad \hat{\boldsymbol{\alpha}}(u) = \hat{\mathbf{a}}. \tag{4.19}$$

We now study the asymptotic properties of $\hat{\alpha}_0(u)$ and $\hat{\boldsymbol{\alpha}}(u)$.

Theorem 4.6. *Under the regularity conditions (B1) – (B6) given in the Appendix, if* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$*, the asymptotic distributions of* $\hat{\alpha}_0(u)$ *and* $\hat{\boldsymbol{\alpha}}(u)$ *are given by*

$$\begin{aligned}\sqrt{nh}\left(\hat{\alpha}_0(u) - \alpha_0(u) - \frac{1}{q}\sum_{k=1}^{q}c_k - \frac{\mu_2 h^2}{2}\alpha_0''(u)\right)& \\ \xrightarrow{\mathcal{L}} N\left(0, \frac{\nu_0}{f_U(u)}\frac{1}{q^2}\mathbf{1}^T\big[\mathbf{D}_2^{-1}(u)\boldsymbol{\Sigma}_2(u)\mathbf{D}_2^{-1}(u)\big]_{11}\mathbf{1}\right)&\end{aligned} \tag{4.20}$$

and

$$\sqrt{nh}\left(\hat{\boldsymbol{\alpha}}(u)-\boldsymbol{\alpha}(u)-\frac{\mu_2 h^2}{2}\boldsymbol{\alpha}''(u)\right)\xrightarrow{\mathcal{L}} N\left(0,\frac{\nu_0}{f_U(u)}\left[\mathbf{D}_2^{-1}(u)\boldsymbol{\Sigma}_2(u)\mathbf{D}_2^{-1}(u)\right]_{22}\right), \tag{4.21}$$

where $[\cdot]_{11}$ *denotes the top-left* $q\times q$ *submatrix and* $[\cdot]_{22}$ *denotes the bottom-right* $d_1\times d_1$ *submatrix.*

Again, the fully iterated backfitting algorithm is not necessary in terms of computationally efficiency. Different from local median regression, the newly proposed estimator has competitive efficiency with respect to traditional least squares estimators. And it is also much more stable and robust than least squares estimators because it utilizes all the information shared across multiple quantile regression.

Remark. The baseline function estimator $\hat{\alpha}_0(u)$ converges to $\alpha_0(u)$ plus the average of uniform quantiles of error. It is consistent for $\alpha_0(u)$ when the error distribution is symmetric, just like the local median estimator.

Note that for model (4.13),

$$E(Y|U)=\alpha_0(U)+E(\mathbf{X}|U)^T\boldsymbol{\alpha}(U)+E(\mathbf{Z}|U)^T\boldsymbol{\beta}.$$

Then it follows that

$$Y=E(Y|U)+\{\mathbf{X}-E(\mathbf{X}|U)\}^T\boldsymbol{\alpha}(U)+\{\mathbf{Z}-E(\mathbf{Z}|U)\}^T\boldsymbol{\beta}+\epsilon.$$

To get insights into the performance of the CQR method, let us consider the situation in which $E(\mathbf{X}|U)=0$ and $E(\mathbf{Z}|U)=0$. Then, all $\mathbf{D}_1(u),\mathbf{D}_2(u),\boldsymbol{\Sigma}_1(u)$

and $\mathbf{\Sigma}_2(u)$ become block diagonal matrices. Thus, from Theorem 4.6, we have

$$\sqrt{nh}\left(\hat{\alpha}_0(u)-\alpha_0(u)-\frac{1}{q}\sum_{k=1}^{q}c_k-\frac{\mu_2 h^2}{2}\alpha_0''(u)\right)\xrightarrow{\mathcal{L}}N\left(0,\frac{\nu_0}{f_U(u)}\frac{1}{q^2}\mathbf{1}^T C^{-1}TC^{-1}\mathbf{1}\right) \tag{4.22}$$

and

$$\sqrt{nh}\left(\hat{\boldsymbol{\alpha}}(u)-\boldsymbol{\alpha}(u)-\frac{\mu_2 h^2}{2}\boldsymbol{\alpha}''(u)\right)\xrightarrow{\mathcal{L}}N\left(0,\frac{t}{c^2}\frac{\nu_0}{f_U(u)}\big\{E(\mathbf{X}\mathbf{X}^T|U=u)\big\}^{-1}\right). \tag{4.23}$$

Note that

$$\boldsymbol{\delta}(u,\mathbf{x},\mathbf{z})=E(\mathbf{Z}\mathbf{X}^T|U=u)\big\{E(\mathbf{X}\mathbf{X}^T|U=u)\big\}^{-1}(\mathbf{1}^T\mathbf{X}\mathbf{Z})^T$$

Thus, all columns of $\boldsymbol{\delta}(u,\mathbf{x},\mathbf{z})$ are the same. Therefore $\mathbf{\Delta}=tE\big[\{\mathbf{Z}-\boldsymbol{\delta}(U,\mathbf{X},\mathbf{Z})\}\{\mathbf{Z}-\boldsymbol{\delta}(U,\mathbf{X},\mathbf{Z})\}^T\big]$, where $t=\sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}$. Note that $E(\boldsymbol{\delta}(U,\mathbf{X},\mathbf{Z})\mathbf{Z}^T)=0$, so we have

$$\mathbf{\Delta}=tE\big[E(\mathbf{Z}\mathbf{Z}^T|U)\{E(\mathbf{Z}\mathbf{Z}^T|U)-E(\mathbf{Z}\mathbf{X}^T|U)E(\mathbf{X}\mathbf{X}^T|U)^{-1}E(\mathbf{X}\mathbf{Z}^T|U)\}^{-1}E(\mathbf{Z}\mathbf{Z}^T|U)\big].$$

Denote $\mathbf{\Delta}_0=\mathbf{\Delta}/t$. Then

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\xrightarrow{\mathcal{L}}N\left(0,\frac{t}{c^2}\mathbf{S}^{-1}\mathbf{\Xi}_0\mathbf{S}^{-1}\right). \tag{4.24}$$

Define

$$R_1(q)=\frac{1}{q^2}\mathbf{1}^T C^{-1}TC^{-1}\mathbf{1}=\frac{1}{q^2}\sum_{k=1}^{q}\sum_{k'=1}^{q}\frac{\tau_{kk'}}{f(c_k)f(c_{k'})}$$

and

$$R_2(q)=\frac{t}{c^2}=\frac{\sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}}{\big\{\sum_{k=1}^{q}f(c_k)\big\}^2},$$

which corresponds to $R_1(q)$ and $R_2(q)$ in Kai, Li, and Zou (2009a), respectively.

4.4 Variable selection

In practice, many variables may be available to include in the full model at the initial stage of modeling. To obtain an interpretable model and to enhance model predictability, it is desirable to exclude useless variables from the full model. Variable selection is an active research area in the recent statistical literature. In this section, we propose variable selection procedures for quantile regression and composite quantile regression using a unified framework.

For quantile regression, we consider the penalized check loss

$$\sum_{i=1}^{n} \rho_\tau\big\{Y_i - \hat{\alpha}_{0,\tau}(U_i) - \mathbf{X}_i^T \hat{\boldsymbol{\alpha}}_\tau(U_i) - \mathbf{Z}_i^T \boldsymbol{\beta}\big\} + n \sum_{j=1}^{d_2} p_{\lambda_n}(|\beta_j|), \tag{4.25}$$

and for composite quantile regression, we consider the penalized CQR loss

$$\sum_{k=1}^{q} \sum_{i=1}^{n} \rho_{\tau_k}\big\{Y_i - \hat{\alpha}_{0k}(U_i) - \mathbf{X}_i^T \hat{\boldsymbol{\alpha}}(U_i) - \mathbf{Z}_i^T \boldsymbol{\beta}\big\} + n \sum_{j=1}^{d_2} p_{\lambda_n}(|\beta_j|), \tag{4.26}$$

where $p_{\lambda_n}(\cdot)$ is a pre-specified penalty function with regularization parameter λ_n. By minimizing the above two objective functions with a proper penalty, we can get a sparse estimator of $\boldsymbol{\beta}$ and achieve the goal of variable selection.

Fan and Li (2001) suggested using a nonconcave penalty. However, optimizing (4.25) or (4.26) is a challenging problem for general types of penalty functions, because the objective function may be non-differentiable and non-concave. Various numerical algorithms have been proposed to address this problem. Fan and Li (2001) suggest using local quadratic approximation (LQA) to substitute for the penalty function and then optimize using Newton-Raphson algorithm. Hunter and Li (2005) further propose a perturbed version of LQA to alleviate one drawback of LQA. Recently, Zou and Li (2008) propose a new unified algorithm by using the

local linear approximation (LLA). They suggest using the one-step LLA estimator, because the one-step LLA automatically adopts a sparse representation and is as efficient as the fully iterative method with a good initial estimator. Thus, it can dramatically reduce the computational cost in minimizing the non-concave penalized form.

Let us use the variable selection procedure for CQR to demonstrate the general idea. We propose to select significant variables in the parametric component by using the CQR loss via the one-step sparse estimate by minimizing

$$G_n(\boldsymbol{\beta}) = \sum_{k=1}^{q}\sum_{i=1}^{n} \rho_{\tau_k}\{Y_i - \hat{\alpha}_{0k}(U_i) - \mathbf{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta}\} + n\sum_{j=1}^{d_2} p'_{\lambda_n}(|\beta_j^{(0)}|)|\beta_j|, \tag{4.27}$$

where the initial estimate $\boldsymbol{\beta}^{(0)}$ is chosen as the un-penalized semi-CQR estimate $\hat{\boldsymbol{\beta}}$ obtained in the previous section. We denote by $\hat{\boldsymbol{\beta}}^{OSE} = \text{argmin}_{\boldsymbol{\beta}} G_n(\boldsymbol{\beta})$ and call it the *one-step sparse semi-CQR* estimator.

In this section, we show that the one-step semi-CQR estimator $\hat{\boldsymbol{\beta}}^{OSE}$ proposed in the previous section enjoys the oracle property. Let $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$ denote the true value of $\boldsymbol{\beta}$, where $\boldsymbol{\beta}_{10}$ is a s-vector. Without loss of generality, we assume $\boldsymbol{\beta}_{20} = \mathbf{0}$ and $\boldsymbol{\beta}_{10}$ contains all nonzero components of $\boldsymbol{\beta}_0$. Furthermore, let $\mathbf{Z}_1$ be the first s elements of $\mathbf{Z}$ and define

$$\boldsymbol{\lambda}(u, \mathbf{x}, \mathbf{z}) = E\big[\mathbf{Z}_1(\mathbf{c}^T, c\mathbf{X}^T, \mathbf{0})|U = u\big]\mathbf{D}_2^{-1}(u)(I_q, \mathbf{1}^T\mathbf{x}, \mathbf{1}^T\mathbf{z})^T$$

Theorem 4.7 (Oracle Property)**.** *Let $p_\lambda(\cdot)$ be the SCAD penalty. Assume that the regularity conditions (B1) — (B6) given in the Appendix hold. If $\sqrt{n}\lambda_n \to \infty$, $\lambda_n \to 0$ and $nh^4 \to 0$, $nh^2/\log(1/h) \to \infty$ as $n \to \infty$, then the one-step semi-CQR estimator $\hat{\boldsymbol{\beta}}^{OSE}$ must satisfy:*

(a) Sparsity: $\hat{\boldsymbol{\beta}}_2^{OSE} = 0$, with probability tending to one;

(b) Asymptotic normality:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_1^{OSE} - \boldsymbol{\beta}_{10}\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{c^2}\mathbf{S}_1^{-1}\boldsymbol{\Lambda}\mathbf{S}_1^{-1}\right),$$

where $\mathbf{S}_1 = E(\mathbf{Z}_1\mathbf{Z}_1^T)$ *and* $\boldsymbol{\Lambda} = \sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}E\big[\{\mathbf{Z}_1 - \boldsymbol{\lambda}_k(U,\mathbf{X},\mathbf{Z}_1)\}\{\mathbf{Z}_1 - \boldsymbol{\lambda}_{k'}(U,\mathbf{X},\mathbf{Z}_1)\}^T\big]$.

The choice of the regularization parameter always plays an important role in penalized variable selection. Various techniques have been proposed in previous studies, such as the generalized cross-validation selector (Fan and Li 2001), BIC selector (Wang et al. 2007), etc. We also propose a similar one here for the CQR loss, which is

$$BIC(\lambda) \;=\; \log\left(\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\big\{Y_i - \hat{a}_{0k}(U_i) - \mathbf{X}_i^T\hat{\mathbf{a}}(U_i) - \mathbf{Z}_i^T\hat{\boldsymbol{\beta}}^{OSE}(\lambda)\big\}\right) + \frac{\log(n)}{n}df_\lambda,$$

where df_λ is the number of non-zero coefficients in the parametric part of the fitted model. The selected regularization parameter $\hat{\lambda}_{BIC} = \operatorname{argmin} BIC(\lambda)$, which can be found by a grid search. The performance of $\hat{\lambda}_{BIC}$ will be examined in our simulation studies.

A variable selection procedure for quantile regression can be performed by minimizing the penalized quantile regression loss

$$G_{n,\tau}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\rho_\tau\big\{Y_i - \hat{\alpha}_{\tau 0}(U_i) - \mathbf{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta}\big\} + n\sum_{j=1}^{d_2}p_{\lambda_j}(|\beta_j|).$$

The procedure is quite similar to the one for CQR. We omit the details here to save space.

Another advantage of using one-step LLA is that (4.27) can be adapted to be solved efficiently by the LARS algorithm (Efron et al. 2004). Define $Y_{i,k}^* = Y_i - \hat{a}_{0k}(U_i) - \mathbf{X}_i^T\hat{\mathbf{a}}(U_i)$ and $r_{i,k} = Y_{i,k}^* - \mathbf{Z}_i^T\boldsymbol{\beta}^{(0)}$. Note that $\rho_\tau(x) = |x|/2 + (\tau - 1/2)x$.

Thus, if ignoring constant terms,

$$\begin{aligned}
G(\boldsymbol{\beta}) &= \sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\{Y_i - \hat{\alpha}_{0k}(U_i) - \mathbf{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta}\} + n\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j| \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n}\left\{|Y_{i,k}^* - \mathbf{Z}_i^T\boldsymbol{\beta}|/2 + (\tau_k - 1/2)(Y_{i,k}^* - \mathbf{Z}_i^T\boldsymbol{\beta})\right\} + n\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j| \\
&\approx \sum_{k=1}^{q}\sum_{i=1}^{n}\left\{(Y_{i,k}^* - \mathbf{Z}_i^T\boldsymbol{\beta})^2/2r_{i,k} + (\tau_k - 1/2)(Y_{i,k}^* - \mathbf{Z}_i^T\boldsymbol{\beta})\right\} + n\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j| \\
&= \sum_{k=1}^{q}\sum_{i=1}^{n}(Y_{i,k}^* + r_{i,k}(\tau_k - 1/2) - \mathbf{Z}_i^T\boldsymbol{\beta})^2/2r_{i,k} + n\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j| \\
&\triangleq \sum_{k=1}^{q}\sum_{i=1}^{n}(Y_{i,k}^{**} - \mathbf{Z}_i^T\boldsymbol{\beta})^2/2r_{i,k} + n\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j|
\end{aligned}$$

Therefore, we can follow the procedures in section 4 of Zou and Li (2008) to apply the LARS algorithm to solve for $\hat{\boldsymbol{\beta}}^{OSE}$.

4.5 Numerical studies

In this section, we conduct simulation studies to assess the finite sample performance of the proposed procedures. In Example 1, we study the proposed estimation procedures for both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. In Example 2, we examine the finite sample performance of the proposed variable selection procedures. Throughout this section we use the Epanechnikov kernel, i.e., $K(u) = \frac{3}{4}(1-u^2)_+$.

Example 4.5.1. In this example, we generate 400 random samples, each consisting of $n = 100$ observations, from following varying coefficient partially linear model

$$Y = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \epsilon, \tag{4.28}$$

where $\alpha_1(U) = \sin(6\pi U)$, $\alpha_2(U) = \sin(2\pi U)$, $\beta_1 = \beta_2 = 1$, $\beta_3 = 0.5$. The covariate U is from the uniform distribution on $[0,1]$. The covariates X_1, X_2, Z_1, Z_2

are jointly normally distributed with mean 0, variance 1 and correlation 2/3, The covariate Z_3 is binary with probability 0.4 to be 1. Furthermore U and $(X_1, X_2, Z_1, Z_2, Z_3)$ are independent. This simulation setting has been used in Fan and Huang (2005). In our simulation, we consider the following error distributions: $N(0,1)$, Laplace, standard Cauchy, t-distribution with 3 degrees of freedom, mixture of normals $0.9N(0,1)+0.1N(0,10^2)$, and lognormal distribution. Because the error is independent of the covariates, the least squares, quantile regression and CQR procedures provide estimates for the same quantity.

Performance of $\hat{\boldsymbol{\beta}}_\tau$ and $\hat{\boldsymbol{\beta}}$.

We first investigate the effect of bandwidth choice. To demonstrate this, we adopt the three bandwidths $h_0 = 0.166, 0.25, 0.375$ used in Fan and Huang (2005). Note that the profile estimates use optimal bandwidths of order $n^{-1/5}$. This does not satisfy the condition in our theorems because undersmoothing is necessary. Thus, we generate the bandwidths for our methodology based on least squares loss by $\hat{h}_{\text{LS}}^{\text{opt}} = h_0 \times n^{-1/10} = O(n^{-3/10})$. For quantile and CQR estimates, we adjust the bandwidth for different error distributions by using the following formula:

$$\begin{aligned} \hat{h}_{CQR}^{\text{opt}} &= \hat{h}_{\text{LS}}^{\text{opt}} \cdot R_2(q)^{1/5}, \\ \hat{h}_{QR,\tau}^{\text{opt}} &= \hat{h}_{\text{LS}}^{\text{opt}} \cdot \left\{ \tau(1-\tau)/f\left[F^{-1}(\tau)\right] \right\}^{1/5}. \end{aligned}$$

In the first study, we only consider normal errors. The mean and standard deviation based on 400 simulations are reported in Table 4.1. We can clearly see that our proposed estimators are not sensitive to the choice of bandwidth. Therefore, in the following studies, we fix $h_0 = 0.25$.

In the second study, we compare the efficiency of $\boldsymbol{\beta}$ of the proposed estimation procedure to the one based on least squares. We report in Table 4.2 the ratio of the MSE (RMSE) of the quantile regression and CQR estimators to the least squares estimator for different error distributions. Table 4.2 shows clearly that the

semi-CQR estimator has many large gains and only small loss relative to the least squares method. When the error follows the normal distribution, the RMSE's of the semi-CQR estimator are slightly less than 1. For all other non-normal distributions in the table, the RMSE's of the semi-CQR estimators can be much greater than one, indicating the gain in efficiency. For quantile regression estimators, the performance varies and depends heavily on the error distribution.

Performance of $\hat{\boldsymbol{\alpha}}_\tau$ and $\hat{\boldsymbol{\alpha}}$.

Now we compare the performance of $\hat{\boldsymbol{\alpha}}$. We compare the performance of the proposed QR and CQR estimates with the least squares estimate using the ratio of average squared errors (RASE). We first let

$$\text{ASE} = \left\{ \frac{1}{n_{\text{grid}}} \sum_{m=1}^{d_1} \sum_{k=1}^{n_{\text{grid}}} \{\hat{a}_m(u_k) - a_m(u_k)\}^2 \right\},$$

where $\{u_k : k = 1, \cdots, n_{\text{grid}}\}$ is a set of grid points uniformly placed on $[0, 1]$ with $n_{\text{grid}} = 200$. Then RASE is defined to be

$$\text{RASE}(\hat{g}) = \frac{\text{ASE}(\hat{g}_{\text{LS}})}{\text{ASE}(\hat{g})} \tag{4.29}$$

for an estimator $\hat{g}$, where $\hat{g}_{\text{LS}}$ is the local polynomial regression estimator under the least squares loss.

The sample mean and the sample standard deviation of the RASEs over 400 simulations are presented in Table 4.3, in which the values in the parentheses are the standard deviations. Table 4.3 clearly demonstrates that the CQR estimator performs almost as well as the least squares estimator when the random error is normally distributed; and the RASE's are much larger than 1 for other error distributions. The efficiency gain can be substantial. Note that for Cauchy random error, the least squares method fails but the CQR estimator still work very well.

Thus, we can conclude that for estimating $\boldsymbol{\alpha}$, our proposed estimator can serve as a nice alternative to least squares estimator.

Table 4.1. Summary of the mean and standard deviation over 400 simulations.

		$Mean(SD_m)$		
h_0	$Method$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
0.166	LSE	0.006(0.187)	−0.008(0.195)	−0.011(0.276)
	CQR_9	0.013(0.185)	−0.008(0.196)	−0.016(0.256)
	$QR_{0.25}$	0.019(0.256)	−0.005(0.269)	−0.257(0.344)
	$QR_{0.50}$	0.011(0.231)	−0.007(0.240)	−0.021(0.327)
	$QR_{0.75}$	0.015(0.252)	−0.026(0.262)	0.250(0.352)
0.250	LSE	0.009(0.175)	−0.012(0.188)	−0.015(0.260)
	CQR_9	0.011(0.184)	−0.008(0.193)	−0.014(0.257)
	$QR_{0.25}$	0.012(0.248)	−0.006(0.259)	−0.160(0.341)
	$QR_{0.50}$	0.017(0.222)	−0.017(0.229)	−0.009(0.306)
	$QR_{0.75}$	0.010(0.247)	−0.012(0.257)	0.130(0.340)
0.375	LSE	0.008(0.184)	−0.012(0.194)	−0.018(0.272)
	CQR_9	0.012(0.188)	−0.011(0.208)	−0.010(0.274)
	$QR_{0.25}$	0.016(0.253)	−0.017(0.257)	−0.108(0.335)
	$QR_{0.50}$	0.019(0.227)	−0.015(0.239)	−0.007(0.319)
	$QR_{0.75}$	0.007(0.255)	−0.016(0.261)	0.082(0.354)

Example 4.5.2. The goal of this example is to compare the performance of the proposed variable selection procedures. In this example, 400 random samples, each consisting of $n = 100$ observations, were generated from the varying coefficient partially linear model

$$Y = \alpha_1(U)X_1 + \alpha_2(U)X_2 + \boldsymbol{\beta}^T\mathbf{Z} + \epsilon, \tag{4.30}$$

where $\boldsymbol{\beta} = [3, 1.5, 0, 0, 2, 0, 0, 0]^T$, and the covariates $X_1, X_2, \mathbf{Z}$ are treated as a single random vector $\mathbf{W}$ and are jointly normally distributed with mean 0, variance

Table 4.2. Summary of the ratio of MSE over 400 simulations.

	RMSE				*RMSE*		
Method	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	*Method*	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
	Standard Normal				*t-distribution with df* $= 3$		
CQR_9	0.908	0.946	1.023	CQR_9	1.321	1.387	1.324
$\text{QR}_{0.25}$	0.500	0.525	0.479	$\text{QR}_{0.25}$	0.710	0.856	0.688
$\text{QR}_{0.50}$	0.621	0.669	0.725	$\text{QR}_{0.50}$	1.043	1.022	1.091
$\text{QR}_{0.75}$	0.505	0.536	0.512	$\text{QR}_{0.75}$	0.842	0.897	0.690
	Laplace				$0.9N(0,1) + 0.1N(0,10^2)$		
CQR_9	1.244	1.219	1.270	CQR_9	4.372	3.953	4.649
$\text{QR}_{0.25}$	0.695	0.616	0.524	$\text{QR}_{0.25}$	2.057	2.090	1.198
$\text{QR}_{0.50}$	0.884	0.886	0.923	$\text{QR}_{0.50}$	3.902	3.611	3.288
$\text{QR}_{0.75}$	0.683	0.685	0.528	$\text{QR}_{0.75}$	2.116	1.944	1.291
	Standard Cauchy				*Log-Normal*		
CQR_9	9199	16643	31602	CQR_9	2.560	2.565	3.243
$\text{QR}_{0.25}$	3949	8610	10596	$\text{QR}_{0.25}$	2.538	2.542	1.383
$\text{QR}_{0.50}$	10398	17108	32048	$\text{QR}_{0.50}$	2.020	1.946	2.326
$\text{QR}_{0.75}$	5193	8077	10415	$\text{QR}_{0.75}$	0.667	0.671	0.674

Table 4.3. Summary of the RASE over 400 simulations.

	Normal	Laplace	Cauchy
CQR_9	0.918(0.114)	1.157(0.206)	14501(199154)
$\text{QR}_{0.25}$	0.604(0.150)	0.701(0.202)	7556(105663)
$\text{QR}_{0.50}$	0.695(0.158)	1.057(0.274)	14248(193966)
$\text{QR}_{0.75}$	0.616(0.164)	0.697(0.191)	6241(95515)
	t_3	Mixture	Log-Normal
CQR_9	1.380(1.169)	3.281(1.308)	2.513(1.740)
$\text{QR}_{0.25}$	0.870(0.632)	1.870(0.843)	3.223(2.759)
$\text{QR}_{0.50}$	1.158(0.998)	2.803(1.093)	1.923(1.393)
$\text{QR}_{0.75}$	0.818(0.304)	1.814(0.764)	0.741(0.602)

1 and correlation $0.5^{|i-j|}(i, j = 1, \cdots, 10)$. Others are exactly the same as those in Example 4.5.1.

Simulation results are summarized in Table 4.4, in which MRME stands for median of ratios of ME of a selected model to that of the ordinary least squares estimate under the full model. Both the columns 'C' and 'IC' are measures of model complexity. Column 'C' shows the average number of nonzero coefficients correctly estimated to be nonzero, and column 'IC' presents the average number of zero coefficients incorrectly estimated to be nonzero. In the column labeled 'U-fit', we present the proportion of trials excluding any nonzero coefficients in 400 replications. Likewise, we report the probability of trials selecting the exact subset model and the probability of trials including all three significant variables and some noise variables in the columns 'C-fit' and 'O-fit', respectively. As can be seen from Table 4.4, both variable selection procedures dramatically reduce model error. However, the CQR One-step SCAD has much better performance compared to the LS One-step SCAD, in terms of all the measurements: MRME, No. of Zeros, and Proportion of fit, and for all the error distributions in Table 4.4. It reflects the advantage of the combination of robustness and efficiency of the new proposed procedure.

4.6 Regularity conditions and proofs

Lemma 4.8 below, which is a direct result of Mack and Silverman (1982) will be repeatedly used in our proofs. Throughout the proofs, terms of the form $G(u) = O_p(a_n)$ always stand for $\sup_{u\in\Omega} |G(u)| = O_p(a_n)$.

Lemma 4.8. *Let* $(\mathbf{X}_1, Y_1), \cdots, (\mathbf{X}_n, Y_n)$ *be i.i.d. random vectors, where the* Y_i*'s are scalar random variables. Assume further that* $E|Y|^r < \infty$ *and that* $\sup_{\mathbf{x}} \int |y|^r f(\mathbf{x}, y)dy < \infty$*, where* f *denotes the joint density of* $(\mathbf{X}, Y)$*. Let* K *be a bounded positive*

Table 4.4. Varialbe selection for semiparametric models with one-step LLA

		No. of Zeros		**Proportion of**		
Method	**MRME**	**C**	**IC**	**U-fit**	**C-fit**	**O-fit**
Standard Normal						
LS One-step SCAD	0.471	4.463	0.005	0.005	0.675	0.320
CQR One-step SCAD	0.407	4.815	0.000	0.000	0.840	0.160
LS Oracle	0.329	5.000	0.000	0.000	1.000	0.000
Laplace						
LS One-step SCAD	0.471	4.425	0.018	0.010	0.685	0.305
CQR One-step SCAD	0.348	4.770	0.007	0.003	0.820	0.177
LS Oracle	0.336	5.000	0.000	0.000	1.000	0.000
Standard Cauchy						
LS One-step SCAD	0.866	2.930	0.870	0.635	0.070	0.295
CQR One-step SCAD	0.055	4.880	0.115	0.105	0.782	0.113
LS Oracle	0.327	5.000	0.000	0.000	1.000	0.000
t-distribution with df $= 3$						
LS One-step SCAD	0.451	4.433	0.000	0.000	0.677	0.323
CQR One-step SCAD	0.316	4.780	0.000	0.000	0.815	0.185
LS Oracle	0.321	5.000	0.000	0.000	1.000	0.000
$0.9N(0,1) + 0.1N(0,10^2)$						
LS One-step SCAD	0.477	4.395	0.033	0.015	0.672	0.313
CQR One-step SCAD	0.148	4.702	0.015	0.005	0.752	0.242
LS Oracle	0.334	5.000	0.000	0.000	1.000	0.000
Log-Normal						
LS One-step SCAD	0.472	4.433	0.005	0.005	0.652	0.343
CQR One-step SCAD	0.202	4.765	0.000	0.000	0.805	0.195
LS Oracle	0.332	5.000	0.000	0.000	1.000	0.000

function with bounded support, satisfying a Lipschitz condition. Then

$$\sup_{\mathbf{x}\in D}\left|n^{-1}\sum_{i=1}^{n}\Big\{K_h(\mathbf{X}_i-\mathbf{x})Y_i-E[K_h(\mathbf{X}_i-\mathbf{x})Y_i]\Big\}\right|=O_p\left(\frac{\log^{1/2}(1/h)}{\sqrt{nh}}\right),$$

provided that $n^{2\epsilon-1}h\to\infty$ *for some* $\epsilon<1-r^{-1}$.

4.6.1 Conditions and proofs for quantile regression

To establish the asymptotic properties of the local quantile regression estimators, the following conditions are imposed.

(A1) The random variable U has a bounded support Ω and its density function $f_U(\cdot)$ is positive and has a continuous second derivative.

(A2) The varying coefficients $\alpha_0(\cdot)$ and $\boldsymbol{\alpha}(\cdot)$ have continuous second derivatives in $u\in\Omega$.

(A3) $K(\cdot)$ is a symmetric density function and the support is bounded.

(A4) The random vector $\mathbf{Z}$ has bounded support.

(A5) $F_\tau(0|u,\mathbf{x},\mathbf{z})=\tau$ for all $(u,\mathbf{x},\mathbf{z})$. And $f_\tau(\cdot|u,\mathbf{x},\mathbf{z})$ is bounded away from zero and continuously differentiable in a neighborhood of 0 for all $(u,\mathbf{x},\mathbf{z})$.

(A6) $\mathbf{A}_1(u)$ and $\mathbf{A}_2(u)$ are non-singular for all $u\in\Omega$.

Let $\eta_{i,\tau}=I(\epsilon_{i,\tau}\le 0)-\tau$ and $\eta^*_{i,\tau}(u)=I\big\{\epsilon_{i,\tau}\le -r_{i,\tau}(u)\big\}-\tau$, where $r_{i,\tau}(u)=\alpha_{0,\tau}(U_i)-\alpha_{0,\tau}(u)-\alpha'_{0,\tau}(u)(U_i-u)+\mathbf{X}_i^T\{\boldsymbol{\alpha}_\tau(U_i)-\boldsymbol{\alpha}_\tau(u)-\boldsymbol{\alpha}'_\tau(u)(U_i-u)\}$. Define $\tilde{\boldsymbol{\theta}}^*_\tau(u)=\sqrt{nh}\big(\tilde{a}_{0,\tau}-\alpha_{0,\tau}(u),(\tilde{\mathbf{a}}_\tau-\boldsymbol{\alpha}_\tau(u))^T,(\tilde{\boldsymbol{\beta}}_\tau-\boldsymbol{\beta}_\tau)^T,h(\tilde{b}_{0,\tau}-\alpha'_{0,\tau}(u)),h(\tilde{\mathbf{b}}_\tau-\boldsymbol{\alpha}'_\tau(u))^T\big)^T$ and $\mathbf{X}^*_i(u)=(1,\mathbf{X}_i^T,\mathbf{Z}_i^T,(U_i-u)/h,\mathbf{X}_i^T(U_i-u)/h)^T$. The following theorem presents the asymptotic representation of $\{\tilde{a}_{0,\tau},\tilde{\mathbf{a}}_\tau,\tilde{\boldsymbol{\beta}}_\tau,\tilde{b}_{0,\tau},\tilde{\mathbf{b}}_\tau\}$.

Lemma 4.9. *Under the regularity conditions given above, if* $h \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$*, then*

$$\tilde{\boldsymbol{\theta}}^*_{\tau}(u) = -f^{-1}_{U}(u)\big\{\mathbf{S}^*_{\tau}(u)\big\}^{-1}\mathbf{W}^*_{n,\tau}(u) + O_p(h^2 + \log^{1/2}(1/h)/\sqrt{nh}) \tag{4.31}$$

holds uniformly for $u \in \Omega$*, where*

$$\mathbf{S}^*_{\tau}(u) = \mathrm{diag}\big\{\mathbf{A}_1(u), \mu_2\mathbf{A}_2(u)\big\}$$

and

$$\mathbf{W}^*_{n,\tau}(u) = \frac{1}{\sqrt{nh}}\sum_{i=1}^{n} K\{(U_i - u)/h\}\eta^*_{i,\tau}(u)\mathbf{X}^*_i(u).$$

Proof. Recall that $\{\tilde{a}_{0,\tau}, \tilde{\mathbf{a}}_{\tau}, \tilde{\boldsymbol{\beta}}_{\tau}, \tilde{b}_{0,\tau}, \tilde{\mathbf{b}}_{\tau}\}$ minimizes

$$\sum_{i=1}^{n} \rho_\tau\Big[Y_i - a_0 - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta}\Big]K_h(U_i - u).$$

Let $\boldsymbol{\theta}^*_{\tau}(u) = \sqrt{nh}\big(a_{0,\tau} - \alpha_{0,\tau}(u), (\mathbf{a}_{\tau} - \boldsymbol{\alpha}_{\tau}(u))^T, (\boldsymbol{\beta}_{\tau} - \boldsymbol{\beta}_{\tau})^T, h(b_{0,\tau} - \alpha'_{0,\tau}(u)), h(\mathbf{b}_{\tau} - \boldsymbol{\alpha}'_{\tau}(u))^T\big)^T$. We write

$$\begin{aligned} & Y_i - a_0 - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta} \\ = \;& \alpha_{0,\tau}(U_i) + \mathbf{X}_i^T\boldsymbol{\alpha}_{\tau}(U_i) + \mathbf{Z}_i^T\boldsymbol{\beta}_{\tau} + \epsilon_{i,\tau} - a_0 - b_0(U_i - u) \\ & -\mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta} \\ = \;& \epsilon_{i,\tau} + r_{i,\tau}(u) - \Delta_{i,\tau}\ , \end{aligned}$$

where $\Delta_{i,\tau} = \big\{\mathbf{X}^*_i(u)\big\}^T\boldsymbol{\theta}^*_{\tau}/\sqrt{nh}$. Then, $\tilde{\boldsymbol{\theta}}^*_{\tau}(u)$ minimizes the function

$$L^*_{n,\tau}(\boldsymbol{\theta}^*_{\tau}) \;=\; \sum_{i=1}^{n} K_i(u)\Big[\rho_\tau\big\{\epsilon_{i,\tau} + r_{i,\tau}(u) - \Delta_{i,\tau}\big\} - \rho_\tau\big\{\epsilon_{i,\tau} + r_{i,\tau}(u)\big\}\Big],$$

where $K_i(u) = K\{(U_i - u)/h\}$. By applying the identity (Knight 1998)

$$\rho_\tau(x-y) - \rho_\tau(x) = y\big\{I(x \le 0) - \tau\big\} + \int_0^y \big\{I(x \le z) - I(x \le 0)\big\}dz, \qquad (4.32)$$

we can write $L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)$ as follows:

$$\begin{aligned} L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) &= \sum_{i=1}^n K_i(u)\Big\{\Delta_{i,\tau}\Big[I\big\{\epsilon_{i,\tau} \le -r_{i,\tau}(u)\big\} - \tau\Big] \\ &\quad + \int_0^{\Delta_{i,\tau}} \Big[I\big\{\epsilon_{i,\tau} \le -r_{i,\tau}(u) + z\big\} - I\big\{\epsilon_{i,\tau} \le -r_{i,\tau}(u)\big\}\Big]dz\Big\} \\ &= \left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n K_i(u)\eta^*_{i,\tau}(u)\mathbf{X}^*_i(u)\right)^T \boldsymbol{\theta}^*_\tau + B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) \\ &= \big\{\mathbf{W}^*_{n,\tau}(u)\big\}^T\boldsymbol{\theta}^*_\tau + B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau), \end{aligned}$$

where

$$B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) = \sum_{i=1}^n K_i(u)\int_0^{\Delta_{i,\tau}} \Big[I\big\{\epsilon_{i,\tau} \le -r_{i,\tau}(u) + z\big\} - I\big\{\epsilon_{i,\tau} \le -r_{i,\tau}(u)\big\}\Big]dz.$$

Since $B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)$ is a summation of i.i.d. random variables of the kernel form, by Lemma 4.8 we have

$$B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) = E\big[B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)\big] + O_p(\log^{1/2}(1/h)/\sqrt{nh}).$$

The conditional expectation of $B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)$ can be calculated as follows

$$\begin{aligned}
& E\big[B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)|U,\mathbf{X},\mathbf{Z}\big] \\
= & \sum_{i=1}^{n} K_i(u)\int_0^{\Delta_{i,\tau}} \Big[F_\tau(-r_{i,\tau}(u)+z|U_i,\mathbf{X}_i,\mathbf{Z}_i) - F_\tau(-r_{i,\tau}(u)|U_i,\mathbf{X}_i,\mathbf{Z}_i)\Big]dz \\
= & \sum_{i=1}^{n} K_i(u)\int_0^{\Delta_{i,\tau}} \Big[z f_\tau(-r_{i,\tau}(u)|U_i,\mathbf{X}_i,\mathbf{Z}_i) + O(z^2)\Big]dz \\
= & \sum_{i=1}^{n} K_i(u)\Big[\Delta^2_{i,\tau} f_\tau(-r_{i,\tau}(u)|U_i,\mathbf{X}_i,\mathbf{Z}_i)/2 + O(\Delta^3_{i,\tau})\Big].
\end{aligned}$$

Note that $\Delta^3_{i,\tau}$ is of the order $1/(\sqrt{nh})^3$, so by using Lemma 4.8, we have

$$\begin{aligned}
& E\big[B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)|U,\mathbf{X},\mathbf{Z}\big] \\
= & \frac{1}{2}(\boldsymbol{\theta}^*_\tau)^T\left(\frac{1}{nh}\sum_{i=1}^{n} K_i(u) f_\tau(-r_{i,\tau}(u)|U_i,\mathbf{X}_i,\mathbf{Z}_i)\big\{\mathbf{X}^*_i(u)\big\}\big\{\mathbf{X}^*_i(u)\big\}^T\right)\boldsymbol{\theta}^*_\tau \\
& + O_p(1/\sqrt{nh}) \\
\triangleq & \frac{1}{2}(\boldsymbol{\theta}^*_\tau)^T\mathbf{S}^*_{n,\tau}(u)\boldsymbol{\theta}^*_\tau + O_p(1/\sqrt{nh}).
\end{aligned}$$

Then,

$$\begin{aligned}
L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) &= \big\{\mathbf{W}^*_{n,\tau}(u)\big\}^T\boldsymbol{\theta}^*_\tau + E\big[B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)\big] + O_p(\log^{1/2}(1/h)/\sqrt{nh}) \\
&= \big\{\mathbf{W}^*_{n,\tau}(u)\big\}^T\boldsymbol{\theta}^*_\tau + E\big\{E\big[B^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)|U,\mathbf{X},\mathbf{Z}\big]\big\} + O_p(\log^{1/2}(1/h)/\sqrt{nh}) \\
&= \big\{\mathbf{W}^*_{n,\tau}(u)\big\}^T\boldsymbol{\theta}^*_\tau + \frac{1}{2}(\boldsymbol{\theta}^*_\tau)^T E\mathbf{S}^*_{n,\tau}(u)\boldsymbol{\theta}^*_\tau + O_p(\log^{1/2}(1/h)/\sqrt{nh}).
\end{aligned}$$

It is easy to check that $E\mathbf{S}^*_{n,\tau}(u) = f_U(u)\mathbf{S}^*_\tau(u) + O_p(h^2)$. Therefore, $L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)$ can be written as

$$L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau) = \big\{\mathbf{W}^*_{n,\tau}(u)\big\}^T\boldsymbol{\theta}^*_\tau + \frac{f_U(u)}{2}(\boldsymbol{\theta}^*_\tau)^T\mathbf{S}^*_\tau(u)\boldsymbol{\theta}^*_\tau + O_p(h^2 + \log^{1/2}(1/h)/\sqrt{nh}). \tag{4.33}$$

By applying the convexity lemma (Pollard 1991) and the quadratic approximation lemma (Fan and Gijbels 1996), the minimizer of $L^*_{n,\tau}(\boldsymbol{\theta}^*_\tau)$ can be expressed as

$$\tilde{\boldsymbol{\theta}}^*_\tau(u) = -f^{-1}_U(u)\{\mathbf{S}^*_\tau(u)\}^{-1}\mathbf{W}^*_{n,\tau}(u) + O_p(h^2 + \log^{1/2}(1/h)/\sqrt{nh}),$$

which holds uniformly for $u \in \Omega$. This completes the proof.

□

Proof of Theorem 4.1. Following the proof of Lemma 4.9, we can obtain that

$$\tilde{\boldsymbol{\theta}}^*_\tau(u) = -f^{-1}_U(u)\{\mathbf{S}^*_\tau(u)\}^{-1}\mathbf{W}^*_{n,\tau}(u) + o_p(1) \tag{4.34}$$

for any point $u \in \Omega$.

Because $\mathbf{W}^*_{n,\tau}(u)$ is a sum of independent and identically distributed random vectors, the asymptotic normality of $\mathbf{W}^*_{n,\tau}(u)$ can be established by the central limit theorem and the Slutsky's theorem. And the asymptotic normality of $\tilde{\boldsymbol{\theta}}^*_\tau(u)$ follows by (4.34). Denote $\mathbf{W}^*_{n,1,\tau}(u) = \frac{1}{\sqrt{nh}}\sum_{i=1}^n K_i(u)\eta^*_{i,\tau}(u)(1, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T$ We now calculate the conditional mean and variance of $\mathbf{W}^*_{n,1,\tau}(u)$.

$$\begin{aligned}
&\frac{1}{\sqrt{nh}}E(\mathbf{W}^*_{n,1,\tau}(u)|U,\mathbf{X},\mathbf{Z}) \\
=\ & \frac{1}{nh}\sum_{i=1}^n K_i(u)\{F_\tau(-r_{i,\tau}(u)|U_i,\mathbf{X}_i,\mathbf{Z}_i) - F_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\}(1,\mathbf{X}_i^T,\mathbf{Z}_i^T)^T \\
=\ & -\frac{1}{nh}\sum_{i=1}^n K_i(u)r_{i,\tau}(u)f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\{1+o(1)\}(1,\mathbf{X}_i^T,\mathbf{Z}_i^T)^T \\
=\ & -\frac{\mu_2 h^2}{2}\mathbf{A}_1(u)\{\alpha''_{0,\tau}(u), \boldsymbol{\alpha}''_\tau(u)^T, \mathbf{0}\}^T + o_p(h^2),
\end{aligned}$$

$$\begin{aligned}
Var(\mathbf{W}^*_{n,1,\tau}(u)|U,\mathbf{X},\mathbf{Z}) &= \frac{1}{nh}\sum_{i=1}^n K_i^2(u)Var(\eta^*_{i,\tau}(u)|U,\mathbf{X},\mathbf{Z})(1,\mathbf{X}_i^T,\mathbf{Z}_i^T)(1,\mathbf{X}_i^T,\mathbf{Z}_i^T)^T \\
&= \nu_0\tau(1-\tau)f_U(u)\mathbf{B}_1(u) + o_p(1).
\end{aligned}$$

Note that $S^*_{\tau} = \text{diag}\{\mathbf{A}_1(u), \mu_2\mathbf{A}_2(u)\}$ is a block diagonal matrix. The asymptotic normality representation of $\{\tilde{\alpha}_{0,\tau}(u), \tilde{\boldsymbol{\alpha}}_{\tau}(u), \tilde{\boldsymbol{\beta}}_{\tau}\}$ follows immediately.

□

Proof of Theorem 4.2. Let $\boldsymbol{\theta}_{\tau} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_{\tau})$. Rewrite

$$\begin{aligned} Y^*_{i,\tau} - \mathbf{Z}_i^T\boldsymbol{\beta} &= \epsilon_{i,\tau} - \{\tilde{\alpha}_{0,\tau}(U_i) - \alpha_{0,\tau}(U_i)\} - \mathbf{X}_i^T\{\tilde{\boldsymbol{\alpha}}_{\tau}(U_i) - \boldsymbol{\alpha}_{\tau}(U_i)\} - \mathbf{Z}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{\tau}) \\ &= \epsilon_{i,\tau} - \tilde{r}_{i,\tau} - \mathbf{Z}_i^T\boldsymbol{\theta}_{\tau}/\sqrt{n}, \end{aligned}$$

where $\tilde{r}_{i,\tau} = \{\tilde{\alpha}_{0,\tau}(U_i) - \alpha_{0,\tau}(U_i)\} + \mathbf{X}_i^T\{\tilde{\boldsymbol{\alpha}}_{\tau}(U_i) - \boldsymbol{\alpha}_{\tau}(U_i)\}$. Then $\hat{\boldsymbol{\theta}}_{\tau}$, which minimizes $\sum_{i=1}^n \rho_{\tau}(Y^*_{i,\tau} - \mathbf{Z}_i^T\boldsymbol{\beta})$, is also the minimizer of

$$L_{n,\tau}(\boldsymbol{\theta}_{\tau}) = \sum_{i=1}^n \left\{\rho_{\tau}(\epsilon_{i,\tau} - \tilde{r}_{i,\tau} - \mathbf{Z}_i^T\boldsymbol{\theta}_{\tau}/\sqrt{n}) - \rho_{\tau}(\epsilon_{i,\tau} - \tilde{r}_{i,\tau})\right\}.$$

By applying the identity (4.32), we can rewrite $L_{n,\tau}(\boldsymbol{\theta}_{\tau})$ as follows:

$$\begin{aligned} L_{n,\tau}(\boldsymbol{\theta}_{\tau}) &= \sum_{i=1}^n \left\{\frac{\mathbf{Z}_i^T\boldsymbol{\theta}_{\tau}}{\sqrt{n}}\big[I(\epsilon_{i,\tau} \le 0) - \tau\big] + \int_{\tilde{r}_{i,\tau}}^{\tilde{r}_{i,\tau} + \mathbf{Z}_i^T\boldsymbol{\theta}_{\tau}/\sqrt{n}} \big[I(\epsilon_{i,\tau} \le z) - I(\epsilon_{i,\tau} \le 0)\big]dz\right\} \\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \eta_{i,\tau}\mathbf{Z}_i\right)^T \boldsymbol{\theta}_{\tau} + B_{n,\tau}(\boldsymbol{\theta}_{\tau}), \end{aligned}$$

where $B_{n,\tau}(\boldsymbol{\theta}_{\tau}) = \sum_{i=1}^n \int_{\tilde{r}_{i,\tau}}^{\tilde{r}_{i,\tau} + \mathbf{Z}_i^T\boldsymbol{\theta}_{\tau}/\sqrt{n}} \big[I(\epsilon_{i,\tau} \le z) - I(\epsilon_{i,\tau} \le 0)\big]dz.$

Simple calculation yields

$$
\begin{aligned}
& E\big[B_{n,\tau}(\boldsymbol{\theta}_\tau)|U,\mathbf{X},\mathbf{Z}\big] \\
= & \sum_{i=1}^{n}\int_{\tilde{r}_{i,\tau}}^{\tilde{r}_{i,\tau}+\mathbf{Z}_i^T\boldsymbol{\theta}_\tau/\sqrt{n}}\Big[F_\tau(z|U_i,\mathbf{X}_i,\mathbf{Z}_i)-F_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\Big]dz \\
= & \sum_{i=1}^{n}\int_{\tilde{r}_{i,\tau}}^{\tilde{r}_{i,\tau}+\mathbf{Z}_i^T\boldsymbol{\theta}_\tau/\sqrt{n}}\Big[zf_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\big\{1+o(1)\big\}\Big]dz \\
= & \frac{1}{2}\boldsymbol{\theta}_\tau^T\left(\frac{1}{n}\sum_{i=1}^{n}f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^T\right)\boldsymbol{\theta}_\tau+\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\tilde{r}_{i,\tau}\mathbf{Z}_i\right)^T\boldsymbol{\theta}_\tau+o_p(1).
\end{aligned}
$$

Define $R_{n,\tau}(\boldsymbol{\theta}_\tau)=B_{n,\tau}(\boldsymbol{\theta}_\tau)-E\big[B_{n,\tau}(\boldsymbol{\theta}_\tau)|U,\mathbf{X},\mathbf{Z}\big]$. By showing $Var\big[B_{n,\tau}(\boldsymbol{\theta}_\tau)|U,\mathbf{X},\mathbf{Z}\big]=o_p(1)$, it is easy to check that $R_{n,\tau}(\boldsymbol{\theta}_\tau)=o_p(1)$. Hence,

$$
\begin{aligned}
L_{n,\tau}(\boldsymbol{\theta}_\tau) &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i,\tau}\mathbf{Z}_i\right)^T\boldsymbol{\theta}_\tau+E\big[B_{n,\tau}(\boldsymbol{\theta}_\tau)|U,\mathbf{X},\mathbf{Z}\big]+R_{n,\tau}(\boldsymbol{\theta}_\tau) \\
&= \frac{1}{2}\boldsymbol{\theta}_\tau^T\left(\frac{1}{n}\sum_{i=1}^{n}f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^T\right)\boldsymbol{\theta}_\tau+\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i,\tau}\mathbf{Z}_i\right)^T\boldsymbol{\theta}_\tau \\
&\quad+\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\tilde{r}_{i,\tau}\mathbf{Z}_i\right)^T\boldsymbol{\theta}_\tau+o_p(1).
\end{aligned}
$$

By Lemma 4.9, the quantity in the third term of the foregoing expression can be expressed as

$$
\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\tilde{r}_{i,\tau}\mathbf{Z}_i \\
= & -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{Z}_i\frac{f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)}{f_U(U_i)}(1,\mathbf{X}_i^T,\mathbf{0})\mathbf{A}_1^{-1}(U_i)\left(\frac{1}{nh}\sum_{j=1}^{n}\eta_{j,\tau}^*(U_i)(1,\mathbf{X}_j^T,\mathbf{Z}_j^T)^TK_j(U_i)\right) \\
& +O_p(h^{3/2}+\log^{1/2}(1/h)/\sqrt{nh^2}) \\
= & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\eta_{j,\tau}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i(1,\mathbf{X}_i^T,\mathbf{0})\frac{f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)}{f_U(U_i)}K_h(U_i-U_j)\right\}\mathbf{A}_1^{-1}(U_j)(1,\mathbf{X}_j^T,\mathbf{Z}_j^T)^T \\
& +O_p(n^{1/2}h^2+\log^{1/2}(1/h)/\sqrt{nh^2}).
\end{aligned}
$$

Using Lemma 4.8 again

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\tilde{r}_{i,\tau}\mathbf{Z}_i \\
= \;& -\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\eta_{j,\tau}\boldsymbol{\xi}_{j,\tau}(U_j,\mathbf{X}_j,\mathbf{Z}_j) + O_p(n^{1/2}h^2+\log^{1/2}(1/h)/\sqrt{nh^2}) \\
= \;& -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i,\tau}\boldsymbol{\xi}_{i,\tau}(U_i,\mathbf{X}_i,\mathbf{Z}_i) + o_p(1),
\end{aligned}$$

where $\boldsymbol{\xi}_{i,\tau}(U_i,\mathbf{X}_i,\mathbf{Z}_i) = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})\mathbf{Z}(1,\mathbf{X}^T,\mathbf{0})|U=U_i\big]\mathbf{A}_1^{-1}(U_i)(1,\mathbf{X}_i^T,\mathbf{Z}_i^T)^T$.

Therefore,

$$\begin{aligned}
L_{n,\tau}(\boldsymbol{\theta}_\tau) &= \frac{1}{2}\boldsymbol{\theta}_\tau^T\left(\frac{1}{n}\sum_{i=1}^{n} f_\tau(0|U_i,\mathbf{X}_i,\mathbf{Z}_i)\mathbf{Z}_i\mathbf{Z}_i^T\right)\boldsymbol{\theta}_\tau \\
&\quad + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_{i,\tau}\big\{\mathbf{Z}_i - \boldsymbol{\xi}_{i,\tau}(U_i,\mathbf{X}_i,\mathbf{Z}_i)\big\}\right)^T\boldsymbol{\theta} + o_p(1) \\
&\triangleq \frac{1}{2}\boldsymbol{\theta}_\tau^T\mathbf{S}_{n,\tau}\boldsymbol{\theta}_\tau + \mathbf{W}_{n,\tau}^T\boldsymbol{\theta}_\tau + o_p(1).
\end{aligned}$$

It is easy to see that $\mathbf{S}_{n,\tau} = E(\mathbf{S}_{n,\tau}) + o_p(1)$, where

$$E(\mathbf{S}_{n,\tau}) = E\big[f_\tau(0|U,\mathbf{X},\mathbf{Z})\mathbf{Z}\mathbf{Z}^T\big] = \mathbf{S}_\tau.$$

Hence,

$$L_{n,\tau}(\boldsymbol{\theta}_\tau) = \frac{1}{2}\boldsymbol{\theta}_\tau^T\mathbf{S}_\tau\boldsymbol{\theta}_\tau + \mathbf{W}_{n,\tau}^T\boldsymbol{\theta}_\tau + o_p(1)$$

Since the convex function $L_{n,\tau}(\boldsymbol{\theta}_\tau) - \mathbf{W}_{n,\tau}^T\boldsymbol{\theta}_\tau$ converges in probability to the convex function $\frac{1}{2}\boldsymbol{\theta}_\tau^T\mathbf{S}_\tau\boldsymbol{\theta}_\tau$, it follows from the convexity lemma (Pollard 1991) that the quadratic approximation to $L_{n,\tau}(\boldsymbol{\theta}_\tau)$ holds uniformly for $\boldsymbol{\theta}_\tau$ in any compact set Θ, which leads to

$$\hat{\boldsymbol{\theta}}_\tau = -\mathbf{S}_\tau^{-1}\mathbf{W}_{n,\tau} + o_p(1). \tag{4.35}$$

By the Cramér-Wald theorem, the Central Limit Theorem for $\mathbf{W}_{n,\tau}$ holds and $Var(\mathbf{W}_{n,\tau}) \xrightarrow{P} \mathbf{\Xi}_\tau = \tau(1-\tau)E\big[\{\mathbf{Z}-\boldsymbol{\xi}_\tau(U,\mathbf{X},\mathbf{Z})\}\{\mathbf{Z}-\boldsymbol{\xi}_\tau(U,\mathbf{X},\mathbf{Z})\}^T\big]$. Therefore, the asymptotic normality of $\hat{\boldsymbol{\beta}}_\tau$ follows

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\right) \xrightarrow{\mathcal{L}} N\left(0, \mathbf{S}_\tau^{-1}\mathbf{\Xi}_\tau\mathbf{S}_\tau^{-1}\right). \tag{4.36}$$

This completes the proof.

□

Proof of Theorem 4.3. The proof is quite similar to the proof of Theorem 4.1. We omit it here.

□

4.6.2 Conditions and proofs for composite quantile regression

Let us continue to the proofs for local CQR estimators. To establish the asymptotic properties of the local CQR estimators, we need the following regularity conditions:

(B1) The random variable U has a bounded support Ω and its density function $f_U(\cdot)$ is positive and has a continuous second derivative.

(B2) The varying coefficients $\alpha_0(\cdot)$ and $\boldsymbol{\alpha}(\cdot)$ have continuous second derivatives in $u \in \Omega$.

(B3) $K(\cdot)$ is a symmetric density function and the support is bounded.

(B4) The random vector $\mathbf{Z}$ has bounded support.

(B5) $f(\cdot)$ is bounded away from zero and has a continuous and uniformly bounded derivative.

(B6) $\mathbf{D}_1(u)$ and $\mathbf{D}_2(u)$ are non-singular for all $u \in \Omega$.

Let $\eta_{i,k} = I(\epsilon_i \le c_k) - \tau_k$ and $\eta^*_{i,k}(u) = I\{\epsilon_i \le c_k - r_i(u)\} - \tau_k$, where $r_i(u) = \alpha_0(U_i) - \alpha_0(u) - \alpha_0'(u)(U_i - u) + \mathbf{X}_i^T\{\boldsymbol{\alpha}(U_i) - \boldsymbol{\alpha}(u) - \boldsymbol{\alpha}'(u)(U_i - u)\}$. Furthermore, let $\tilde{\boldsymbol{\theta}}^*(u) = \sqrt{nh}\big\{\tilde{a}_{01} - \alpha_0(u) - c_1, \cdots, \tilde{a}_{0q} - \alpha_0(u) - c_q, \{\tilde{\mathbf{a}} - \boldsymbol{\alpha}(u)\}^T, \{\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\}^T, h\{\tilde{b}_0 - \alpha_0'(u)\}, h\{\tilde{\mathbf{b}} - \boldsymbol{\alpha}'(u)\}^T\big\}^T$ and $\mathbf{X}^*_{i,k}(u) = \big\{\mathbf{e}_k^T, \mathbf{X}_i^T, \mathbf{Z}_i^T, (U_i - u)/h, \mathbf{X}_i^T(U_i - u)/h\big\}^T$, where $\mathbf{e}_k$ is a q-vector with 1 on the k^{th} position and 0 elsewhere.

In the proof of Theorem 4.4, we will first show the following asymptotic representation of $\{\tilde{\mathbf{a}}_0, \tilde{b}_0, \tilde{\mathbf{a}}, \tilde{\mathbf{b}}, \tilde{\boldsymbol{\beta}}\}$.

$$\tilde{\boldsymbol{\theta}}^*(u) = -f_U^{-1}(u)\{\mathbf{S}^*(u)\}^{-1}\mathbf{W}^*_n(u) + o_p(1), \tag{4.37}$$

where $\mathbf{S}^*(u) = \mathrm{diag}\{\mathbf{D}_1(u), c\mu_2\mathbf{B}_2(u)\}$ and $\mathbf{W}^*_n = \frac{1}{\sqrt{nh}}\sum_{k=1}^q\sum_{i=1}^n K_i(u)\eta^*_{i,k}(u)\mathbf{X}^*_{i,k}(u)$.

Proof of Theorem 4.4. Recall that $\{\tilde{\mathbf{a}}_0, \tilde{\mathbf{a}}, \tilde{\boldsymbol{\beta}}, \tilde{b}_0, \tilde{\mathbf{b}}\}$ minimizes

$$\sum_{k=1}^q\sum_{i=1}^n \rho_{\tau_k}\Big[Y_i - a_{0k} - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta}\Big]K_h(U_i - u).$$

Rewrite

$$\begin{aligned} & Y_i - a_{0k} - b_0(U_i - u) - \mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta} \\ = \;& \alpha_0(U_i) + \mathbf{X}_i^T\boldsymbol{\alpha}(U_i) + \mathbf{Z}_i^T\boldsymbol{\beta}_0 + \epsilon_i - a_{0k} - b_0(U_i - u) \\ & -\mathbf{X}_i^T\big\{\mathbf{a} + \mathbf{b}(U_i - u)\big\} - \mathbf{Z}_i^T\boldsymbol{\beta} \\ = \;& (\epsilon_i - c_k) + r_i - \Delta_{i,k}\,, \end{aligned}$$

where $\Delta_{i,k} = \{\mathbf{X}^*_{i,k}(u)\}^T\boldsymbol{\theta}^*(u)/\sqrt{nh}$. Then, $\tilde{\boldsymbol{\theta}}^*$ is also the minimizer of

$$L^*_n(\boldsymbol{\theta}^*) \;=\; \sum_{k=1}^q\sum_{i=1}^n K_i(u)\Big[\rho_{\tau_k}\big\{(\epsilon_i - c_k) + r_i(u) - \Delta_{i,k}\big\} - \rho_{\tau_k}\big\{(\epsilon_i - c_k) + r_i(u)\big\}\Big].$$

By applying the identity (4.32), we can rewrite $L_n^*(\boldsymbol{\theta}^*)$ as follows:

$$\begin{aligned} L_n^*(\boldsymbol{\theta}^*) &= \sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\Big\{\Delta_{i,k}\Big[I\big\{\epsilon_i \le c_k - r_i(u)\big\} - \tau_k\Big] \\ &\quad + \int_0^{\Delta_{i,k}} \Big[I\big\{\epsilon_i \le c_k - r_i(u) + z\big\} - I\big\{\epsilon_i \le c_k - r_i(u)\big\}\Big]dz\Big\} \\ &= \left(\frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\eta_{i,k}^*(u)\mathbf{X}_{i,k}^*(u)\right)^T \boldsymbol{\theta}^* + \sum_{k=1}^{q} B_{n,k}^*(\boldsymbol{\theta}^*) \\ &= \{\mathbf{W}_n^*(u)\}^T\boldsymbol{\theta}^*(u) + \sum_{k=1}^{q} B_{n,k}^*(\boldsymbol{\theta}^*), \end{aligned}$$

where

$$B_{n,k}^*(\boldsymbol{\theta}^*) = \sum_{i=1}^{n} K_i(u)\int_0^{\Delta_{i,k}} \Big[I\big\{\epsilon_i \le c_k - r_i(u) + z\big\} - I\big\{\epsilon_i \le c_k - r_i(u)\big\}\Big]dz.$$

Since $B_{n,k}^*(\boldsymbol{\theta}^*)$ is a summation of i.i.d. random variables of the kernel form, by Lemma 4.8 we have

$$B_{n,k}^*(\boldsymbol{\theta}^*) = E[B_{n,k}^*(\boldsymbol{\theta}^*)] + O_p(\log^{1/2}(1/h)/\sqrt{nh}).$$

The conditional expectation of $\sum_{k=1}^{q} B_{n,k}^*(\boldsymbol{\theta}^*)$ can be calculated as

$$\begin{aligned} &\sum_{k=1}^{q} E[B_{n,k}^*(\boldsymbol{\theta}^*)|U,\mathbf{X},\mathbf{Z}] \\ =\ &\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\int_0^{\Delta_{i,k}} \Big[F\big(c_k - r_i(u) + z\big) - F\big(c_k - r_i(u)\big)\Big]dz \\ =\ &\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\int_0^{\Delta_{i,k}} \Big[zf\big(c_k - r_i(u)\big) + O(z^2)\Big]dz \\ =\ &\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\Big[\Delta_{i,k}^2 f\big(c_k - r_i(u)\big)/2 + O(\Delta_{i,k}^3)\Big] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}(\boldsymbol{\theta}^*)^T \left(\frac{1}{nh} \sum_{k=1}^{q} \sum_{i=1}^{n} K_i(u) f\big(c_k - r_i(u)\big) \{\mathbf{X}^*_{i,k}(u)\} \{\mathbf{X}^*_{i,k}(u)\}^T \right) \boldsymbol{\theta}^* \\
&\quad + O_p(\log^{1/2}(1/h)/\sqrt{nh}) \\
&\triangleq \frac{1}{2}(\boldsymbol{\theta}^*)^T \mathbf{S}^*_n(u) \boldsymbol{\theta}^* + O_p(\log^{1/2}(1/h)/\sqrt{nh}).
\end{aligned}
$$

Then,

$$
\begin{aligned}
L^*_n(\boldsymbol{\theta}^*) &= \{\mathbf{W}^*_n(u)\}^T \boldsymbol{\theta}^* + \sum_{k=1}^{q} E[B^*_{n,k}(\boldsymbol{\theta}^*)] + O_p(\log^{1/2}(1/h)/\sqrt{nh}) \\
&= \{\mathbf{W}^*_n(u)\}^T \boldsymbol{\theta}^* + \sum_{k=1}^{q} E\big\{ E[B^*_{n,k}(\boldsymbol{\theta}^*) | U, \mathbf{X}, \mathbf{Z}] \big\} + O_p(\log^{1/2}(1/h)/\sqrt{nh}) \\
&= \{\mathbf{W}^*_n(u)\}^T \boldsymbol{\theta}^* + \frac{1}{2}(\boldsymbol{\theta}^*)^T E\mathbf{S}^*_n(u) \boldsymbol{\theta}^* + O_p(\log^{1/2}(1/h)/\sqrt{nh}).
\end{aligned}
$$

It is easy to check that $E\mathbf{S}^*_n(u) = f_U(u)\mathbf{S}^*(u) + O(h^2)$. Therefore, we can write $L_n(\boldsymbol{\theta}^*)$ as

$$
L^*_n(\boldsymbol{\theta}^*) = \{\mathbf{W}^*_n(u)\}^T \boldsymbol{\theta}^* + \frac{f_U(u)}{2}(\boldsymbol{\theta}^*)^T \mathbf{S}^*(u) \boldsymbol{\theta}^* + O_p(h^2 + \log^{1/2}(1/h)/\sqrt{nh}). \quad (4.38)
$$

By applying the convexity lemma (Pollard 1991) and the quadratic approximation lemma (Fan and Gijbels 1996), the minimizer of $L^*_n(\boldsymbol{\theta}^*)$ can be expressed as

$$
\tilde{\boldsymbol{\theta}}^* = -f_U^{-1}(u)\{\mathbf{S}^*(u)\}^{-1} \mathbf{W}^*_n(u) + O_p(h^2 + \log^{1/2}(1/h)/\sqrt{nh}), \quad (4.39)
$$

which holds uniformly for $u \in \Omega$. Meanwhile, for any point $u \in \Omega$, we have

$$
\tilde{\boldsymbol{\theta}}^* = -f_U^{-1}(u)\{\mathbf{S}^*(u)\}^{-1} \mathbf{W}^*_n(u) + o_p(1). \quad (4.40)
$$

Note that $S^* = \text{diag}\{\mathbf{D}_1(u), c\mu_2\mathbf{B}_2(u)\}$ is a quasi-diagonal matrix. So

$$\sqrt{nh}\begin{pmatrix} \tilde{\mathbf{a}}_0 - \boldsymbol{\alpha}_0(u) \\ \tilde{\mathbf{a}} - \boldsymbol{\alpha}(u) \\ \tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \end{pmatrix} = -f_U^{-1}(u)\mathbf{D}_1^{-1}(u)\mathbf{W}^*_{n,1}(u) + o_p(1), \tag{4.41}$$

where $\mathbf{W}^*_{n,1}(u) = \frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\eta^*_{i,k}(u)(\mathbf{e}_k^T, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T$. Let

$$\mathbf{W}^{\#}_{n,1}(u) = \frac{1}{\sqrt{nh}}\sum_{k=1}^{q}\sum_{i=1}^{n} K_i(u)\eta_{i,k}(\mathbf{e}_k^T, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T.$$

Note that

$$Cov(\eta_{i,k}, \eta_{i,k'}) = \tau_{kk'}, \qquad Cov(\eta_{i,k}, \eta_{j,k'}) = 0, \quad \text{if} \quad i \neq j.$$

It is easy to calculate that $E[\mathbf{W}^{\#}_{n,1}(u)] = \mathbf{0}$ and $Var[\mathbf{W}^{\#}_{n,1}(u)] \to f_U(u)\nu_0\boldsymbol{\Sigma}_1(u)$. By the Cramér-Wald theorem, it is easy to see that the CLT for $\mathbf{W}_{n,1}(u)$ holds. Therefore

$$\mathbf{W}^{\#}_{n,1}(u) \xrightarrow{\mathcal{L}} N(\mathbf{0}, f_U(u)\nu_0\boldsymbol{\Sigma}_1(u)).$$

Moreover, we have $Var\big(\mathbf{W}^*_{n,1}(u) - \mathbf{W}^{\#}_{n,1}(u)|U, \mathbf{X}, \mathbf{Z}\big) \leq \frac{q^2}{nh}\sum_{i=1}^{n} K_i^2(u)(\mathbf{e}_k^T, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T (\mathbf{e}_k^T, \mathbf{X}_i^T, \mathbf{Z}_i^T)\max_k\{F(c_k + |r_i|) - F(c_k)\} = o_p(1)$, thus

$$Var\big(\mathbf{W}^*_{n,1}(u) - \mathbf{W}^{\#}_{n,1}(u)\big) = o(1).$$

So by Slutsky's theorem, conditioning on $\{U, \mathbf{X}, \mathbf{Z}\}$, we have

$$\mathbf{W}^*_{n,1}(u) - E[\mathbf{W}^*_{n,1}(u)] \xrightarrow{\mathcal{L}} N(\mathbf{0}, f_U(u)\nu_0\boldsymbol{\Sigma}_1(u)). \tag{4.42}$$

We now calculate the conditional mean of $\mathbf{W}^*_{n,1}(u)$.

$$\begin{aligned}
\frac{1}{\sqrt{nh}}E[\mathbf{W}^*_{n,1}(u)|U,\mathbf{X},\mathbf{Z}] &= \frac{1}{nh}\sum_{k=1}^{q}\sum_{i=1}^{n}K_i(u)\big\{F\big(c_k - r_i(u)\big) - F(c_k)\big\}(\mathbf{e}_k^t, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T \\
&= -\frac{1}{nh}\sum_{k=1}^{q}\sum_{i=1}^{n}K_i(u)r_i(u)f(c_k)\{1+o(1)\}(\mathbf{e}_k^t, \mathbf{X}_i^T, \mathbf{Z}_i^T)^T \\
&= -\frac{\mu_2 h^2}{2}f_U(u)\mathbf{D}_1(u)\begin{pmatrix}\boldsymbol{\alpha}_0''(u)\\ \boldsymbol{\alpha}''(u)\\ \mathbf{0}\end{pmatrix} + o_p(h^2). \qquad (4.43)
\end{aligned}$$

The proof is completed by combining (4.41), (4.42) and (4.43).

□

Proof of Theorem 4.5. Let $\boldsymbol{\theta} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. Rewrite

$$\begin{aligned}
& Y_i - \tilde{a}_{0k}(U_i) - \mathbf{X}_i^T\tilde{\mathbf{a}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta} \\
= \;& \alpha_0(U_i) + \mathbf{X}_i^T\boldsymbol{\alpha}(U_i) + \mathbf{Z}_i^T\boldsymbol{\beta}_0 + \epsilon_i - m_k - \tilde{a}_0(U_i) - \mathbf{X}_i^T\tilde{\mathbf{a}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta} \\
= \;& \epsilon_i - c_k - \{\tilde{a}_{0k}(U_i) - \alpha_0(U_i) - c_k\} - \mathbf{X}_i^T\{\tilde{\mathbf{a}}(U_i) - \boldsymbol{\alpha}(U_i)\} - \mathbf{Z}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
= \;& \epsilon_i - c_k - \tilde{r}_{i,k} - \mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n},
\end{aligned}$$

where $\tilde{r}_{i,k} = \{\tilde{a}_{0k}(U_i) - \alpha_0(U_i) - c_k\} + \mathbf{X}_i^T\{\tilde{\mathbf{a}}(U_i) - \boldsymbol{\alpha}(U_i)\}$. Then $\hat{\boldsymbol{\theta}} = \operatorname{argmin}\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}(Y_i - \tilde{a}_{0k}(U_i) - \mathbf{X}_i^T\tilde{\mathbf{a}}(U_i) - \mathbf{Z}_i^T\boldsymbol{\beta})$ is also the minimizer of

$$L_n(\boldsymbol{\theta}) = \sum_{k=1}^{q}\sum_{i=1}^{n}\left\{\rho_{\tau_k}(\epsilon_i - c_k - \tilde{r}_{i,k} - \mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n}) - \rho_{\tau_k}(\epsilon_i - c_k - \tilde{r}_{i,k})\right\}.$$

By applying the identity (4.32), we can rewrite $L_n(\boldsymbol{\theta})$ as follows:

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \sum_{k=1}^{q}\sum_{i=1}^{n}\left\{\frac{\mathbf{Z}_i^T\boldsymbol{\theta}}{\sqrt{n}}\big[I(\epsilon_i \le c_k) - \tau_k\big] + \int_{\tilde{r}_{i,k}}^{\tilde{r}_{i,k}+\mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n}} \big[I(\epsilon_i \le c_k + z) - I(\epsilon_i \le c_k)\big]dz\right\} \\ &= \left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\eta_{i,k}\mathbf{Z}_i\right)^T\boldsymbol{\theta} + B_n(\boldsymbol{\theta}), \end{aligned}$$

where $B_n(\boldsymbol{\theta}) = \sum_{k=1}^{q}\sum_{i=1}^{n}\int_{\tilde{r}_{i,k}}^{\tilde{r}_{i,k}+\mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n}} \big[I(\epsilon_i \le c_k+z) - I(\epsilon_i \le c_k)\big]dz$. Let us calculate the conditional expectation of $B_n(\boldsymbol{\theta})$.

$$\begin{aligned} & E[B_n(\boldsymbol{\theta})|U,\mathbf{X},\mathbf{Z}] \\ =\ & \sum_{k=1}^{q}\sum_{i=1}^{n}\int_{\tilde{r}_{i,k}}^{\tilde{r}_{i,k}+\mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n}} \big[F(c_k+z) - F(c_k)\big]dz \\ =\ & \sum_{k=1}^{q}\sum_{i=1}^{n}\int_{\tilde{r}_{i,k}}^{\tilde{r}_{i,k}+\mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n}} \big[zf(c_k)\{1+o(1)\}\big]dz \\ =\ & \frac{1}{2}\boldsymbol{\theta}^T\left(\frac{1}{n}\sum_{k=1}^{q}\sum_{i=1}^{n}f(c_k)\mathbf{Z}_i\mathbf{Z}_i^T\right)\boldsymbol{\theta} - \left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}f(c_k)\tilde{r}_{i,k}\mathbf{Z}_i\right)^T\boldsymbol{\theta} + o_p(1). \end{aligned}$$

Define $R_n(\boldsymbol{\theta}) = B_n(\boldsymbol{\theta}) - E\big[B_n(\boldsymbol{\theta})|U,\mathbf{X},\mathbf{Z}\big]$. It can be shown that $R_n(\boldsymbol{\theta}) = o_p(1)$. Hence,

$$\begin{aligned} L_n(\boldsymbol{\theta}) &= \left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\eta_{i,k}\mathbf{Z}_i\right)^T\boldsymbol{\theta} + E\big[B_n(\boldsymbol{\theta})|U,\mathbf{X},\mathbf{Z}\big] + R_n(\boldsymbol{\theta}) \\ &= \frac{1}{2}\boldsymbol{\theta}^T\mathbf{S}_n\boldsymbol{\theta} + \left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\eta_{i,k}\mathbf{Z}_i\right)^T\boldsymbol{\theta} - \left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}f(c_k)\tilde{r}_{i,k}\mathbf{Z}_i\right)^T\boldsymbol{\theta} + o_p(1), \end{aligned}$$

where $\mathbf{S}_n = \frac{1}{n}\sum_{k=1}^{q}\sum_{i=1}^{n}f(c_k)\mathbf{Z}_i\mathbf{Z}_i^T$. By (4.39), the third term in the foregoing expression can be expressed as

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}f(c_k)\tilde{r}_{i,k}\mathbf{Z}_i$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\frac{f(c_k)}{f_U(U_i)}\mathbf{Z}_i(\mathbf{e}_k^T,\mathbf{X}_i^T,\mathbf{0})\mathbf{D}_1^{-1}(U_i)\left(\frac{1}{nh}\sum_{k'=1}^{q}\sum_{i'=1}^{n}\eta_{i',k'}^*(U_i)(\mathbf{e}_{k'}^T,\mathbf{X}_{i'}^T,\mathbf{Z}_{i'}^T)^T K_{i'}(U_i)\right)\\
&\quad +O_p(h^{3/2}+\log^{1/2}(1/h)/\sqrt{nh^2})\\
&= \frac{1}{\sqrt{n}}\sum_{k'=1}^{q}\sum_{i'=1}^{n}\eta_{i',k'}\boldsymbol{\xi}_{k'}(U_{i'},\mathbf{X}_{i'},\mathbf{Z}_{i'})+O_p(n^{1/2}h^2+\log^{1/2}(1/h)/\sqrt{nh^2})\\
&= \frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\eta_{i,k}\boldsymbol{\xi}_k(U_i,\mathbf{X}_i,\mathbf{Z}_i)+o_p(1),
\end{aligned}
$$

where

$$
\boldsymbol{\xi}(U_i,\mathbf{X}_i,\mathbf{Z}_i)=E\big[\mathbf{Z}(\mathbf{c}^T,c\mathbf{X}^T,\mathbf{0})|U=U_i\big]\mathbf{D}_1^{-1}(U_i)(I_q,\mathbf{1}^T\mathbf{X}_i,\mathbf{1}^T\mathbf{Z}_i)^T.
$$

Therefore,

$$
\begin{aligned}
L_n(\boldsymbol{\theta}) &= \frac{1}{2}\boldsymbol{\theta}^T\mathbf{S}_n\boldsymbol{\theta}+\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{q}\sum_{i=1}^{n}\eta_{i,k}\big\{\mathbf{Z}_i-\boldsymbol{\xi}_k(U_i,\mathbf{X}_i,\mathbf{Z}_i)\big\}\right)^T\boldsymbol{\theta}+o_p(1)\\
&\triangleq \frac{1}{2}\boldsymbol{\theta}^T\mathbf{S}_n\boldsymbol{\theta}+\mathbf{W}_n^T\boldsymbol{\theta}+o_p(1).
\end{aligned}
$$

It is easy to see that $\mathbf{S}_n=E(\mathbf{S}_n)+o_p(1)=c\mathbf{S}+o_p(1)$. Hence,

$$
L_n(\boldsymbol{\theta})=\frac{c}{2}\boldsymbol{\theta}^T\mathbf{S}\boldsymbol{\theta}+\mathbf{W}_n^T\boldsymbol{\theta}+o_p(1)
$$

Since the convex function $L_n(\boldsymbol{\theta})-\mathbf{W}_n^T\boldsymbol{\theta}$ converges in probability to the convex function $\frac{c}{2}\boldsymbol{\theta}^T\mathbf{S}\boldsymbol{\theta}$, it follows from the convexity lemma (Pollard 1991) that the quadratic approximation to $L_n(\boldsymbol{\theta})$ holds uniformly for $\boldsymbol{\theta}$ in any compact set Θ, which leads to

$$
\hat{\boldsymbol{\theta}}=-\frac{1}{c}\mathbf{S}^{-1}\mathbf{W}_n+o_p(1). \tag{4.44}
$$

By the Cramér-Wald theorem, the Central Limit Theorem for $\mathbf{W}_n$ holds and $Var(\mathbf{W}_n) \to \mathbf{\Xi} = \sum_{k=1}^{q}\sum_{k'=1}^{q}\tau_{kk'}E\{\mathbf{Z}-\boldsymbol{\xi}_k(U,\mathbf{X},\mathbf{Z})\}\{\mathbf{Z}-\boldsymbol{\xi}_{k'}(U,\mathbf{X},\mathbf{Z})\}^T$. Therefore, the asymptotic normality of $\hat{\boldsymbol{\beta}}$ is followed by

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{c^2}\mathbf{S}^{-1}\mathbf{\Xi}\mathbf{S}^{-1}\right).$$

□

Proof of Theorem 4.6. The asymptotic normality of $\hat{\alpha}_0(u)$ and $\hat{\boldsymbol{\alpha}}(u)$ can be obtained by following the ideas in the proof of Theorem 4.4.

□

Proof of Theorem 4.7. Use the same notation in the proof of Theorem 4.6. Minimizing

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_{\tau_k}\left\{Y_i-\hat{a}_{0k}(U_i)-\mathbf{X}_i^T\hat{\mathbf{a}}(U_i)-\mathbf{Z}_i^T\boldsymbol{\beta}\right\}+n\sum_{j=1}^{d}p'_{\lambda_j}(|\beta_j^{(0)}|)|\beta_j|$$

is equivalent to minimizing

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \sum_{k=1}^{q}\sum_{i=1}^{n}\left\{\rho_{\tau_k}(\epsilon_i-c_k-\hat{r}_{i,k}-\mathbf{Z}_i^T\boldsymbol{\theta}/\sqrt{n})-\rho_{\tau_k}(\epsilon_i-c_k-\hat{r}_{i,k})\right\} \\ &\quad +\sum_{j=1}^{d}p'_{\lambda_j}(|\beta_j^{(0)}|)(|\beta_j|-|\beta_{0j}|) \\ &= \frac{c}{2}\boldsymbol{\theta}^T\mathbf{S}\boldsymbol{\theta}+\mathbf{W}_n^T\boldsymbol{\theta}+\sum_{j=1}^{d}p'_{\lambda_j}(|\beta_j^{(0)}|)(|\beta_j|-|\beta_{0j}|)+o_p(1), \end{aligned}$$

where $\boldsymbol{\theta}=\sqrt{n}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$ and $\hat{r}_{i,k}=\{\hat{a}_{0k}(U_i)-\alpha_0(U_i)-c_k\}+\mathbf{X}_i^T\{\hat{\mathbf{a}}(U_i)-\boldsymbol{\alpha}(U_i)\}$. Similar to the derivation in the proof of Theorem 5 in Zou and Li (2008), the third

term above can be expressed as

$$\sum_{j=1}^{d} p'_{\lambda_j}(|\beta_j^{(0)}|)(|\beta_j| - |\beta_{0j}|) \xrightarrow{P} \begin{cases} 0, & \text{if } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{20}, \\ \infty, & \text{otherwise.} \end{cases} \tag{4.45}$$

Therefore, by the epi-convergence results (Geyer 1994; Knight and Fu 2000), we have $\hat{\boldsymbol{\beta}}_2^{OSE} \xrightarrow{P} 0$ and the asymptotic results for $\hat{\boldsymbol{\beta}}_1^{OSE}$ holds.

To prove sparsity, we only need to show that $\hat{\boldsymbol{\beta}}_2^{OSE} = 0$ with probability tending to one. It suffices to prove that if $\beta_{0j} = 0$, $P(\hat{\beta}_j^{OSE} \neq 0) \to 0$. By using the fact $|\frac{\rho_\tau(t_1)-\rho_\tau(t_2)}{t_1-t_2}| \leq \max(\tau, 1-\tau) < 1$, if $\hat{\beta}_j^{OSE} \neq 0$, then we must have $\sqrt{n}p'_{\lambda_j}(|\beta_j^{(0)}|) < \frac{1}{n}\sum_{i=1}^{n}|Z_{ij}|$. Thus we have $P(\hat{\beta}_j^{OSE} \neq 0) \leq P(\sqrt{n}p'_{\lambda_j}(|\beta_j^{(0)}|) < \frac{1}{n}\sum_{i=1}^{n}|Z_{ij}|)$. But under the assumptions, we have $\sqrt{n}p'_{\lambda_j}(|\beta_j^{(0)}|) \to \infty$, Therefore $P(\hat{\beta}_j^{OSE} \neq 0) \to 0$. This completes the proof.

□

Chapter 5

Local Rank Inference for Varying Coefficient Models

5.1 Introduction

As introduced in Cleveland, Grosse, and Shyu (1992) and Hastie and Tibshirani (1993), the varying coefficient model provides a natural and useful extension of the classical linear regression model by allowing the regression coefficients to depend on certain covariates. Due to its flexibility to explore the dynamic features which may exist in the data and its easy interpretation, the varying coefficient model has been widely applied in many scientific areas. It has also experienced rapid developments in both theory and methodology; see Fan and Zhang (2008) for a comprehensive survey. Fan and Zhang (1999) proposed a two-step estimation procedure for the varying coefficient model when the coefficient functions have possibly different degrees of smoothness. Kauermann and Tutz (1999) investigated the use of varying coefficient models for diagnosing the lack-of-fit of regression, regarding the varying coefficient model as an alternative to a parametric null model. Cai et al. (2000) developed a more efficient estimation procedure for varying coefficient models in the framework of generalized linear models. As special cases of varying coefficient models, time-varying coefficient models are particularly appealing in longitudinal studies, survival analysis and time series data since they allow one to explore the time-varying effect of covariates over the response. Pioneering works on novel applications of time-varying coefficient models to longitudinal data include Brumback and Rice (1998), Hoover et al. (1998), Wu et al. (1998) and Fan and Zhang (2000), among others. For more details, readers are referred to Fan and Li (2006a) and the references therein. Time-varying coefficient models are

also popular in modeling and predicting nonlinear time series data and survival data; see Fan and Zhang (2008) for related literature.

Estimation procedures in the aforementioned papers are built on either of the local least squares type or the local likelihood type method. Although these estimators remain asymptotically normal for a large class of random error distributions, their efficiency can deteriorate dramatically when the true error distribution deviates from normality. Furthermore, these estimators are very sensitive to outliers. Even a few outlying data points may introduce undesirable artificial features in the estimated functions. These considerations motivate us to develop a novel local rank estimation procedure that is highly efficient, robust and computationally simple. In particular, the proposed local rank regression estimator may achieve the nonparametric convergence rate even when the local linear least squares method fails to consistently estimate the regression coefficient functions due to infinite random error variance, which occurs for instance when the random error has a Cauchy distribution.

The new approach can substantially improve upon the commonly used local linear least squares procedure for a wide class of error distributions. Theoretical analysis reveals that the asymptotic relative efficiency (ARE), measured by the asymptotic mean squared error (or the asymptotic mean integrated squared error), of the local rank regression estimator in comparison with the local linear least squares estimator has an expression that is closely related to that of the Wilcoxon-Mann-Whitney rank test in comparison with the two-sample t-test. However, different from the two-sample test scenario, where the efficiency is completely determined by the asymptotic variance, in the current setting of estimating an infinite-dimensional parameter, both bias and variance contribute to the asymptotic efficiency. The value of ARE is often significantly greater than one. For example, the ARE is 167% for estimating the regression coefficient functions

when the random error has a t_3 distribution, is 240% for an exponential random error distribution, and is 493% for a lognormal random error distribution.

A striking feature of the local rank procedure is that its pronounced efficiency gain comes with only a little loss when the random error actually has a normal distribution, for which the ARE of the local rank regression estimator relative to the local linear least squares estimator is above 96% for estimating both the coefficient functions and their derivatives. For estimating the regression coefficient functions, the ARE has a sharp lower bound of 88.96%, which implies that the efficiency loss is at most 11.04% in the worst case scenario. For estimating the first derivative of the regression coefficient functions, the ARE possesses a lower bound of 89.91%. Kim (2007) developed a quantile regression procedure for varying coefficient models when the random errors are assumed to have a certain quantile equal to zero. She used the regression splines method and derived the convergence rate, but the lack of an asymptotic normality result does not allow the comparison of the relative efficiency. On the other hand, one may extend the local quantile regression approach (Yu and Jones 1998) to the varying coefficient models. However, this is expected to yield an estimator which still suffers from loss of efficiency and may have near zero ARE relative to the local linear least squares estimator in the worst case scenario.

The new estimator proposed in this chapter minimizes a convex objective function based on local ranks. The implementation of the minimization can be conveniently carried out using existing functions in the R statistical software package via a simple algorithm (§4.1). The objective function has the form of a generalized U-statistic whose kernel varies with the sample size. Under some mild conditions, we establish the asymptotic representation of the proposed estimator and further prove its asymptotic normality. We derive the formula of the asymptotic relative efficiency of the local rank estimator relative to the local linear least squares estimator, which confirms the efficiency advantage of the local rank approach. We

also extend a resampling approach, which perturbs the objective function repeatedly, to the generalized U-statistics setting; and demonstrate that it can accurately estimate the asymptotic covariance matrix.

This chapter is organized as follows. Section 2 presents the local rank procedure for estimating the varying coefficient models. Section 3 discusses its large sample properties and proposes a resampling method for estimating the asymptotic covariance matrix. In Section 4, we address issues related to practical implementation and present Monte Carlo simulation results. We further illustrate the proposed procedure via analyzing an environmental data set. Regularity conditions and technical proofs are presented in Section 5.

5.2 Local rank estimation procedure

Let Y be a response variable, and U and $\mathbf{X}$ be the covariates. The varying coefficient model is defined by

$$Y = a_0(U) + \mathbf{X}^T \mathbf{a}(U) + \epsilon, \tag{5.1}$$

where $a_0(\cdot)$ and $\mathbf{a}(\cdot)$ are both unknown smooth functions. In this chapter, it is assumed that U is a scalar and $\mathbf{X}$ is a p-dimensional vector. The proposed procedures can be extended to the case of multivariate U with more complicated notations by following the same idea in this chapter.

Suppose that $\{U_i, \mathbf{X}_i, Y_i\}$, $i = 1, \ldots, n$, is a random sample from model (5.1). Write $X_i = (X_{i1}, \ldots, X_{ip})^T$ and $\mathbf{a}(\cdot) = (a_1(\cdot), \ldots, a_p(\cdot))^T$. For u in a neighborhood of any given u_0, we locally approximate the coefficient function by a Taylor expansion

$$a_m(u) \approx a_m(u_0) + a_m'(u_0)(u - u_0), \quad m = 0, 1, \ldots, p. \tag{5.2}$$

Denote $\alpha_1 = a_0(u_0)$, $\alpha_2 = a_0'(u_0)$, $\beta_m = a_m(u_0)$ and $\beta_{p+m} = a_m'(u_0)$, for $m = 1, \ldots, p$. Based on the above approximation, we obtain the residual for estimating Y_i at $U_i = u_0$

$$e_i = Y_i - \alpha_1 - \alpha_2(U_i - u_0) - \sum_{m=1}^{p} \big[\beta_m + \beta_{p+m}(U_i - u_0)\big] X_{im}. \tag{5.3}$$

We define the local rank objective function to be

$$Q_n(\boldsymbol{\beta}, \alpha_2) = \frac{1}{n(n-1)} \sum_{1 \le i,j \le n} |e_i - e_j| K_h(U_i - u_0) K_h(U_j - u_0), \tag{5.4}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_{2p})^T$, and for a given kernel function $K(\cdot)$ and a bandwidth h, $K_h(t) = h^{-1}K(t/h)$. Note that $Q_n(\boldsymbol{\beta}, \alpha_2)$ does not depend on α_1 because α_1 is canceled out in $e_i - e_j$. The objective function $Q_n(\boldsymbol{\beta}, \alpha_2)$ is a local version of Gini's mean difference, which is a classical measure of concentration or dispersion (David 1998). Without the kernel functions, $[n(n-1)]^{-1} \sum_{1 \le i,j \le n} |e_j - e_j|$ is the global rank objective function that leads to the classical rank estimator in linear models based on Wilcoxon scores. Rank-based statistical procedures have played a fundamental role in nonparametric analysis of linear models due to their high efficiency and robustness. We refer to the review paper of McKean (2004) for many useful references.

For any given u_0, minimizing $Q_n(\boldsymbol{\beta}, \alpha_2)$ yields the local Wilcoxon rank estimator for $(\boldsymbol{\beta}_0^T, \alpha_2)^T$, where $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(u_0) = (a_1(u_0), \ldots, a_p(u_0), a_1'(u_0), \ldots, a_p'(u_0))^T$. Denote the minimizer of $Q_n(\boldsymbol{\beta}, \alpha_2)$ by $(\widehat{\boldsymbol{\beta}}^T, \widehat{\alpha}_2)^T$. Then for $m = 1, \cdots, p$,

$$\widehat{a}_m(u_0) = \widehat{\beta}_m, \quad \widehat{a}_m'(u_0) = \widehat{\beta}_{p+m} \text{ and } \widehat{a}_0'(u_0) = \widehat{\alpha}_2,$$

In the sequel, we also use the vector notation $\widehat{\mathbf{a}}(u_0) = (\widehat{a}_1(u_0), \cdots, \widehat{a}_p(u_0))^T$ and $\widehat{\mathbf{a}}'(u_0) = (\widehat{a}_1'(u_0), \cdots, \widehat{a}_p'(u_0))^T$.

The location parameter $a_0(u_0)$ needs to be estimated separately. This is analogous to the scenario of global rank estimation of the intercept in a linear regression model. In order to make the intercept identifiable, it is essential to have additional location constraints on the random errors. We adopt the commonly used constraint that ϵ_i has median zero. Given $(\widehat{\boldsymbol{\beta}}^T, \widehat{\alpha}_2)^T$, we estimate $a_0(u_0)$ by $\widehat{\alpha}_1$, the value of α_1 that minimizes

$$n^{-1}\sum_{i=1}^{n}\Big|Y_i-\alpha_1-\widehat{\alpha}_2(U_i-u_0)-\sum_{i=1}^{p}\big[\widehat{\beta}_m+\widehat{\beta}_{p+m}(U_i-u_0)\big]X_{im}\Big|K_h(U_i-u_0), \quad (5.5)$$

which is a local version of a weighted L_1-norm objective function.

5.3 Theoretical properties

5.3.1 Large sample distributions

In this subsection, we investigate the asymptotic properties of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}_2$. The main challenge comes from the non-smoothness of the objective function $Q_n(\boldsymbol{\beta}, \alpha_2)$. To overcome this difficulty, we first derive an asymptotic representation of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}_2$ via a quadratic approximation of $Q_n(\boldsymbol{\beta}, \alpha_2)$, which holds uniformly in a local neighborhood of the true parameter values. Aided by this asymptotic representation, we further establish the asymptotic normality of the local rank estimator.

Let us begin with some new notation. Let $\gamma_n = (nh)^{-1/2}$, and define

$$\begin{aligned}
\boldsymbol{\beta}^* &= \gamma_n^{-1}\big(\beta_1 - a_1(u_0), \ldots, \beta_p - a_p(u_0), h(\beta_{p+1} - a_1'(u_0)), \ldots, h(\beta_{2p} - a_p'(u_0))\big)^T,\\
\boldsymbol{\alpha}^* &= \big(\alpha_1^*, \alpha_2^*\big)^T = \gamma_n^{-1}(\alpha_1 - a_0(u_0), h(\alpha_2 - a_0'(u_0)))^T,\\
\Delta_i(u_0) &= \sum_{m=1}^{p}\big[a_m(U_i) - a_m(u_0) - a_m'(u_0)(U_i - u_0)\big]X_{im}\\
&\qquad + \big[a_0(U_i) - a_0(u_0) - a_0'(u_0)(U_i - u_0)\big].
\end{aligned}$$

Let $(\widehat{\boldsymbol{\beta}}^{*T}_{n}, \widehat{\alpha}^{*}_{2n})^T$ be the value of $(\boldsymbol{\beta}^{*T}, \alpha^{*}_{2})^T$ that minimizes the following reparametrized objective function

$$\begin{aligned} Q^{*}_{n}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) &= \frac{1}{n(n-1)} \sum_{1\le i,j\le n} \Big| \big(\epsilon_i - \gamma_n \alpha^{*}_{2}(U_i - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0)\big) \\ &\quad - \big(\epsilon_j - \gamma_n \alpha^{*}_{2}(U_j - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\big)\Big| K_h(U_i - u_0) K_h(U_j - u_0), \end{aligned} \tag{5.6}$$

where $\mathbf{Z}_i = \big(\mathbf{X}^T_i, ((U_i - u_0)/h)\mathbf{X}^T_i\big)^T$. Let $\mathbf{H} = \mathrm{diag}(1, h) \otimes \mathbf{I}_p$, where $\otimes$ denotes the operation of Kronecker product and $\mathbf{I}_p$ denotes a $p \times p$ identity matrix. Then it can be easily seen that

$$\widehat{\boldsymbol{\beta}}^{*}_{n} = \sqrt{nh}\mathbf{H}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \ \text{ and } \ \widehat{\alpha}^{*}_{2n} = \sqrt{nh^3}\big[\widehat{\alpha}_2 - a^{'}_{0}(u_0)\big].$$

We next show that the non-smooth function $Q^{*}_{n}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2})$ can be locally approximated by a quadratic function of $(\boldsymbol{\beta}^{*T}, \alpha^{*}_{2})^T$. Let $\mu_i = \int t^i K(t)dt$, $i = 1, 2$, and $\nu_i = \int t^i K^2(t)dt$, $i = 0, 1, 2$. In this chapter we assume that the kernel function $K(\cdot)$ is symmetric. This is not restrictive considering that most of the commonly used kernel functions, such as the Epanechnikov kernel $K(t) = 0.75(1 - t^2)I(|t| < 1)$, are symmetric. We use $S_n(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) = \big(S^T_{n1}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}), S_{n2}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2})\big)^T$ to denote the gradient function of $Q^{*}_{n}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2})$, i.e., $S_{n1}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) = \nabla_{\boldsymbol{\beta}^{*}} Q^{*}_{n}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2})$ and $S_{n2}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) = \nabla_{\alpha^{*}_{2}} Q^{*}_{n}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2})$. More specifically,

$$\begin{aligned} S_{n1}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) &= 2\gamma_n [n(n-1)]^{-1} \sum_{i\neq j} \big[I\big(\epsilon_i - \gamma_n \alpha^{*}_{2}(U_i - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0) \le \epsilon_j \\ &\quad - \gamma_n \alpha^{*}_{2}(U_j - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\big) - 1/2\big](\mathbf{Z}_i - \mathbf{Z}_j) K_h(U_i - u_0) K_h(U_j - u_0) \end{aligned}$$

and

$$\begin{aligned} S_{n2}(\boldsymbol{\beta}^{*}, \alpha^{*}_{2}) &= 2\gamma_n [n(n-1)]^{-1} \sum_{i\neq j} \big[I\big(\epsilon_i - \gamma_n \alpha^{*}_{2}(U_i - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0) \le \epsilon_j \\ &\quad - \gamma_n \alpha^{*}_{2}(U_j - u_0)/h - \gamma_n \boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\big) - 1/2\big]((U_i - U_j)/h) K_h(U_i - u_0) K_h(U_j - u_0). \end{aligned}$$

Furthermore, we consider the following quadratic function of $(\boldsymbol{\beta}^{*T}, \alpha_2^*)^T$:

$$\begin{aligned} B_n(\boldsymbol{\beta}^*, \alpha_2^*) &= \gamma_n^{-1}(\boldsymbol{\beta}^{*T}, \alpha_2^*) \begin{pmatrix} S_{n1}(\mathbf{0}, 0) \\ S_{n2}(\mathbf{0}, 0) \end{pmatrix} + \frac{1}{2}\gamma_n(\boldsymbol{\beta}^{*T}, \alpha_2^*)\mathbf{A} \begin{pmatrix} \boldsymbol{\beta}^* \\ \alpha_2^* \end{pmatrix} \\ &\quad + \gamma_n^{-1} Q_n^*(\mathbf{0}, 0), \end{aligned} \tag{5.7}$$

where

$$\mathbf{A} = 4\tau f^2(u_0) \begin{pmatrix} \Sigma(u_0) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mu_2\Sigma(u_0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mu_2 \end{pmatrix}, \tag{5.8}$$

$\Sigma(u_0) = E[\mathbf{X}_i\mathbf{X}_i^T | U_i = u_0]$, $\mathbf{0}$ denotes a matrix (or vector) of zeroes whose dimension is determined by the context, $\tau = \int g^2(t)dt$ is the Wilcoxon constant, and $g(\cdot)$ is the density function of the random error ϵ.

Lemma 5.1. *Suppose that Conditions (C1)—(C4) in the Appendix hold. Then* $\forall\ \epsilon > 0$, $\forall\ c > 0$,

$$P\left[\sup_{\|(\boldsymbol{\beta}^{*T}, \alpha_2^*)\| \le c} \left|\gamma_n^{-1} Q_n^*(\boldsymbol{\beta}^*, \alpha_2^*) - B_n(\boldsymbol{\beta}^*, \alpha_2^*)\right| \ge \epsilon\right] \to 0,$$

where $||\cdot||$ *denotes the Euclidean norm.*

Lemma 5.1 implies that the non-smooth objective function $Q_n^*(\boldsymbol{\beta}^*, \alpha_2^*)$ can be uniformly approximated by a quadratic function $B_n(\boldsymbol{\beta}^*, \alpha_2^*)$ in a neighborhood around $\mathbf{0}$. In the appendix, it is also shown that the minimizer of $B_n(\boldsymbol{\beta}^*, \alpha_2^*)$ is asymptotically within a $o(1)$ neighborhood of $(\widehat{\boldsymbol{\beta}}_n^{*T}, \widehat{\alpha}_{n2}^*)^T$. This further allows us to derive the asymptotic distribution.

The local linear Wilcoxon estimator of $\mathbf{a}(u_0) = \big(a_1(u_0), \ldots, a_p(u_0)\big)^T$ is $\widehat{\mathbf{a}}(u_0)$. The theorem below provides an asymptotic representation of $\widehat{\mathbf{a}}(u_0)$ and the

asymptotic normal distribution. Let $S_{n1}(\mathbf{0},0) = (S^T_{n11}(\mathbf{0},0), S^T_{n12}(\mathbf{0},0))^T$, where $S_{n11}(\mathbf{0},0)$ and $S_{n12}(\mathbf{0},0)$ are both $p \times 1$ vectors.

Theorem 5.2. *Suppose that Conditions (C1)—(C4) in the Appendix hold. Then we have the asymptotic representation*

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0)\right] = -\gamma_n^{-2}[4\tau f^2(u_0)\Sigma(u_0)]^{-1} S_{n11}(\mathbf{0},0) + o_P(1), \quad (5.9)$$

where $f(u)$ is the density function of U. Furthermore,

$$\sqrt{nh}\left[\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) - \frac{\mu_2}{2}\mathbf{a}''(u_0)h^2 + o(h^2)\right] \to \mathbf{N}\left(0, \frac{\nu_0}{12\tau^2 f(u_0)}\Sigma^{-1}(u_0)\right) \quad (5.10)$$

in distribution, where $\mathbf{a}''(u_0) = (a_1''(u_0), \ldots, a_p''(u_0))^T$.

Remark. For the estimators of the derivatives of the coefficient functions, we have the following asymptotic representations:

$$\sqrt{nh^3}\left[\widehat{\alpha}_2 - a_0'(u_0)\right] = -\gamma_n^{-2}[4\tau f^2(u_0)\mu_2]^{-1} S_{n2}(\mathbf{0},0) + o_P(1), \quad (5.11)$$

$$\sqrt{nh^3}\left[\widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0)\right] = -\gamma_n^{-2}[4\tau f^2(u_0)\mu_2\Sigma(u_0)]^{-1} S_{n12}(\mathbf{0},0) + o_P(1). \quad (5.12)$$

Following a similar proof to that for Theorem 5.2 in the appendix, it can be shown that $\sqrt{nh^3}\left[\widehat{\alpha}_{n2} - a_0'(u_0)\right]$ and $\sqrt{nh^3}\left[\widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0)\right]$ are both asymptotically normal. The proof of the asymptotic normality of $\hat{\alpha}_2$ and $\hat{\mathbf{a}}'(u_0)$ is given in Appendix B.

5.3.2 Asymptotic relative efficiency

We now compare the estimation efficiency of the local rank estimator (denoted by $\widehat{\mathbf{a}}_R(u_0)$) with that of the local linear least squares estimator (denoted by $\widehat{\mathbf{a}}_{LS}(u_0)$) for estimating $\mathbf{a}(u_0)$ in the varying coefficient model. To measure efficiency, we consider both the asymptotic mean squared error (MSE) at a given

u_0 and the asymptotic mean integrated squared error (MISE) to assess the global performance. When evaluating both criteria, we plug in the theoretical optimal bandwidth.

Zhang and Lee (2000) gives the asymptotic MSE of $\widehat{\mathbf{a}}_{LS}(u_0)$ for estimating $\mathbf{a}(u_0)$:

$$\mathrm{MSE}_{LS}(h;u_0) = E\|\widehat{\mathbf{a}}_{LS}(u_0) - \mathbf{a}(u_0)\|^2 = \frac{\mu_2^2\|\mathbf{a}''(u_0)\|^2}{4}h^4 + \frac{\nu_0\sigma^2}{f(u_0)}\mathrm{tr}\{\Sigma^{-1}(u_0)\}\frac{1}{nh},$$

where $\sigma^2 = \mathrm{var}(\epsilon)$ is assumed to be finite and positive. Thus, the theoretical optimal bandwidth, which minimizes the asymptotic MSE of $\widehat{\mathbf{a}}_{LS}(u_0)$, is

$$h_{LS}^{opt}(u_0) = \left[\frac{\nu_0\sigma^2\mathrm{tr}\{\Sigma^{-1}(u_0)\}}{\mu_2^2\|\mathbf{a}''(u_0)\|^2 f(u_0)}\right]^{1/5} n^{-1/5}. \tag{5.13}$$

From (5.10), the asymptotic MSE of the local rank estimator $\widehat{\mathbf{a}}_R(u_0)$ is

$$\mathrm{MSE}_R(h;u_0) = E\|\widehat{\mathbf{a}}_R(u_0) - \mathbf{a}(u_0)\|^2 = \frac{\mu_2^2\|\mathbf{a}''(u_0)\|^2}{4}h^4 + \frac{\nu_0}{12\tau^2 f(u_0)}\mathrm{tr}\{\Sigma^{-1}(u_0)\}\frac{1}{nh}.$$

The theoretical optimal bandwidth for the local rank estimator thus is

$$h_R^{opt}(u_0) = \left[\frac{\nu_0\mathrm{tr}\{\Sigma^{-1}(u_0)\}}{12\tau^2\mu_2^2\|\mathbf{a}''(u_0)\|^2 f(u_0)}\right]^{1/5} n^{-1/5}. \tag{5.14}$$

This allows us to calculate the local asymptotic relative efficiency.

Theorem 5.3. *The asymptotic relative efficiency of the local rank estimator to the local linear least squares estimator for* $\mathbf{a}(u_0)$ *is*

$$ARE(u_0) = \frac{MSE_{LS}\{h_{LS}^{opt}(u_0), u_0\}}{MSE_R\{h_R^{opt}(u_0), u_0\}} = (12\sigma^2\tau^2)^{4/5}.$$

This asymptotic relative efficiency has a lower bound of 0.8896, which is attained at the random error density $f(t) = \frac{3}{20\sqrt{5}}(5 - x^2)I(|x| \leq 5)$.

Remark. Alternatively, we may consider relative efficiency obtained by comparing the MISE, which is defined as $\text{MISE}(h) = \int E\|\hat{\mathbf{a}}(u) - \mathbf{a}(u)\|^2 w(u)\,du$ with a weight function $w(\cdot)$. This provides a global measurement. Interestingly, it leads to the same relative efficiency. This can be easily seen by noting that the theoretical optimal global bandwidths for the local linear least squares estimator and the local rank estimator are

$$h_{LS}^{opt} = \left[\frac{\nu_0 \sigma^2 \int w(u)\text{tr}\{\Sigma^{-1}(u)\}/f(u)\,du}{\mu_2^2 \int \|\mathbf{a}''(u)\|^2 w(u)\,du}\right]^{1/5} n^{-1/5} \tag{5.15}$$

and

$$h_{R}^{opt} = \left[\frac{\nu_0 \int w(u)\text{tr}\{\Sigma^{-1}(u)\}/f(u)\,du}{12\tau^2\mu_2^2 \int \|\mathbf{a}''(u)\|^2 w(u)\,du}\right]^{1/5} n^{-1/5}, \tag{5.16}$$

respectively. Thus, with the theoretical optimal bandwidths,

$$\text{ARE} = \frac{\text{MISE}_{LS}(h_{LS}^{opt})}{\text{MISE}_{R}(h_{R}^{opt})} = (12\sigma^2\tau^2)^{4/5}.$$

Define $\phi = (12\sigma^2\tau^2)^{4/5}$. Then $\text{ARE}(u_0) = \text{ARE} = \phi$.

Note that the above ARE is closely related to the relative efficiency of the Wilcoxon-Mann-Whitney rank test in comparison with the two-sample t-test. Table 5.1 depicts the value of ϕ for some commonly used error distributions. It can be seen that the desirable high efficiency of traditional rank methods for estimating a finite-dimensional parameter completely carries over to the local rank method for estimating an infinite dimensional parameter.

By a similar calculation, we can show that the asymptotic relative efficiencies of the local rank estimator to the local linear estimator for $\mathbf{a}'(u_0)$ and $\mathbf{a}'(\cdot)$

Table 5.1. Asymptotic relative efficiency

Error	Normal	Laplace	t_3	Exponential	Log N	Cauchy
ϕ	0.9638	1.3832	1.6711	2.4082	4.9321	∞
ψ	0.9671	1.3430	1.5949	2.2233	4.2661	∞

both equal $\psi = (12\sigma^2\tau^2)^{8/11}$, which has a lower bound of 0.8991. This value is also reported in Table 1 for some common error distributions.

5.3.3 Asymptotic normality of $\widehat{\alpha}_1$

Following (5.5), $\widehat{\alpha}_1^* = \sqrt{nh}\{\widehat{\alpha}_1 - a_0(u_0)\}$ is the value of α_1^* that minimizes

$$Q_{n0}^*(\alpha_1^*, \widehat{\alpha}_2, \widehat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n}\Big|\epsilon_i - \gamma_n\alpha_1^* - (\widehat{\alpha}_2 - a_0'(u_0))(U_i - u_0)$$
$$-\sum_{m=1}^{p}\big[(\widehat{\beta}_m - a_m(u_0)) + (\widehat{\beta}_{p+m} - a_m'(u_0))(U_i - u_0)\big]X_{im} + \Delta_i(u_0)\Big|K_h(U_i - u_0).$$

Similarly to Lemma 5.1, we can show that the following local quadratic approximation holds uniformly in a neighborhood around 0:

$$\gamma_n^{-1}Q_{n0}^*(\alpha_1^*, \widehat{\alpha}_2, \widehat{\boldsymbol{\beta}}) = \gamma_n^{-1}\alpha_1^* S_{n0} + \gamma_n g(0)f(u_0)\alpha_1^{*2} + \gamma_n^{-1}Q_{n0}^*(0, a_0'(u_0), \boldsymbol{\beta}_0) + o_p(1),$$

where

$$S_{n0} = 2\gamma_n n^{-1}\sum_{i=1}^{n}[I(\epsilon_i \le -\Delta_i(u_0)) - 1/2]K_h(U_i - u_0). \tag{5.17}$$

This allows us to further establish an asymptotic representation of $\widehat{\alpha}_1$:

$$\sqrt{nh}(\widehat{\alpha}_1 - a_0(u_0)) = -\gamma_n^{-2}[2g(0)f(u_0)]^{-1}S_{n0} + o_p(1). \tag{5.18}$$

The theorem below gives the asymptotic distribution of $\widehat{\alpha}_1$.

Theorem 5.4. *Under the conditions of Theorem 5.2, we have*

$$\sqrt{nh}\left[\widehat{\alpha}_1 - a_0(u_0) - \frac{\mu_2 a_0''(u_0)}{2}h^2 + o(h^2)\right] \to \mathbf{N}\Big(0, \left[12g^2(0)f(u_0)\right]^{-1}\nu_0\Big).$$

5.3.4 Estimation of the standard errors

To make a statistical inference based on the local rank methodology, one needs to estimate the standard error of the resulting estimator. As indicated by Theorem 5.2, the asymptotic covariance matrix of the local rank estimator is rather complex and involves unknown functions. Here we propose a standard error estimator using a simple resampling method proposed by Jin, Ying, and Wei (2001).

Let $V_1, \ldots, V_n$ be independent and identically distributed nonnegative random variables with mean 1/2 and variance 1. We consider a stochastic perturbation of (5.4):

$$\overline{Q}_n(\boldsymbol{\beta}, \alpha_2) = \frac{1}{n(n-1)} \sum_{1\le i,j\le n} (V_i + V_j)|e_i - e_j|K_h(U_i - u_0)K_h(U_j - u_0), \quad (5.19)$$

where e_i is defined in (5.3). Note that in $\overline{Q}_n(\boldsymbol{\beta}, \alpha_2)$ the data $\{Y_i, U_i, \mathbf{X}_i\}$ are considered to be fixed, and the randomness comes from the V_i's. Let $(\overline{\beta}^T, \overline{\alpha}_2)^T$ be the value of $(\boldsymbol{\beta}^T, \alpha_2)^T$ that minimizes $\overline{Q}_n(\boldsymbol{\beta}, \alpha_2)$. It is easy to obtain $(\overline{\beta}^T, \overline{\alpha}_2)^T$ by applying a simple algorithm described in Section 3.1.

Jin, Ying, and Wei (2001) established the validity of the resampling method when the objective function has a U-statistic structure. Although their theory covers many important applications, they require that the U-statistic has a fixed kernel. We extend their result to our setting, where the U-statistic involves a variable kernel due to nonparametric smoothing. Let $\overline{\mathbf{a}}(u_0)$ be the local rank estimator of $\mathbf{a}(u_0)$ based on the perturbed objective function (5.19), i.e., it is the

subvector that consists of the first p components of $\overline{\boldsymbol{\beta}}$. Its asymptotic normality is given in the theorem below.

Theorem 5.5. *Under the conditions of Lemma 5.1, conditional on almost surely every sequence of data* $\{Y_i, U_i, \mathbf{X}_i\}$ *,*

$$\sqrt{nh}\left[\overline{\mathbf{a}}(u_0) - \widehat{\mathbf{a}}(u_0)\right] \to \mathbf{N}\left(0, \frac{\nu_0}{12\tau^2 f(u_0)}\Sigma^{-1}(u_0)\right)$$

in distribution.

This theorem suggests that to estimate the asymptotic covariance matrix of $\widehat{\mathbf{a}}(u_0)$, one can repeatedly perturb (5.4) by generating a large number of independent random samples $\{V_i\}_{i=1}^n$. For each perturbed objective function, one solves for $\overline{\mathbf{a}}(u_0)$. The sample covariance matrix of $\overline{\mathbf{a}}(u_0)$ based on a large number of independent perturbations provides a good approximation. The accuracy of the resulting standard error estimate will be tested in the next section.

The perturbed estimator has conditional bias equal to zero. It has been found that a standard bootstrap method, which resamples from the empirical distribution of the data, also estimates the bias as zero when estimating nonparametric curves (Hall and Kang 2001). It is possible to use a more delicate bootstrap technique to estimate the bias of a nonparametric curve estimator. Although some of the ideas may be adapted to the method of perturbing the objective function, this is beyond the scope of our research and is not pursued further here.

5.4 Numerical studies

5.4.1 A pseudo-observation algorithm

The local rank estimator can be obtained by applying an efficient and reliable algorithm. Note that the local rank estimator of $(\boldsymbol{\beta}_0^T, a_0'(u_0))^T$ can be found by fitting a weighted L_1 regression on $\frac{n(n-1)}{2}$ pseudo observations $(\mathbf{x}_i^* - \mathbf{x}_j^*, Y_i - Y_j)$

with weights $w_{ij} = K\big((U_i - u_0)/h\big)K\big((U_j - u_0)/h\big)$, where $\mathbf{x}_i^* = (U_i - u_0, X_i^T, (U_i - u_0)X_i^T)^T$, $1 \le i < j \le n$. Given $(\widehat{\boldsymbol{\beta}}^T, \widehat{\alpha}_2)^T$, the estimator of $a_0(u_0)$ can be obtained by another weighted L_1 regression on $(1, Y_i - \widehat{\alpha}_2(U_i - u_0) - \sum_{m=1}^{p}\big[\widehat{\beta}_m + \widehat{\beta}_{p+m}(U_i - u_0)\big]X_{im})$ with weights $w_i = K\big((U_i - u_0)/h\big)$, $1 \le i \le n$. Many statistical software packages can implement weighted L_1 regression. In our numerical studies, we use the function "rq" in the R package *quantreg*.

5.4.2 Bandwidth selection

Bandwidth selection is an important issue for all statistical models that involve nonparametric smoothing. Although we have derived the theoretical optimal bandwidth for the local rank estimator in (5.14) and (5.16), it is difficult to use the "plug-in" method to estimate it due to many unknown quantities.

We propose below an alternative bandwidth selection method that is practically feasible. This approach is based on the relationship between h_R^{opt} and h_{LS}^{opt}. From Section 2.3, we see that

$$h_R^{opt}(u_0) = \left(\frac{1}{12\tau^2\sigma^2}\right)^{1/5} h_{LS}^{opt}(u_0) \quad \text{and} \quad h_R^{opt} = \left(\frac{1}{12\tau^2\sigma^2}\right)^{1/5} h_{LS}^{opt}. \tag{5.20}$$

Thus, we can first use existing bandwidth selectors (e.g. Zhang and Lee 2000) to estimate $h_{LS}^{opt}(u_0)$ or h_{LS}^{opt}, and then use the residuals from local least squares fitting to estimate σ^2 and τ. See Hettmansperger and McKean (1998, p.181) for more details on how to estimate τ, which can be obtained by the function "wilcoxontau" in the R software developed by Terpstra and McKean (2005). In the end, we plug these estimators into (5.20) to get the selected bandwidth for the local rank estimator.

5.4.3 Examples

We conduct Monte Carlo simulations to access the finite sample performance of the proposed procedures and illustrate the proposed methodology by an analysis of a real environmental data set. In the analysis, the Epanechnikov kernel $K(u) = .75(1-u^2)I(|u|<1)$ is used.

Example 5.4.1. We generate random data from

$$Y = a_0(U) + a_1(U)X_1 + a_2(U)X_2 + \epsilon,$$

where $a_0(U) = \exp(2U-1)$, $a_1(U) = 8U(1-U)$ and $a_2(U) = 2\sin^2(2\pi U)$. The covariate U follows a uniform distribution on $[0,1]$, and is independent from (X_1, X_2), where the covariates X_1 and X_2 are standard normal random variables with correlation coefficient $2^{-1/2}$. The coefficient functions and the mechanism to generate U and (X_1, X_2) were used in Cai, Fan, and Li (2000). In this example, we consider six error distributions: $N(0,1)$, Laplace, standard Cauchy, t-distribution with 3 degrees of freedom, mixture of normals $0.9N(0,1) + 0.1N(0,10^2)$, and lognormal distribution. Except for the Cauchy error, all other error distributions are standardized to have median 0 and variance 1. To make a fair comparison with the local least squares method, we set the bandwidth to be the theoretical optimal value h^{opt} for both the local rank estimator and the local least squares estimator. Optimal bandwidths are calculated using (5.15) and (5.16). To demonstrate that the proposed methodology performs well with a wide range of bandwidths, we also consider the undersmoothing case by setting the bandwidth to be $0.5h^{opt}$ and the oversmoothing scenario by taking the bandwidth to be $2h^{opt}$. In our simulation, we consider the sample sizes n =400 and 800, and we conduct 400 simulations for each case.

We compare the performance of the proposed local rank estimate with the local least squares estimate using the square root of average squared errors (RASE),

defined by

$$\text{RASE} = \left\{ \frac{1}{n_{\text{grid}}} \sum_{m=1}^{p} \sum_{k=1}^{n_{\text{grid}}} \{\hat{a}_m(u_k) - a_m(u_k)\}^2 \right\}^{1/2},$$

where $\{u_k : k = 1, \cdots, n_{\text{grid}}\}$ is a set of grid points uniformly placed on $[0, 1]$ with $n_{\text{grid}} = 200$.

Table 5.2. Summary of the RASE over 400 simulations. LS denotes the local least squares estimator and R denotes the local rank estimator.

h		Normal	Laplace	Cauchy	t_3	Mixture	Log-Normal
				$n = 400$			
$.5h^{opt}$	LS	.431(.079)	.423(.085)	32.4(173)	.419(.136)	.423(.121)	.425(.215)
	R	.450(.088)	.420(.094)	.968(.549)	.382(.098)	.343(.333)	.405(.342)
h^{opt}	LS	.311(.066)	.307(.068)	21.7(109)	.304(.108)	.305(.087)	.296(.117)
	R	.321(.069)	.280(.064)	.564(.169)	.249(.053)	.161(.060)	.191(.101)
$2h^{opt}$	LS	.404(.052)	.400(.052)	14.6(58.0)	.400(.068)	.399(.057)	.398(.071)
	R	.402(.053)	.344(.048)	.597(.103)	.313(.043)	.194(.026)	.205(.037)
				$n = 800$			
$.5h^{opt}$	LS	.295(.045)	.298(.051)	19.1(5.59)	.288(.083)	.296(.059)	.289(.072)
	R	.303(.046)	.277(.045)	.548(.126)	.243(.038)	.164(.054)	.187(.035)
h^{opt}	LS	.225(.044)	.223(.043)	13.8(38.3)	.217(.063)	.224(.047)	.222(.058)
	R	.230(.045)	.199(.040)	.386(.091)	.176(.033)	.106(.020)	.119(.023)
$2h^{opt}$	LS	.313(.036)	.312(.036)	1.57(31.1)	.310(.041)	.313(.036)	.313(.040)
	R	.312(.037)	.267(.035)	.470(.063)	.242(.028)	.147(.017)	.152(.019)

The sample mean and the sample standard deviation of the RASEs over 400 simulations are presented in Table 5.2, in which the value in the parenthesis is the standard deviation. Table 5.2 clearly demonstrates that the local rank estimator performs almost as well as the local least squares estimator when the random error is normally distributed; and has smaller RASE than the local least squares estimator for other error distributions. The efficiency gain can be substantial. For

example, for the mixture error distribution, the observed relative efficiency of the local rank estimator to the local least squares estimator is $(0.305/0.161)^2 = 3.5888$ for $n = 400$, and is $(0.224/0.106)^2 = 4.4656$ for $n = 800$. Note that for the Cauchy random error, the local least squares method yields an inconsistent estimator, however the local rank estimator still results in a $\sqrt{n}$-consistent estimator. This explains why the RASE of the local least squares estimator is very large.

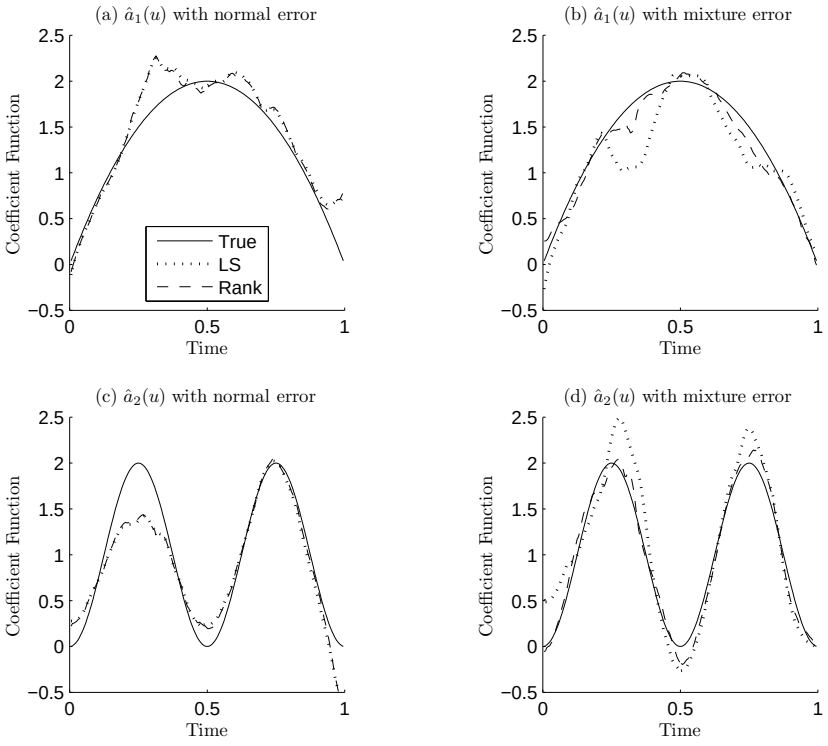


Fig. 5.1. Plots of estimated coefficient functions for a typical data set

Figure 5.1 depicts the estimated coefficient functions for the normal random error and the mixture random error for a typical sample, which is selected in such a way that its RASE value is the median of the 400 RASE values. From Figure 5.1 (a) and (c), it can be seen that the resulting local least squares estimator and the local rank estimator are almost identical when the random error is normal. From

Figure 5.1 (b) and (d), we observe that the bias of the local rank estimator is smaller than that of the local least squares estimator. Furthermore, the local rank estimator can improve over the local least squares estimator in terms of variance, as shown in Figure 5.2, which plots the estimated coefficient functions for all 400 simulations when the random error has a mixture of normals distribution.

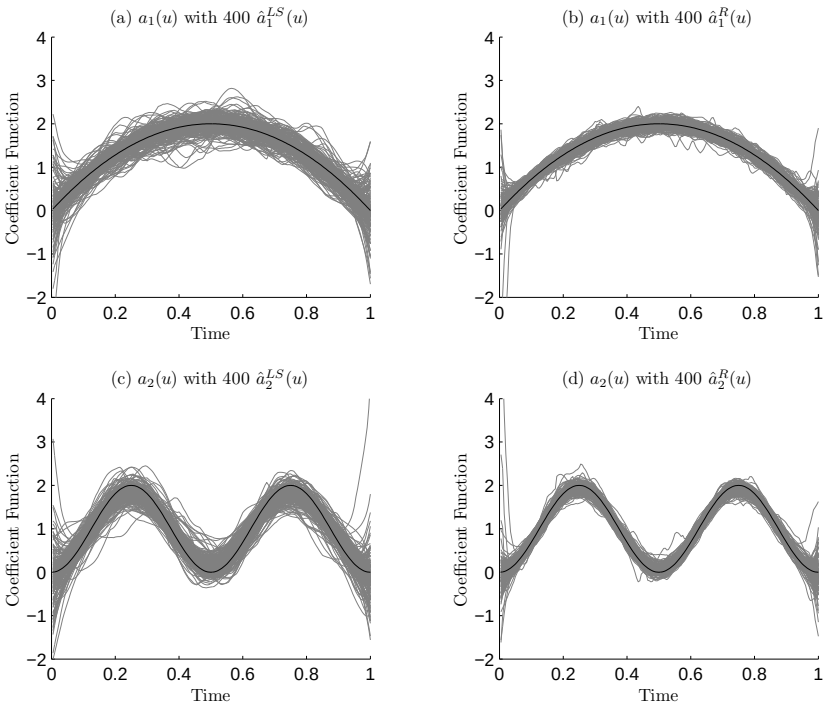


Fig. 5.2. (a) and (c) are plots of 400 local least squares estimators of $a_1(\cdot)$ and $a_2(\cdot)$ over 400 simulations, respectively. (b) and (d) are plots of 400 local rank estimators of $a_1(\cdot)$ and $a_2(\cdot)$, respectively.

We now test the accuracy of the standard error estimator proposed in Section 2.5. We randomly perturb the objective function 1000 times; each time the random variables V_i in (5.19) are generated from the $Gamma(0.25, 2)$ distribution. Table 5.3 summarizes the simulation results at three points, $u_0 = 0.25$, 0.50 and 0.75. In the table, 'SD' denotes the standard deviation of 400 estimated $\hat{a}_m(u_0)$ and can be regarded as the true standard error; 'SE(std(SE))' denotes the mean

Table 5.3. Standard deviations of the local rank estimators with $n = 400$

		$\hat{a}_1(u)$		$\hat{a}_2(u)$	
Error	u_0	SD	$SE(Std(SE))$	SD	$SE(Std(SE))$
Normal	0.25	0.189	0.159(0.032)	0.197	0.160(0.032)
	0.5	0.183	0.159(0.030)	0.180	0.162(0.031)
	0.75	0.191	0.162(0.033)	0.195	0.163(0.032)
Laplace	0.25	0.175	0.151(0.037)	0.174	0.151(0.037)
	0.5	0.168	0.153(0.039)	0.173	0.154(0.039)
	0.75	0.168	0.150(0.037)	0.177	0.150(0.037)
Mixture	0.25	0.095	0.107(0.051)	0.092	0.107(0.049)
	0.5	0.095	0.109(0.057)	0.091	0.109(0.055)
	0.75	0.095	0.108(0.061)	0.093	0.109(0.055)
t_3	0.25	0.144	0.137(0.039)	0.145	0.138(0.036)
	0.5	0.148	0.133(0.035)	0.152	0.136(0.037)
	0.75	0.158	0.137(0.039)	0.155	0.139(0.042)
Log N	0.25	0.111	0.112(0.047)	0.112	0.114(0.049)
	0.5	0.106	0.114(0.047)	0.107	0.119(0.050)
	0.75	0.118	0.117(0.058)	0.118	0.120(0.060)

(standard deviation) of 400 estimated standard errors from the resampling method. Bandwidths are set to be the optimal ones. Table 5.3 indicates that the proposed resampling method estimates the standard error very well. The true standard deviations all fall within one standard deviation away from the estimated standard errors.

Example 5.4.2. As an illustration, we apply our proposed procedure to the environmental data set analyzed in Fan and Zhang (1999). This data set was collected in Hong Kong from January 1, 1994 to December 31, 1995. An objective of the study is to understand the association between levels of pollutants and the number of total hospital admissions for circulatory and respiratory problems. The covariates considered here are the level of sulfur dioxide (X_1), the level of nitrogen dioxide (X_2) and the level of dust (X_3), and the response is taken to be the logarithm of the number of total hospital admissions. A scatter plot of the response variable over time is shown in Figure 5.3(a). Here we analyze this data set with a varying coefficient model

$$Y = a_0(u) + a_1(u)X_1 + a_2(u)X_2 + a_3(u)X_3 + \epsilon,$$

where u denotes time and is scaled to the interval [0,1].

We select the bandwidth via the relation (5.20). More specifically, we first use leave-one-out cross validation to select a bandwidth h_{LS} for the local least squares estimator. We then use the kernel density estimate to infer the error density $f(\cdot)$ based on the residuals from the local least squares estimator and estimate $(12\sigma^2\tau^2)^{-1}$. This leads to the selected bandwidth for the local rank estimator: $h_R = 0.06$.

The estimated coefficient functions are depicted in Figures 5.3(b), (c) and (d), where the two dashed curves around the solid line are the estimated function plus/minus twice the standard errors estimated by the resampling method. These

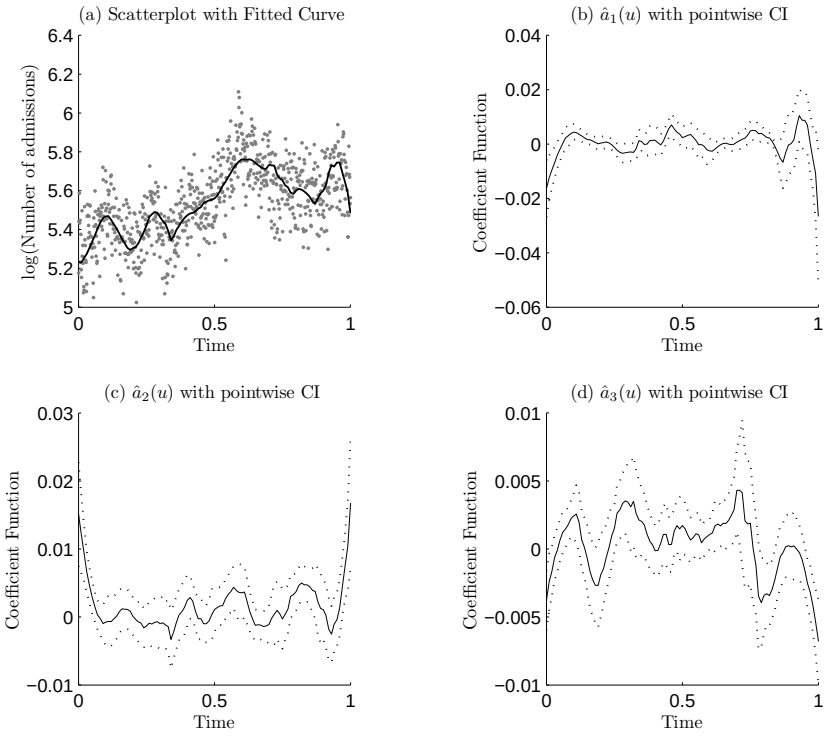


Fig. 5.3. (a) Scatterplot of the log of number of total hospital admissions over time, and the solid curve is an estimator of the expected number of hospital admissions over time at the average pollutant levels, i.e., $\hat{a}_0(u)+\hat{a}_1(u)\bar{X}_1+\hat{a}_2(u)\bar{X}_2+\hat{a}_3(u)\bar{X}_3$. (b), (c) and (d) are the estimated coefficient functions via the local rank estimator for $a_k(\cdot)$, $k = 1$, 2, and 3, respectively.

two dashed lines can be regarded as a pointwise confidence interval with bias ignored. Figure 5.3(a) indicates a clearly increasing trend with some seasonal pattern in the number of hospital admissions.

5.5 Proofs

5.5.1 Proofs of the main theorems

We first impose some regularity conditions. These conditions are used to facilitate the proofs, but may not be the weakest ones.

Regularity conditions:

(C1). Assume that $\{U_i, \mathbf{X}_i, Y_i\}$ are independent and identically distributed from model (5.1). Furthermore, the random error ϵ and covariate $\{U, \mathbf{X}\}$ are independent. Assume that ϵ has probability density function $g(\cdot)$ which has finite Fisher information, i.e., $\int \{g(x)\}^{-1} g'(x)^2 dx < \infty$; and U has probability density function $f(\cdot)$.

(C2). The function $a_m(\cdot)$, $m = 0, 1, \ldots, p$, has continuous second-order derivative in a neighborhood of u_0 .

(C3). Assume that $E(X_i|U_i = u_0) = 0$ and that $\Sigma(u) = E(X_i X_i^T|U_i = u)$ is continuous at $u = u_0$. The matrix $\Sigma(u_0)$ is positive definite.

(C4). The kernel function $K(\cdot)$ is symmetric about the origin and has a bounded support. Assume that $h \to 0$ and $nh^2 \to \infty$, as $n \to \infty$.

In our proofs, we will use some results on generalized U-statistics, where the kernel function in the U-statistic is allowed to depend on the sample size n. The generalized U-statistic has the form $U_n = [n(n-1)]^{-1} \sum \sum_{i \neq j} H_n(D_i, D_j)$, where $\{D_i\}_{i=1}^n$ is a random sample and H_n is symmetric in its arguments, i.e., $H_n(D_i, D_j) = H_n(D_j, D_i)$. In this chapter, $D_i = (\mathbf{X}_i^T, U_i, \epsilon_i)^T$. Define $r_n(D_i) =$

$E[H_n(D_i, D_j)|D_i]$, $\overline{r}_n = E[r_n(D_i)]$ and $\widehat{U}_n = \overline{r}_n + 2n^{-1}\sum_{i=1}^{n}[r_n(D_i) - \overline{r}_n]$. We will repeatedly use the following lemma taken from Powell, Stock, and Stoker (1989).

Lemma 5.6. *If* $E[||H_n(D_i, D_j)||^2] = o(n)$, *then* $\sqrt{n}(U_n - \widehat{U}_n) = o_p(1)$ *and* $U_n = \overline{r}_n + o_p(1)$.

We need the following two lemmas to prove Lemma 5.1. Denote

$$\begin{aligned}
\mathbf{A}_{n11} &= 2h^{-2}E\left\{(\mathbf{Z}_i - \mathbf{Z}_j)(\mathbf{Z}_i - \mathbf{Z}_j)^T K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\}, \\
\mathbf{A}_{n12} &= 2h^{-2}E\left\{(\mathbf{Z}_i - \mathbf{Z}_j)[(U_i - U_j)/h]K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\}, \\
\mathbf{A}_{n21} &= \mathbf{A}_{n12}^T, \\
A_{n22} &= 2h^{-2}E\left\{[(U_i - U_j)^2/h^2]K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\},
\end{aligned}$$

and define

$$\mathbf{A}_n = \tau \begin{pmatrix} \mathbf{A}_{n11} & \mathbf{A}_{n12} \\ \mathbf{A}_{n21} & A_{n22} \end{pmatrix}.$$

Lemma 5.7. *Suppose that Conditions (C1)—(C4) hold. Then* $\mathbf{A}_n \to \mathbf{A}$, *where* $\mathbf{A}$ *is defined in (5.8).*

Proof. We can write $\mathbf{A}_{n11} = \begin{pmatrix} \mathbf{A}_{n11}^1 & \mathbf{A}_{n11}^2 \\ \mathbf{A}_{n11}^3 & A_{n11}^4 \end{pmatrix}$. Let

$$\mathbf{A}_{n11}^1 = 2h^{-2}E\left[(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right].$$

Calculating the expectation by conditional on U_i and U_j first, $\mathbf{A}_{n11}$ becomes

$$2h^{-2}\int E\big[(\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T | U_i = u, U_j = v\big] K\left(\frac{u - u_0}{h}\right) K\left(\frac{v - u_0}{h}\right) f(u) f(v)\, du\, dv.$$

Using Condition (C3), straightforward calculation gives $\mathbf{A}^1_{n11} \to 4f^2(u_0)\Sigma(u_0)$. Let

$$\mathbf{A}^2_{n11} = 2h^{-2}E\left\{(\mathbf{X}_i - \mathbf{X}_j)\big[\mathbf{X}_i(U_i - u_0)/h - \mathbf{X}_j(U_j - u_0)/h\big]^T K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\}.$$

Using Condition (C3) and noticing that $K(\cdot)$ is symmetric, it can be shown that $\mathbf{A}^2_{n11} \to \mathbf{0}$. By symmetry, $\mathbf{A}^3_{n11} \to \mathbf{0}$. Similarly, we have

$$\begin{aligned}\mathbf{A}^4_{n11} &= 2h^{-2}E\Bigg\{\big[\mathbf{X}_i(U_i - u_0)/h - \mathbf{X}_j(U_j - u_0)/h\big]\big[\mathbf{X}_i(U_i - u_0)/h - \mathbf{X}_j(U_j - u_0)/h\big]^T \\ &\qquad K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\Bigg\} \\ &\to 4f^2(u_0)\Sigma(u_0)\mu_2.\end{aligned}$$

Thus $\mathbf{A}_{n11} \to 4f^2(u_0)\Sigma(u_0)\begin{pmatrix}\mathbf{I}_p & \mathbf{0}\\ \mathbf{0} & \mu_2\mathbf{I}_p\end{pmatrix}$. Similarly, we can show that $\mathbf{A}_{n12} = \mathbf{A}^T_{n21} \to \mathbf{0}$, and

$$A_{n22} = 2\int (t_1 - t_2)^2 K(t_1)K(t_2)f(u_0 + t_1h)f(u_0 + t_2h)dt_1dt_2 \to 4f^2(u_0)\mu_2. \quad \square$$

Lemma 5.8. *Under Conditions (C1)—(C4), we have*

$$\gamma_n^{-1}[S_n(\boldsymbol{\beta}^*, \alpha_2^*) - S_n(\mathbf{0}, 0)] = \gamma_n\mathbf{A}\begin{pmatrix}\boldsymbol{\beta}^*\\ \alpha_2^*\end{pmatrix} + o_p(1).$$

Proof. Let $U_n = \gamma_n^{-1}[S_n(\boldsymbol{\beta}^*, \alpha_2^*) - S_n(\mathbf{0}, 0)]$. Write $U_n = [n(n-1)]^{-1}\sum\sum_{i\neq j} W_n(D_i, D_j)$, where

$$\begin{aligned}W_n(D_i, D_j) &= 2\big[I\left(\epsilon_i - \gamma_n\alpha_2^*(U_i - u_0)/h - \gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0) \le \epsilon_j - \gamma_n\alpha_2^*(U_j - u_0)/h\right. \\ &\qquad \left. -\gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\right) - 1/2]\begin{pmatrix}\mathbf{Z}_i - \mathbf{Z}_j\\ (U_i - U_j)/h\end{pmatrix}K_h(U_i - u_0)K_h(U_j - u_0).\end{aligned}$$

Let $H_n(D_i, D_j) = [W_n(D_i, D_j) + W_n(D_j, D_i)]/2$, then $U_n = [n(n-1)]^{-1} \sum\sum_{i \neq j} H_n(D_i, D_j)$ has the form of a generalized U-statistic. We next check the condition of Lemma 5.6. Note that

$$E[||H_n(D_i, D_j)||^2] \leq \frac{1}{2}E[||W_n(D_i, D_j)||^2] + \frac{1}{2}E[||W_n(D_i, D_j)||^2] = E[||W_n(D_i, D_j)||^2].$$

Furthermore,

$$\begin{aligned}
& E[||W_n(D_i, D_j)||^2] \\
\leq\ & 4h^{-4}E\Big\{\Big[(\mathbf{Z}_i - \mathbf{Z}_j)^T(\mathbf{Z}_i - \mathbf{Z}_j) + [(U_i - U_j)/h]^2\Big] K^2\left(\frac{U_i - u_0}{h}\right) K^2\left(\frac{U_j - u_0}{h}\right)\Big\} \\
=\ & O(h^{-2}) = o(n)
\end{aligned}$$

as $nh^2 \to \infty$ by assumption. Thus, by Lemma 5.6, $U_n = E[H_n(D_i, D_j)] + o_p(1)$. Note that $E[H_n(D_i, D_j)] = E[W_n(D_i, D_j)]$. Thus,

$$\begin{aligned}
& E[H_n(D_i, D_j)] \\
=\ & 2h^{-2}E\Bigg\{\int \Big[G\Big(\epsilon + \Delta_j(u_0) - \Delta_i(u_0) - \gamma_n\alpha_2^*(U_j - U_i)/h - \gamma_n\boldsymbol{\beta}^{*T}(\mathbf{Z}_j - \mathbf{Z}_i)\Big) - G(\epsilon)\Big] \\
& g(\epsilon)d\epsilon \begin{pmatrix} \mathbf{Z}_i - \mathbf{Z}_j \\ (U_i - U_j)/h \end{pmatrix} K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\Bigg\} \\
=\ & 2h^{-2}\gamma_n E\Bigg\{\int g\Big[\epsilon + \Delta_j(u_0) - \Delta_i(u_0)\Big] g(\epsilon)d\epsilon \begin{pmatrix} \mathbf{Z}_i - \mathbf{Z}_j \\ (U_i - U_j)/h \end{pmatrix} \\
& \Big(\mathbf{Z}_i^T - \mathbf{Z}_j^T, (U_i - U_j)/h\Big) K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\Bigg\} \begin{pmatrix} \boldsymbol{\beta}^* \\ \alpha_2^* \end{pmatrix} (1 + o(1)) \\
=\ & \gamma_n \mathbf{A}_n \begin{pmatrix} \boldsymbol{\beta}^* \\ \alpha_2^* \end{pmatrix} \{1 + o(1)\}.
\end{aligned}$$

The proof is completed by using Lemma 5.7. □

Proof of Lemma 5.1. In view of Lemma 5.8, it follows that

$$\bigtriangledown \left[\gamma_n^{-1} Q_n^*(\boldsymbol{\beta}^*, \alpha_2^*) - B_n(\boldsymbol{\beta}^*, \alpha_2^*)\right] = \gamma_n^{-1}[S_n(\boldsymbol{\beta}^*, \alpha_2^*) - S_n(\mathbf{0}, 0)] - \gamma_n \mathbf{A} \begin{pmatrix} \boldsymbol{\beta}^* \\ \alpha_2^* \end{pmatrix} = o_p(1).$$

The proof follows along the same lines as the proof of Theorem A.3.7. of Hettmansperger and McKean (1998), which uses a "diagonal subsequencing" argument and properties of convex functions. □

Proof of Theorem 5.2. By Lemma 5.1, $\gamma_n^{-1} Q_n^*(\mathbf{s}_1, s_2) = B_n(\mathbf{s}_1, s_2) + r_n(\mathbf{s}_1, s_2)$, where $r_n(\mathbf{s}_1, s_2) \xrightarrow{p} 0$ uniformly over any bounded set. Note that $\gamma_n^{-1} Q_n^*(\mathbf{s}_1, s_2)$ is minimized by $\left(\widehat{\boldsymbol{\beta}}_n^{*T}, \widehat{\alpha}_{2n}^*\right)^T$, and $B_n(\mathbf{s}_1, s_2)$ is minimized by $\left(\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*\right)^T = -\gamma_n^{-2} \mathbf{A}^{-1} (S_{n1}^T(\mathbf{0}, 0), S_{n2}(\mathbf{0}, 0))^T$. We first establish the asymptotic representation by following a similar argument to Hjort and Pollard (1993). For any constant $c > 0$, define

$$\begin{aligned} T_n &= \inf_{||(\mathbf{s}_1^T, s_2) - (\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)|| = c} B_n(\mathbf{s}_1, s_2) - B_n(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*) \\ R_n &= \sup_{||(\mathbf{s}_1^T, s_2) - (\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)|| \le c} |\gamma_n^{-1} Q_n^*(\mathbf{s}_1, s_2) - B_n(\mathbf{s}_1, s_2)|, \end{aligned}$$

then $R_n \xrightarrow{p} 0$ as $n \to \infty$. Let $(\mathbf{s}_1^T, s_2)^T$ be an arbitrary point outside the ball $\left\{(\mathbf{s}_1^T, s_2)^T : ||(\mathbf{s}_1^T, s_2) - (\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)|| \le c\right\}$, then we can write $(\mathbf{s}_1^T, s_2)^T = (\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)^T + l\mathbf{1}_{2p+1}$, where $l > c$ is a positive constant and $\mathbf{1}_d$ denotes a unit vector of length d. Write

$$\begin{aligned} & \frac{c}{l}\left[\gamma_n^{-1} Q_n^*(\mathbf{s}_1, s_2) - \gamma_n^{-1} Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*)\right] \\ = \; & \frac{c}{l}\gamma_n^{-1} Q_n^*(\mathbf{s}_1, s_2) + \left(1 - \frac{c}{l}\right)\gamma_n^{-1} Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*) - \gamma_n^{-1} Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*). \quad (5.21) \end{aligned}$$

By the convexity of $\gamma_n^{-1}Q_n^*(\mathbf{s}_1, s_2)$, we have

$$\begin{aligned} & \frac{c}{l}\left[\gamma_n^{-1}Q_n^*(\mathbf{s}_1, s_2) - \gamma_n^{-1}Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*)\right] \\ \geq \; & \gamma_n^{-1}Q_n^*\left(\frac{c}{l}(\mathbf{s}_1, s_2) + \left(1 - \frac{c}{l}\right)(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*)\right) - \gamma_n^{-1}Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*). \end{aligned} \tag{5.22}$$

Thus,

$$\begin{aligned} & \frac{c}{l}\left[\gamma_n^{-1}Q_n^*(\mathbf{s}_1, s_2) - \gamma_n^{-1}Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*)\right] \geq \gamma_n^{-1}Q_n^*\left(\widetilde{\boldsymbol{\beta}}_n^* + c\mathbf{1}_{2p}, \widetilde{\alpha}_{2n}^* + c\right) - \gamma_n^{-1}Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*) \\ = \; & B_n(\widetilde{\boldsymbol{\beta}}_n^* + c\mathbf{1}_{2p}, \widetilde{\alpha}_{2n}^* + c) + r_n(\widetilde{\boldsymbol{\beta}}_n^* + c\mathbf{1}_{2p}, \widetilde{\alpha}_{2n}^* + c) - B_n(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*) - r_n(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*) \\ \geq \; & T_n - 2R_n. \end{aligned}$$

If $R_n \leq \frac{1}{2}T_n$, then $\gamma_n^{-1}Q_n^*(\mathbf{s}_1, s_2) > \gamma_n^{-1}Q_n^*(\widetilde{\boldsymbol{\beta}}_n^*, \widetilde{\alpha}_{2n}^*)$ for all $(\mathbf{s}_1^T, s_2)^T$ outside the ball. This implies the if $R_n \leq \frac{1}{2}T_n$, then the minimizer of $\gamma_n^{-1}Q_n^*$ must be inside the ball. Thus,

$$P\left(\|(\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)^T - (\widehat{\boldsymbol{\beta}}_n^{*T}, \widehat{\alpha}_{2n}^*)^T\| \geq c\right) \leq P\left(R_n \geq \frac{1}{2}T_n\right) = P\left(R_n \geq \frac{1}{2}\lambda c^2\right) \to 0,$$

where λ is the smallest eigenvalue of $\mathbf{A}$. Therefore, $(\widehat{\boldsymbol{\beta}}_n^{*T}, \widehat{\alpha}_{2n}^*)^T = (\widetilde{\boldsymbol{\beta}}_n^{*T}, \widetilde{\alpha}_{2n}^*)^T + o_p(1)$. This in particular implies the asymptotic representations (5.9), (5.11) and (5.12).

We next show the asymptotic normality of $\widehat{\mathbf{a}}(u_0)$. From (5.9), we have

$$\sqrt{nh}\big(\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0)\big) = -\gamma_n^{-2}\big(4\tau f^2(u_0)\Sigma(\mu_0)\big)^{-1}S_{n11}(\mathbf{0}, 0) + o_p(1), \tag{5.23}$$

where

$$\begin{aligned} S_{n11}(\mathbf{0}, 0) \; = \; & 2\gamma_n[n(n-1)]^{-1}\sum_{i\neq j}\left[I\left(\epsilon_i + \Delta_i(u_0) \leq \epsilon_j + \Delta_j(u_0)\right) - 1/2\right](\mathbf{X}_i - \mathbf{X}_j) \\ & K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned} \tag{5.24}$$

By (5.24), let us rewrite $-\gamma_n^2 S_{n11}(\mathbf{0},0) = S_{n\mathbf{a}1}(\mathbf{0},0) + S_{n\mathbf{a}2}(\mathbf{0},0)$, where

$$\begin{aligned} S_{n\mathbf{a}}(\mathbf{0},0) &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\left[I\left(\epsilon_i \leq \epsilon_j\right) - 1/2\right](\mathbf{X}_j - \mathbf{X}_i)K_h(U_i - u_0)K_h(U_j - u_0), \\ S_{n\mathbf{b}}(\mathbf{0},0) &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\big[I\big(\epsilon_i + \Delta_i(u_0) \leq \epsilon_j + \Delta_j(u_0)\big) - I\big(\epsilon_i \leq \epsilon_j\big)\big](\mathbf{X}_j - \mathbf{X}_i) \\ &\quad K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned}$$

We next prove that

$$S_{n\mathbf{a}}(\mathbf{0},0) \to N\left(0, \frac{4}{3}f^3(u_0)\nu_0\Sigma(u_0)\right) \quad \text{in distribution.} \tag{5.25}$$

Note that we can write $S_{n\mathbf{a}}(\mathbf{0},0) = \sqrt{n}[n(n-1)]^{-1}\sum_{i\neq j} H_n(D_i, D_j)$, where $H_n(D_i, D_j) = W_n(D_i, D_j) + W_n(D_j, D_i)$ with

$$W_n(D_i, D_j) = h^{-3/2}\left[I\left(\epsilon_i \leq \epsilon_j\right) - 1/2\right](\mathbf{X}_j - \mathbf{X}_i)K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right).$$

Similarly to the arguments in the proof of Lemma 5.8, it can be shown that $E[||H_n(D_i, D_j)||^2] = o(n)$. By Lemma 5.6, this implies that $S_{n\mathbf{a}}(\mathbf{0},0) = 2n^{-1}\sum_{i=1}^n r_n(D_i) + o_p(1)$ since it is easy to check that $\overline{r}_n = 0$. We have

$$\begin{aligned} r_n(D_i) &= E[H_n(D_i, D_j)|D_i] \\ &= 2h^{-3/2}\big[G(\epsilon_i) - 1/2\big]K\left(\frac{U_i - u_0}{h}\right)E\left\{(\mathbf{X}_i - \mathbf{X}_j)K\left(\frac{U_j - u_0}{h}\right)\Big|\mathbf{X}_i, U_i, \epsilon_i\right\} \\ &= 2h^{-1/2}[G(\epsilon_i) - 1/2]K\left(\frac{U_i - u_0}{h}\right)\left[\left(\int K(t)f(u_0 + th)dt\right)\mathbf{X}_i\right. \\ &\quad \left. - \int E(X_j|U_j = u_0 + th)K(t)f(u_0 + th)dt\right]. \end{aligned}$$

Furthermore,

$$\begin{aligned} & E\left[r_n(D_i)r_n(D_i)^T\right] \\ = \; & \frac{1}{3}h^{-1}E\Bigg\{K^2\left(\frac{U_i-u_0}{h}\right)\Big[\left(\int K(t)f(u_0+th)dt\right)\mathbf{X}_i \\ & -\int E(X_j|U_j=u_0+th)K(t)f(u_0+th)dt\Big] \\ & \left[\left(\int K(t)f(u_0+th)dt\right)\mathbf{X}_i^T-\int E(X_j^T|U_j=u_0+th)K(t)f(u_0+th)dt\right]\Bigg\} \\ \to \; & \frac{1}{3}f^3(u_0)\nu_0\Sigma(u_0). \end{aligned}$$

To prove the asymptotic normality of $S_{n\mathbf{a}}(\mathbf{0},0)$, it is sufficient to check the Lindeberg-Feller condition: $\forall\ \epsilon>0$, $n^{-1}\sum_{i=1}^n E\{r_n(D_i)r_n(D_i)^T I(||r_n(D_i)||>\epsilon\sqrt{n})\}\to 0$. This can be easily verified by applying the dominated convergence theorem.

Next we show that

$$S_{n\mathbf{b}}(\mathbf{0},0)=\frac{2h^2}{\gamma_n}\left[\tau f^2(u_0)\mu_2\Sigma(u_0)\mathbf{a}''(u_0)+o(1)\right]+o_p(1). \tag{5.26}$$

We may write $S_{n\mathbf{b}}(\mathbf{0},0)=[n(n-1)]^{-1}\sum_{i\neq j}H_n^*(D_i,D_j)$, where $H_n^*(D_i,D_j)=W_n^*(D_i,D_j)+W_n^*(D_j,D_i)$ with

$$\begin{aligned} W_n^*(D_i,D_j) \;=\; & nh^{-1}\gamma_n\left[I\left(\epsilon_i+\Delta_i(u_0)\le\epsilon_j+\Delta_j(u_0)\right)-I\left(\epsilon_i\le\epsilon_j\right)\right](\mathbf{X}_j-\mathbf{X}_i) \\ & K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right). \end{aligned}$$

Note that

$$\begin{aligned} & \Delta_j(u_0)-\Delta_i(u_0) \\ = \; & \frac{1}{2}\left[(U_j-u_0)^2\mathbf{X}_j^T-(U_i-u_0)^2\mathbf{X}_i^T\right]\mathbf{a}''(u_0)+\frac{1}{2}\left[(U_j-u_0)^2-(U_i-u_0)^2\right]a_0''(u_0) \\ & +o((U_i-u_0)^2)+o((U_j-u_0)^2). \end{aligned}$$

By applying Lemma 5.6, it can be shown that $S_{n\mathbf{b}}(\mathbf{0},0) = E[H_n^*(D_i,D_j)] + o_p(1)$. It follows by using the same arguments as those in the proof of Lemma 5.7 that

$$\begin{aligned}
& E\big[H_n^*(D_i,D_j)\big] \\
= \;& 2nh^{-1}\gamma_n E\left\{\int\big[G(\epsilon+\Delta_j(u_0)-\Delta_i(u_0))-G(\epsilon)\big]g(\epsilon)d\epsilon\right. \\
& \left.(\mathbf{X}_j-\mathbf{X}_i)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\} \\
= \;& 2nh^{-1}\gamma_n[\tau+O(h)]E\left[\big(\Delta_j(u_0)-\Delta_i(u_0)\big)(\mathbf{X}_j-\mathbf{X}_i)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right] \\
& (1+o(1)) \\
= \;& \frac{2h^2}{\gamma_n}\left[\tau f^2(u_0)\mu_2\Sigma(u_0)\mathbf{a}''(u_0)+o(1)\right].
\end{aligned}$$

This proves (5.26). By combining (5.25) and (5.26) and using the approximation given in (5.23), we obtain (5.10). □

Proof of Theorem 5.3. A result of Hodges and Lehmann (1956) indicates that the ARE has a lower bound $0.864^{4/5} = 0.8896$, with this lower bound being attained at the density $f(t) = \frac{3}{20\sqrt{5}}(5-x^2)I(|x|\le 5)$. □

Proof of Theorem 5.4. Let

$$\begin{aligned}
& V_n(\alpha_1^*,\xi_1,\boldsymbol{\xi}_2,\boldsymbol{\xi}_3) \\
= \;& (nh)^{-1}\sum_{i=1}^n\left|\epsilon_i-\gamma_n\alpha_1^*-\xi_1(U_i-u_0)-\boldsymbol{\xi}_2^T\mathbf{X}_i-\boldsymbol{\xi}_3^T(U_i-u_0)\mathbf{X}_i+\Delta_i(u_0)\right| \\
& K\left(\frac{U_i-u_0}{h}\right),
\end{aligned}$$

where $\alpha_1^* = \gamma_n^{-1}\big(\alpha_1 - a_0(u_0)\big)$, $\xi_1 \in \mathbb{R}$, $\boldsymbol{\xi}_2 \in \mathbb{R}^p$ and $\boldsymbol{\xi}_3 \in \mathbb{R}^p$. The subgradient of $V_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3)$ with respect to α_1^* is

$$\begin{aligned} & S_n^*(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) \\ = \; & \frac{2\gamma_n}{nh}\sum_{i=1}^{n}\Big[I\big(\epsilon_i \le \gamma_n\alpha_1^* + \xi_1(U_i - u_0) + \boldsymbol{\xi}_2^T\mathbf{X}_i + \boldsymbol{\xi}_3^T(U_i - u_0)\mathbf{X}_i - \Delta_i(u_0)\big) - 1/2\Big] \\ & K\left(\frac{U_i - u_0}{h}\right). \end{aligned}$$

We have $S_n^*(0,0,\mathbf{0},\mathbf{0}) = 2\gamma_n(nh)^{-1}\sum_{i=1}^{n}[I(\epsilon_i \le \Delta_i(u_0)) - 1/2]K\left(\frac{U_i-u_0}{h}\right)$, which is the same as the S_{n0} defined in (5.17). Let $U_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = \gamma_n^{-1}[S_n^*(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) - S_n^*(0,0,\mathbf{0},\mathbf{0})]$, then

$$\begin{aligned} & U_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) \\ = \; & 2(nh)^{-1}\sum_{i=1}^{n}\Big[I\big(\epsilon_i \le \gamma_n\alpha_1^* + \xi_1(U_i - u_0) + \boldsymbol{\xi}_2^T\mathbf{X}_i + \boldsymbol{\xi}_3^T(U_i - u_0)\mathbf{X}_i - \Delta_i(u_0)\big) \\ & - I\big(\epsilon_i \le \Delta_i(u_0)\big)\Big]K\left(\frac{U_i - u_0}{h}\right). \end{aligned}$$

For any positive constants c_i, $i = 1,2,3$ and $\forall\ \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3$ such that $\xi_1 \le c_1 h^{-1}\gamma_n$, $||\boldsymbol{\xi}_2|| \le c_2\gamma_n$ and $||\boldsymbol{\xi}_3|| \le c_3 h^{-1}\gamma_n$, we have

$$U_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = 2\gamma_n g(0) f(u_0)\alpha_1^* + o_p(1). \tag{5.27}$$

This can be proved by directly checking the mean and variance. More specifically,

$$\begin{aligned}
& E\big[U_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3)\big] \\
= \;& 2h^{-1}E\Bigg\{\Big[G\big(\gamma_n\alpha_1^* + \xi_1(U_i - u_0) + \boldsymbol{\xi}_2^T\mathbf{X}_i + \boldsymbol{\xi}_3^T(U_i - u_0)\mathbf{X}_i - \Delta_i(u_0)\big) \\
& -G\big(-\Delta_i(u_0)\big)\Big]K\left(\frac{U_i - u_0}{h}\right)\Bigg\} \\
= \;& 2h^{-1}g(0)E\left\{\Big[\gamma_n\alpha_1^* + \xi_1(U_i - u_0) + \boldsymbol{\xi}_2^T\mathbf{X}_i + \boldsymbol{\xi}_3^T(U_i - u_0)\mathbf{X}_i\Big]K\left(\frac{U_i - u_0}{h}\right)\right\} \\
& (1 + O(h)) \\
= \;& 2\gamma_n g(0)f(u_0)\alpha_1^*\big(1 + O(h)\big).
\end{aligned}$$

And

$$\begin{aligned}
& Var\big[U_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3)\big] \\
\leq \;& 4n^{-1}h^{-2}E\Bigg\{\Big[I\big(\epsilon_i \leq \gamma_n\alpha_1^* + \xi_1(U_i - u_0) + \boldsymbol{\xi}_2^T\mathbf{X}_i + \boldsymbol{\xi}_3^T(U_i - u_0)\mathbf{X}_i - \Delta_i(u_0)\big) \\
& -I\big(\epsilon_i \leq \Delta_i(u_0)\big)\Big]^2K^2\left(\frac{U_i - u_0}{h}\right)\Bigg\} \\
\leq \;& 4n^{-1}h^{-2}E\left\{K^2\left(\frac{U_i - u_0}{h}\right)\right\} = O(n^{-1}h^{-1}) = o(1).
\end{aligned}$$

By (5.27) and similar proof as that for Lemma 5.1, we have

$$\gamma_n^{-1}V_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) = V_n^*(\alpha_1^*) + o_p(1), \tag{5.28}$$

where $V_n^*(\alpha_1^*) = \gamma_n^{-1}S_n^*(0, 0, \mathbf{0}, \mathbf{0})\alpha_1^* + \gamma_n g(0)f(u_0)\alpha_1^{*2} + \gamma_n^{-1}V_n(0, 0, \mathbf{0}, \mathbf{0})$. Because the function $V_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3)$ is convex in its arguments, (5.28) can be strengthened to uniform convergence (convexity lemma, see Pollard 1991) , i.e.,

$$\sup_{\substack{\alpha_1^* \in \mathbb{C},\ ||\xi_1|| \leq c_1h^{-1}\gamma_n \\ ||\boldsymbol{\xi}_2|| \leq c_2\gamma_n,\ ||\boldsymbol{\xi}_3|| \leq c_3h^{-1}\gamma_n}} |\gamma_n^{-1}V_n(\alpha_1^*, \xi_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3) - V_n^*(\alpha_1^*)| = o_p(1),$$

where $\mathbb{C}$ is a compact set in $\mathbb{R}$. By Theorem 5.2, $\widehat{\alpha}_2 - a_0'(u_0) = O_p(h^{-1}\gamma_n)$, $\widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0) = O_p(\gamma_n)$ and $\widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0) = O_p(h^{-1}\gamma_n)$, we thus have

$$\sup_{\alpha_1^* \in \mathbb{C}} \left| \gamma_n^{-1} V_n\big(\alpha_1^*, \widehat{\alpha}_2 - a_0'(u_0), \widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0), \widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0)\big) - V_n^*(\alpha_1^*) \right| = o_p(1).$$

Note that $V_n(\alpha_1^*, \widehat{\alpha}_2 - a_0'(u_0), \widehat{\mathbf{a}}(u_0) - \mathbf{a}(u_0), \widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0)) = Q_{n0}^*(\alpha_1^*, \widehat{\alpha}_2, \widehat{\boldsymbol{\beta}})$, $S_n^*(0, 0, \mathbf{0}, \mathbf{0}) = S_{n0}$, where Q_{n0}^* and S_{n0} are defined in Section 2.4. The quadratic function $V_n^*(\alpha_1^*)$ is minimized by $\widetilde{\alpha}_{1n}^* = \frac{1}{2}\gamma_n^{-2}[g(0)f(u_0)]^{-1}S_{n0}$. As similar argument to that for Theorem 5.2 shows that $\widehat{\alpha}_{1n}^* = \widetilde{\alpha}_{1n}^* + o_p(1)$. Thus we have (5.18). We can write $\gamma_n^{-2}S_{n0} = T_{1n} + T_{2n}$, where

$$\begin{aligned} T_{1n} &= \frac{2\gamma_n^{-1}}{nh} \sum_{i=1}^{n} \big[I(\epsilon_i \le 0) - 1/2\big] K\left(\frac{U_i - u_0}{h}\right), \\ T_{2n} &= \frac{2\gamma_n^{-1}}{nh} \sum_{i=1}^{n} \big[I(\epsilon_i \le -\Delta_i(u_0)) - I(\epsilon_i \le 0)\big] K\left(\frac{U_i - u_0}{h}\right). \end{aligned}$$

By the Lindeberg-Feller central limit theorem, $T_{1n} \to N(0, f(u_0)\nu_0/3)$ in distribution. By checking mean and variance, we have

$$T_{2n} = -\frac{h^2}{\gamma_n} g(0) f(u_0) a_0''(u_0) \mu_2 (1 + o(1)) + o_p(1).$$

Combining the above results and using (5.18), the proof is completed. □

To prove Theorem 5.5, we first extend Lemma 5.6 to almost sure convergence.

Lemma 5.9. *If $E[||H_n(D_i, D_j)||^2] = O(h^{-2})$, then $U_n - \widehat{U}_n = o(1)$ almost surely and $U_n = \overline{r}_n + o(1)$ a.s.*

Proof. The proof of Powell, Stock, and Stoker (1989) for Lemma 5.6 suggests that $E[||U_n - \widehat{U}_n||^2] = O(n^{-2}h^{-2})$. By Theorem 1.3.5 of Serfling (1980), $\sum_{i=1}^{n} E[||U_n - \widehat{U}_n||^2] = O(n^{-1}h^{-2}) < \infty$. This implies that $U_n - \widehat{U}_n = o(1)$ almost surely. The

second result follows by an application of the strong law of large numbers to $\widehat{U}_n$. □

Proof of Theorem 5.5. Let $\boldsymbol{\beta}^*$ and α_2^* be defined the same as before. We introduce the reparametrized objective function $\overline{Q}_n^*(\boldsymbol{\beta}^*, \alpha_2^*)$. Let $\overline{S}_n(\boldsymbol{\beta}^*, \alpha_2^*) = \big(\overline{S}_{n1}^T(\boldsymbol{\beta}^*, \alpha_2^*), \overline{S}_{n2}(\boldsymbol{\beta}^*, \alpha_2^*)\big)^T$ denote the gradient function of $\overline{Q}_n^*(\boldsymbol{\beta}^*, \alpha_2^*)$, which is defined similarly as in Section 2.2. We first show that $\overline{S}_n(\boldsymbol{\beta}^*, \alpha_2^*)$ has a similar local linear approximation to the one stated in Lemma 5.8. To make the proof concise, we prove this for $\overline{S}_{n1}(\boldsymbol{\beta}^*, \alpha_2^*)$, where

$$\begin{aligned} & \overline{S}_{n1}(\boldsymbol{\beta}^*, \alpha_2^*) \\ = \; & 2\gamma_n[n(n-1)]^{-1}\sum_{i\neq j}\big(V_i+V_j\big)\big[I\big(\epsilon_i - \gamma_n\alpha_2^*(U_i-u_0)/h - \gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0) \le \epsilon_j \\ & -\gamma_n\alpha_2^*(U_j-u_0)/h - \gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\big) - 1/2\big](\mathbf{Z}_i-\mathbf{Z}_j)K_h(U_i-u_0)K_h(U_j-u_0). \end{aligned}$$

Let $U_n = \gamma_n^{-1}[\overline{S}_{n1}(\boldsymbol{\beta}^*, \alpha_2^*) - \overline{S}_{n1}(\mathbf{0}, 0)] = [n(n-1)]^{-1}\sum_{i\neq j}\big(V_i+V_j\big)M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*)$, where $M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*) = \frac{1}{2}\big[m_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*) + m_n(D_j, D_i, \boldsymbol{\beta}^*, \alpha_2^*)\big]$ and

$$\begin{aligned} & m_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*) \\ = \; & 2\big[I\big(\epsilon_i - \gamma_n\alpha_2^*(U_i-u_0)/h - \gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_i + \Delta_i(u_0) \le \epsilon_j - \gamma_n\alpha_2^*(U_j-u_0)/h \\ & -\gamma_n\boldsymbol{\beta}^{*T}\mathbf{Z}_j + \Delta_j(u_0)\big) - 1/2\big](\mathbf{Z}_i-\mathbf{Z}_j)K_h(U_i-u_0)K_h(U_j-u_0). \end{aligned}$$

Note that $U_n = 2n^{-1}\sum_{i=1}^n V_i\big[(n-1)^{-1}\sum_{j=1,j\neq i}^n M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*)\big]$. Conditional on $\{D_i\}_{i=1}^n$, this is a weighted average of V_i. Note that

$$\begin{aligned} E(U_n|\{D_i\}_{i=1}^n) &= [n(n-1)]^{-1}\sum_{i\neq j} M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*), \\ Var(U_n|\{D_i\}_{i=1}^n) &= n^{-2}\sum_{i=1}^n\left[(n-1)^{-1}\sum_{j=1,j\neq i}^n M_n(d_i, d_j, \boldsymbol{\beta}^*, \alpha_2^*)\right]^2. \end{aligned}$$

By Lemma 5.9, it can be shown that $[n(n-1)]^{-1}\sum_{i\neq j} M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*) = \gamma_n A^* \boldsymbol{\beta}^* + o(1)$ almost surely, where $\mathbf{A}^* = 4\tau f^2(u_0)\mathrm{diag}(\mathbf{I}_p, \mu_2 \mathbf{I}_p) \otimes \Sigma(\mu_0)$. It is also easy to check that $n^{-2}\sum_{i=1}^n \big[(n-1)^{-1}\sum_{j=1, j\neq i}^n M_n(D_i, D_j, \boldsymbol{\beta}^*, \alpha_2^*)\big]^2 = o(1)$ almost surely. Thus for almost surely every sequence $\{D_i\}_{i=1}^n$, $U_n = \gamma_n A^* \boldsymbol{\beta}^* + o_p(1)$, where $o_p(1)$ is in the probability space generated by $\{V_i\}_{i=1}^n$. The proofs of Lemma 5.1 and the asymptotic representation in Theorem 5.2 can be similarly carried out to show that for almost surely every sequence $\{D_i\}_{i=1}^n$,

$$\sqrt{nh}\big[\overline{\mathbf{a}}_n(u_0) - \mathbf{a}(u_0)\big] = -\gamma_n^{-2}\big(4\tau f^2(u_0)\Sigma(\mu_0)\big)^{-1}\overline{S}^*_{n\mathbf{a}}(\mathbf{0}, 0) + o_p(1), \qquad (5.29)$$

where $o_p(1)$ is in the probability space generated by $\{V_i\}_{i=1}^n$, and

$$\begin{aligned}\overline{S}^*_{n\mathbf{a}}(\mathbf{0}, 0) &= 2\gamma_n[n(n-1)]^{-1}\sum_{i\neq j}(V_i + V_j)\left[I\left(\epsilon_i + \Delta_i(u_0) \leq \epsilon_j + \Delta_j(u_0)\right) - 1/2\right] \\ &\quad (\mathbf{X}_i - \mathbf{X}_j)K_h(U_i - u_0)K_h(U_j - u_0).\end{aligned}$$

The approximation (5.23) can be strengthened to almost sure convergence, i.e.,

$$\sqrt{nh}\big[\widehat{\mathbf{a}}_n(u_0) - \mathbf{a}(u_0)\big] = -\gamma_n^{-2}\big(4\tau f^2(u_0)\Sigma(\mu_0)\big)^{-1}S^*_{n\mathbf{a}}(\mathbf{0}, 0) + o(1) \quad \text{a. s..} \quad (5.30)$$

Combining (5.17) and (5.29), we have that for almost surely every sequence $\{D_i\}_{i=1}^n$,

$$\sqrt{nh}\big[\overline{\mathbf{a}}_n(u_0) - \widehat{\mathbf{a}}_n(u_0)\big] = -\gamma_n^{-2}\big(4\tau f^2(u_0)\Sigma(\mu_0)\big)^{-1}\big[\overline{S}^*_{n\mathbf{a}}(\mathbf{0}, 0) - S^*_{n\mathbf{a}}(\mathbf{0}, 0)\big] + o_p(1).$$

Note that

$$
\begin{aligned}
& \gamma_n^{-2}\big[\overline{S}^*_{n\mathbf{a}}(\mathbf{0},0) - S^*_{n\mathbf{a}}(\mathbf{0},0)\big] \\
= \;& 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\big[(V_i - 1/2) + (V_j - 1/2)\big] \\
& \Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\Big) - 1/2\Big](\mathbf{X}_i - \mathbf{X}_j)K_h(U_i - u_0)K_h(U_j - u_0) \\
= \;& 4\gamma_n^{-1}n^{-1}\sum_{i=1}^{n}(V_i - 1/2)\Big\{(n-1)^{-1}\sum_{j=1,j\neq i}^{n}\Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\Big) - 1/2\Big] \\
& (\mathbf{X}_i - \mathbf{X}_j)K_h(U_i - u_0)K_h(U_j - u_0)\Big\}.
\end{aligned}
$$

And $vE\left\{\gamma_n^{-2}\big[\overline{S}^*_{n\mathbf{a}}(\mathbf{0},0) - S^*_{n\mathbf{a}}(\mathbf{0},0)\big]\big|\{D_i\}_{i=1}^n\right\} = \mathbf{0}$. We have

$$
\begin{aligned}
& Var\left\{\gamma_n^{-2}\big[\overline{S}^*_{n\mathbf{a}}(\mathbf{0},0) - S^*_{n\mathbf{a}}(\mathbf{0},0)\big]\big|\{D_i\}_{i=1}^n\right\} \\
= \;& 16\gamma_n^{-2}n^{-2}(n-1)^2\sum_{i=1}^{n}\Big\{\sum_{j=1,j\neq i}^{n}\Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\Big) - 1/2\Big] \\
& (\mathbf{X}_i - \mathbf{X}_j)K_h(U_i - u_0)K_h(U_j - u_0)\Big\}^2 \\
= \;& W_1 + W_2,
\end{aligned}
$$

where

$$
\begin{aligned}
W_1 = \;& 16\gamma_n^{-2}n^{-2}(n-1)^2h^{-4}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\Big) - 1/2\Big]^2 \\
& (\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^T K^2\big((U_i - u_0)/h\big)K^2\big((U_j - u_0)/h\big), \\
W_2 = \;& 16\gamma_n^{-2}n^{-2}(n-1)^2h^{-4}\sum_{i=1}^{n}\sum_{j_1\neq i}^{n}\sum_{j_2\neq i,j_1}^{n}\Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_{j_1} + \Delta_{j_1}(u_0)\Big) - 1/2\Big] \\
& \Big[I\Big(\epsilon_i + \Delta_i(u_0) \le \epsilon_{j_2} + \Delta_{j_2}(u_0)\Big) - 1/2\Big](\mathbf{X}_i - \mathbf{X}_{j_1})(\mathbf{X}_i - \mathbf{X}_{j_2})^T \\
& K^2\big((U_i - u_0)/h\big)K\big((U_{j_1} - u_0)/h\big)K\big((U_{j_2} - u_0)/h\big).
\end{aligned}
$$

Lemma 5.9 can be used to show that $W_1 = o(1)$ almost surely; and a minor extension of Lemma 5.9 to a third-order U-statistic can be used to show that $W_2 = \frac{4}{3}f^3(u_0)\nu_0\Sigma(u_0) + o(1)$ almost surely. The asymptotic normality of $\gamma_n^{-2}\left[\overline{S}^*_{n\mathbf{a}}(\mathbf{0},0) - S^*_{n\mathbf{a}}(\mathbf{0},0)\right]$ follows by showing that the condition of the Lindeberg-Feller central limit theorem for triangular arrays holds almost surely. We have, for almost surely every sequence $\{D_i\}_{i=1}^n$,

$$\gamma_n^{-2}\left[\overline{S}^*_{n\mathbf{a}}(\mathbf{0},0) - S^*_{n\mathbf{a}}(\mathbf{0},0)\right] \to N\left(\mathbf{0}, \frac{4}{3}f^3(u_0)\nu_0\Sigma(u_0)\right)$$

in distribution. This completes the proof. □

5.5.2 Asymptotic normality of $\widehat{\alpha}_2$ and $\widehat{\mathbf{a}}'$

Asymptotic normality of $\widehat{\alpha}_2$.
From (5.11), we have $\sqrt{nh^3}\left[\widehat{\alpha}_2 - a_0'(u_0)\right] = -\gamma_n^{-2}[4\tau f^2(u_0)\mu_2]^{-1}S_{n2}(\mathbf{0},0) + o_P(1)$, where

$$\begin{aligned} S_{n2}(\mathbf{0},0) &= 2\gamma_n[n(n-1)]^{-1}\sum_{i\neq j}\left[I\left(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\right) - 1/2\right]\left(\frac{U_i - U_j}{h}\right) \\ &\quad K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned} \tag{5.31}$$

By (5.31), let us rewrite $-\gamma_n^{-2}S_{n2}(\mathbf{0},0) = C_1 + C_2$, where

$$\begin{aligned} C_1 &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\left[I\left(\epsilon_i \le \epsilon_j\right) - 1/2\right]\left(\frac{U_j - U_i}{h}\right)K_h(U_i - u_0)K_h(U_j - u_0), \\ C_2 &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\left[I\big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\big) - I\big(\epsilon_i \le \epsilon_j\big)\right]\left(\frac{U_j - U_i}{h}\right) \\ &\quad K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned}$$

Similar to the proof of Theorem 3.2, we can show that $C_1 \to N\left(0, \frac{4}{3}f^3(u_0)\nu_2\right)$ in distribution. Next, note that we can write $C_2 = [n(n-1)]^{-1}\sum_{i\neq j}H_n(D_i, D_j)$,

where

$$H_n(D_i, D_j) = 2\gamma_n^{-1}h^{-2}\left[I\left(\epsilon_i + \Delta_i(u_0) \leq \epsilon_j + \Delta_j(u_0)\right) - I\left(\epsilon_i \leq \epsilon_j\right)\right]\left(\frac{U_j - U_i}{h}\right) K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right).$$

As before, we can show that $C_2 = E[H_n(D_i, D_j)] + o_p(1)$. Now

$$\begin{aligned} & E\big[H_n(D_i, D_j)\big] \\ = \; & 2\gamma_n^{-1}h^{-2}E\left\{\int\big[G(\epsilon + \Delta_j(u_0) - \Delta_i(u_0)) - G(\epsilon)\big]g(\epsilon)d\epsilon \right. \\ & \left.\left(\frac{U_j - U_i}{h}\right) K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\} \\ = \; & 2\gamma_n^{-1}h^{-2}[\tau + O(h)]E\left[\left(\Delta_j(u_0) - \Delta_i(u_0)\right)\left(\frac{U_j - U_i}{h}\right) K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right] \\ & (1 + o(1)). \end{aligned}$$

Note that

$$\begin{aligned} & \Delta_j(u_0) - \Delta_i(u_0) \\ = \; & \frac{1}{2}\left[(U_j - u_0)^2\mathbf{X}_j^T - (U_i - u_0)^2\mathbf{X}_i^T\right]\mathbf{a}''(u_0) + \frac{1}{6}\left[(U_j - u_0)^3\mathbf{X}_j^T - (U_i - u_0)^3\mathbf{X}_i^T\right]\mathbf{a}'''(u_0) \\ & + \frac{1}{2}\left[(U_j - u_0)^2 - (U_i - u_0)^2\right]a_0''(u_0) + \frac{1}{6}\left[(U_j - u_0)^3 - (U_i - u_0)^3\right]a_0'''(u_0) \\ & + o((U_i - u_0)^3) + o((U_j - u_0)^3). \end{aligned}$$

Thus, we have

$$
\begin{aligned}
& E\big[H_n(D_i, D_j)\big] \\
= \; & \gamma_n^{-1} h^{-2}[\tau + O(h)] E\Bigg\{ \Big[(U_i - u_0)^2 \mathbf{X}_i^T - (U_j - u_0)^2 \mathbf{X}_j^T\Big] \left(\frac{U_i - U_j}{h}\right) K\left(\frac{U_i - u_0}{h}\right) \\
& K\left(\frac{U_j - u_0}{h}\right)\Bigg\} \mathbf{a}''(u_0) \\
& + \frac{1}{3}\gamma_n^{-1} h^{-2}[\tau + O(h)] E\Bigg\{ \Big[(U_i - u_0)^3 \mathbf{X}_i^T - (U_j - u_0)^3 \mathbf{X}_j^T\Big] \left(\frac{U_i - U_j}{h}\right) K\left(\frac{U_i - u_0}{h}\right) \\
& K\left(\frac{U_j - u_0}{h}\right)\Bigg\} \mathbf{a}'''(u_0) \\
& + \gamma_n^{-1} h^{-2}[\tau + O(h)] E\Bigg\{ \Big[(U_i - u_0)^2 - (U_j - u_0)^2\Big] \left(\frac{U_i - U_j}{h}\right) K\left(\frac{U_i - u_0}{h}\right) \\
& K\left(\frac{U_j - u_0}{h}\right)\Bigg\} a_0''(u_0) \\
& + \frac{1}{3}\gamma_n^{-1} h^{-2}[\tau + O(h)] E\Bigg\{ \Big[(U_i - u_0)^3 - (U_j - u_0)^3\Big] \left(\frac{U_i - U_j}{h}\right) K\left(\frac{U_i - u_0}{h}\right) \\
& K\left(\frac{U_j - u_0}{h}\right)\Bigg\} a_0'''(u_0) \\
& \triangleq E_1 + E_2 + E_3 + E_4.
\end{aligned}
$$

where the definition of E_i, $i = 1, \ldots, 4$, should be clear from the context. Below, we use $m(u)$ to denote $E(X_i | U_i = u)$. Note that $m(u_0) = 0$. Assume that $m(u)$ is

continuously differentiable around $u = u_0$. Then

$$\begin{aligned}
E_1 &= 2\gamma_n^{-1}h^{-2}[\tau + O(h)]E\Bigg\{(U_i - u_0)^2\mathbf{X}_i^T\left[\frac{U_i - U_0}{h} - \frac{U_j - U_0}{h}\right]K\left(\frac{U_i - u_0}{h}\right) \\
&\quad K\left(\frac{U_j - u_0}{h}\right)\Bigg\}\mathbf{a}''(u_0) \\
&= 2\gamma_n^{-1}h^{-3}[\tau + O(h)]E\left\{(U_i - u_0)^3 m(U_i)^T K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\}\mathbf{a}''(u_0) \\
&\quad -2\gamma_n^{-1}h^{-2}[\tau + O(h)]E\left\{(U_i - u_0)^2 m(U_i)^T\left(\frac{U_j - u_0}{h}\right)K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\} \\
&\quad \mathbf{a}''(u_0) \\
&= 2\gamma_n^{-1}h^2[\tau + O(h)]\Big[\int t^3 m(u_0 + th)^T K(t) f(u_0 + th)dt\Big]\Big[\int K(t)f(u_0 + th)dt\Big]\mathbf{a}''(u_0) \\
&\quad -2\gamma_n^{-1}h^2[\tau + O(h)]\Big[\int t^2 m(u_0 + th)^T K(t) f(u_0 + th)dt\Big]\Big[\int tK(t)f(u_0 + th)dt\Big]\mathbf{a}''(u_0) \\
&= 2\gamma_n^{-1}h^2[\tau + O(h)]\big[hm'(u_0)^T f(u_0)\mu_4 + o(h)\big]\big[f(u_0) + O(h)\big]\mathbf{a}''(u_0) \\
&\quad -2\gamma_n^{-1}h^2[\tau + O(h)]\big[hm'(u_0)^T f(u_0)\mu_3 + o(h)\big]\big[hf'(u_0)\mu_2 + o(h)\big]\mathbf{a}''(u_0) \\
&= \frac{2h^3}{\gamma_n}\tau f^2(u_0)\mu_4 m'(u_0)^T\mathbf{a}''(u_0)\big(1 + o(1)\big).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
E_2 &= \frac{2}{3}\gamma_n^{-1}h^{-3}[\tau + O(h)]E\left\{(U_i - u_0)^4 m(U_i)^T K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\}\mathbf{a}'''(u_0) \\
&\quad -\frac{2}{3}\gamma_n^{-1}h^{-2}[\tau + O(h)]E\left\{(U_i - u_0)^3 m(U_i)^T\left(\frac{U_j - u_0}{h}\right)K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\} \\
&\quad \mathbf{a}'''(u_0) \\
&= \frac{2}{3}\gamma_n^{-1}h^3[\tau + O(h)]\Big[\int t^4 m(u_0 + th)^T K(t) f(u_0 + th)dt\Big]\Big[\int K(t)f(u_0 + th)dt\Big]\mathbf{a}'''(u_0) \\
&\quad -\frac{2}{3}\gamma_n^{-1}h^3[\tau + O(h)]\Big[\int t^3 m(u_0 + th)^T K(t) f(u_0 + th)dt\Big]\Big[\int tK(t)f(u_0 + th)dt\Big]\mathbf{a}'''(u_0) \\
&= \frac{2h^3}{\gamma_n}o(1).
\end{aligned}$$

Also,

$$\begin{aligned}
E_3 &= 2\gamma_n^{-1}h^{-3}[\tau+O(h)]E\left\{(U_i-u_0)^3K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0''(u_0)\\
&\quad -2\gamma_n^{-1}h^{-2}[\tau+O(h)]E\left\{(U_i-u_0)^2\left(\frac{U_j-u_0}{h}\right)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0''(u_0)\\
&= 2\gamma_n^{-1}h^2[\tau+O(h)]\Big[\int t^3K(t)f(u_0+th)dt\Big]\Big[\int K(t)f(u_0+th)dt\Big]a_0''(u_0)\\
&\quad -2\gamma_n^{-1}h^2[\tau+O(h)]\Big[\int t^2K(t)f(u_0+th)dt\Big]\Big[\int tK(t)f(u_0+th)dt\Big]a_0''(u_0)\\
&= 2\gamma_n^{-1}h^2[\tau+O(h)]\big[hf'(u_0)\mu_4+o(h)\big]\big[f(u_0)+O(h)\big]a_0''(u_0)\\
&\quad -2\gamma_n^{-1}h^2[\tau+O(h)]\big[f(u_0)\mu_2+O(h)\big]\big[hf'(u_0)\mu_2+o(h)\big]a_0''(u_0)\\
&= \frac{2h^3}{\gamma_n}\tau f(u_0)f'(u_0)(\mu_4-\mu_2^2)a_0''(u_0)\big(1+o(1)\big).
\end{aligned}$$

Finally,

$$\begin{aligned}
E_4 &= \frac{2}{3}\gamma_n^{-1}h^{-3}[\tau+O(h)]E\left\{(U_i-u_0)^4K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0'''(u_0)\\
&\quad -\frac{2}{3}\gamma_n^{-1}h^{-2}[\tau+O(h)]E\left\{(U_i-u_0)^3\left(\frac{U_j-u_0}{h}\right)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0'''(u_0)\\
&= \frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int t^4K(t)f(u_0+th)dt\Big]\Big[\int K(t)f(u_0+th)dt\Big]a_0'''(u_0)\\
&\quad -\frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int t^3K(t)f(u_0+th)dt\Big]\Big[\int tK(t)f(u_0+th)dt\Big]a_0''(u_0)\\
&= \frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\big[f(u_0)\mu_4+O(h)\big]\big[f(u_0)+O(h)\big]a_0'''(u_0)\\
&\quad -\frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\big[hf'(u_0)\mu_4+o(h)\big]\big[hf'(u_0)\mu_2+o(h)\big]a_0'''(u_0)\\
&= \frac{2h^3}{3\gamma_n}\tau f^2(u_0)\mu_4a_0'''(u_0)\big(1+o(1)\big).
\end{aligned}$$

In summary, $E[H_n(D_i,D_j)]=\frac{2h^3}{\gamma_n}\tau f(u_0)\xi(u_0)(1+o(1))$, where

$$\xi(u_0)=f(u_0)\mu_4m'(u_0)^T\mathbf{a}''(u_0)+f'(u_0)(\mu_4-\mu_2^2)a_0''(u_0)+\frac{1}{3}f(u_0)\mu_4a_0'''(u_0).$$

This leads to $\sqrt{nh^3}\left[\widehat{\alpha}_2 - a_0'(u_0) - \frac{h^4}{2f(u_0)\mu_2}\xi(u_0) + o(h^4)\right] \to N\Big(0, \left[12\tau^2 f(u_0)\mu_2^2\right]^{-1}\nu_2\Big)$ in distribution. □

Asymptotic normality of $\widehat{\mathbf{a}}'(u_0)$.

From (5.12), $\sqrt{nh^3}\left[\widehat{\mathbf{a}}'(u_0) - \mathbf{a}'(u_0)\right] = -\gamma_n^{-2}[4\tau f^2(u_0)\mu_2\Sigma(u_0)]^{-1}S_{n12}(\mathbf{0},0) + o_P(1)$, where

$$\begin{aligned} S_{n12}(\mathbf{0},0) &= 2\gamma_n[n(n-1)]^{-1}\sum_{i\neq j}\left[I\left(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\right) - 1/2\right] \\ &\quad \left[\frac{U_i - u_0}{h}X_i - \frac{U_j - u_0}{h}X_j\right] K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned} \tag{5.32}$$

By (5.32), let us rewrite $-\gamma_n^{-2}S_{n12}(\mathbf{0},0) = C_1 + C_2$, where

$$\begin{aligned} C_1 &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j} \\ &\quad \left[I\left(\epsilon_i \le \epsilon_j\right) - 1/2\right]\left[\frac{U_j - u_0}{h}X_j - \frac{U_i - u_0}{h}X_i\right] K_h(U_i - u_0)K_h(U_j - u_0), \\ C_2 &= 2\gamma_n^{-1}[n(n-1)]^{-1}\sum_{i\neq j}\left[I\big(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\big) - I\big(\epsilon_i \le \epsilon_j\big)\right] \\ &\quad \left[\frac{U_j - u_0}{h}X_j - \frac{U_i - u_0}{h}X_i\right] \\ &\quad K_h(U_i - u_0)K_h(U_j - u_0). \end{aligned}$$

Following the proof of Theorem 3.2, we can show that $C_1 \to N\left(0, \frac{4}{3}f^3(u_0)\Sigma(u_0)\nu_2\right)$ in distribution. Furthermore, we can write $C_2 = [n(n-1)]^{-1}\sum_{i\neq j}H_n(D_i, D_j)$, where

$$\begin{aligned} H_n(D_i, D_j) &= 2\gamma_n^{-1}h^{-2}\left[I\left(\epsilon_i + \Delta_i(u_0) \le \epsilon_j + \Delta_j(u_0)\right) - I\left(\epsilon_i \le \epsilon_j\right)\right] \\ &\quad \left[\frac{U_j - u_0}{h}X_j - \frac{U_i - u_0}{h}X_i\right] K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right). \end{aligned}$$

As before, we can show that $C_2 = E[H_n(D_i, D_j)] + o_p(1)$. Now

$$
\begin{aligned}
& E\big[H_n(D_i, D_j)\big] \\
= \; & 2\gamma_n^{-1} h^{-2}[\tau + O(h)] E\left[\left(\frac{U_j - u_0}{h} X_j - \frac{U_i - u_0}{h} X_i\right)\left(\Delta_j(u_0) - \Delta_i(u_0)\right)\right. \\
& \left. K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right] \\
= \; & \gamma_n^{-1} h^{-2}[\tau + O(h)] E\left\{\left(\frac{U_j - u_0}{h} X_j - \frac{U_i - u_0}{h} X_i\right)\left[(U_j - u_0)^2 \mathbf{X}_j^T - (U_i - u_0)^2 \mathbf{X}_i^T\right]\right. \\
& \left. K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\} \mathbf{a}''(u_0) \\
& + \frac{1}{3}\gamma_n^{-1} h^{-2}[\tau + O(h)] E\left\{\left(\frac{U_j - u_0}{h} X_j - \frac{U_i - u_0}{h} X_i\right)\left[(U_j - u_0)^3 \mathbf{X}_j^T - (U_i - u_0)^3 \mathbf{X}_i^T\right]\right. \\
& \left. K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\} \mathbf{a}'''(u_0) \\
& + \gamma_n^{-1} h^{-2}[\tau + O(h)] E\left\{\left(\frac{U_j - u_0}{h} X_j - \frac{U_i - u_0}{h} X_i\right)\left[(U_j - u_0)^2 - (U_i - u_0)^2\right]\right. \\
& \left. K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\} a_0''(u_0) \\
& + \frac{1}{3}\gamma_n^{-1} h^{-2}[\tau + O(h)] E\left\{\left(\frac{U_j - u_0}{h} X_j - \frac{U_i - u_0}{h} X_i\right)\left[(U_j - u_0)^3 - (U_i - u_0)^3\right]\right. \\
& \left. K\left(\frac{U_i - u_0}{h}\right) K\left(\frac{U_j - u_0}{h}\right)\right\} a_0'''(u_0) \\
& \triangleq E_1 + E_2 + E_3 + E_4.
\end{aligned}
$$

As before, let $m(u) = E(X_i|U_i = u)$ and $\Sigma(u) = E(X_i X_i^T|U_i = u)$, and assume that they are continuously differentiable around $u = u_0$. We have

$$
\begin{aligned}
& E_1 = 2\gamma_n^{-1}h^{-2}[\tau + O(h)]E\Bigg\{\frac{U_j - u_0}{h}X_j\left[(U_j - u_0)^2\mathbf{X}_j^T - (U_i - u_0)^2\mathbf{X}_i^T\right] \\
& \qquad K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\Bigg\}\mathbf{a}''(u_0) \\
= \;& 2\gamma_n^{-1}h^{-3}[\tau + O(h)]E\left\{(U_j - u_0)^3\Sigma(U_j)K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\}\mathbf{a}''(u_0) \\
& -2\gamma_n^{-1}h^{-2}[\tau + O(h)]E\left\{(U_i - u_0)^2 m(U_j)m(U_i)^T\left(\frac{U_j - u_0}{h}\right)K\left(\frac{U_i - u_0}{h}\right)K\left(\frac{U_j - u_0}{h}\right)\right\} \\
& \mathbf{a}''(u_0) \\
= \;& 2\gamma_n^{-1}h^2[\tau + O(h)]\Big[\int t^3\Sigma(u_0 + th)K(t)f(u_0 + th)dt\Big]\Big[\int K(t)f(u_0 + th)dt\Big]\mathbf{a}''(u_0) \\
& -2\gamma_n^{-1}h^2[\tau + O(h)]\Big[\int tm(u_0 + th)^T K(t)f(u_0 + th)dt\Big]\Big[\int t^2 m(u_0 + th)^T K(t)f(u_0 + th)dt\Big] \\
& \mathbf{a}''(u_0) \\
= \;& \frac{2h^3}{\gamma_n}\tau f(u_0)\big[\Sigma'(u_0)f(u_0) + \Sigma(u_0)f'(u_0)\big]\mu_4\mathbf{a}''(u_0)\big(1 + o(1)\big).
\end{aligned}
$$

$$
\begin{aligned}
E_2 &= \frac{2}{3}\gamma_n^{-1}h^{-2}[\tau+O(h)]E\Bigg\{\frac{U_j-u_0}{h}X_j\left[(U_j-u_0)^3\mathbf{X}_j^T-(U_i-u_0)^3\mathbf{X}_i^T\right]\\
&\qquad K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\Bigg\}\mathbf{a}'''(u_0)\\
&= \frac{2}{3}\gamma_n^{-1}h^{-3}[\tau+O(h)]E\left\{(U_j-u_0)^4\Sigma(U_j)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}\mathbf{a}'''(u_0)\\
&\quad -\frac{2}{3}\gamma_n^{-1}h^{-2}[\tau+O(h)]E\left\{(U_i-u_0)^3m(U_j)m(U_i)^T\left(\frac{U_j-u_0}{h}\right)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}\\
&\quad \mathbf{a}'''(u_0)\\
&= \frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int t^4\Sigma(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int K(t)f(u_0+th)dt\Big]\mathbf{a}'''(u_0)\\
&\quad -\frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int tm(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int t^3m(u_0+th)^TK(t)f(u_0+th)dt\Big]\\
&\quad \mathbf{a}'''(u_0)\\
&= \frac{2h^3}{3\gamma_n}\tau f^2(u_0)\Sigma(u_0)\mu_4\mathbf{a}'''(u_0)\big(1+o(1)\big).
\end{aligned}
$$

$$
\begin{aligned}
E_3 &= 2\gamma_n^{-1}h^{-2}[\tau+O(h)]E\Bigg\{\frac{U_j-u_0}{h}X_j\left[(U_j-u_0)^2-(U_i-u_0)^2\right]\\
&\quad K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\Bigg\}a_0''(u_0)\\
&= 2\gamma_n^{-1}h^{-3}[\tau+O(h)]E\left\{(U_j-u_0)^3m(U_j)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0''(u_0)\\
&\quad -2\gamma_n^{-1}h^{-2}[\tau+O(h)]E\left\{(U_i-u_0)^2m(U_j)\left(\frac{U_j-u_0}{h}\right)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\right\}a_0''(u_0)\\
&= 2\gamma_n^{-1}h^2[\tau+O(h)]\Big[\int t^3m(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int K(t)f(u_0+th)dt\Big]a_0''(u_0)\\
&\quad -2\gamma_n^{-1}h^2[\tau+O(h)]\Big[\int tm(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int t^2K(t)f(u_0+th)dt\Big]a_0''(u_0)\\
&= \frac{2h^3}{\gamma_n}\tau f^2(u_0)m'(u_0)\big[\mu_4-\mu_2^2\big]a_0''(u_0)\big(1+o(1)\big).
\end{aligned}
$$

$$
\begin{aligned}
E_4 &= \frac{2}{3}\gamma_n^{-1}h^{-2}[\tau+O(h)]E\Bigg\{\frac{U_j-u_0}{h}X_j\left[(U_j-u_0)^3-(U_i-u_0)^3\right] \\
&\qquad K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\Bigg\}a_0'''(u_0) \\
&= \frac{2}{3}\gamma_n^{-1}h^{-3}[\tau+O(h)]E\Bigg\{(U_j-u_0)^4m(U_j)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\Bigg\}a_0'''(u_0) \\
&\quad -\frac{2}{3}\gamma_n^{-1}h^{-2}[\tau+O(h)]E\Bigg\{(U_i-u_0)^3m(U_j)\left(\frac{U_j-u_0}{h}\right)K\left(\frac{U_i-u_0}{h}\right)K\left(\frac{U_j-u_0}{h}\right)\Bigg\} \\
&\quad a_0'''(u_0) \\
&= \frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int t^4m(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int K(t)f(u_0+th)dt\Big]a_0'''(u_0) \\
&\quad -\frac{2}{3}\gamma_n^{-1}h^3[\tau+O(h)]\Big[\int tm(u_0+th)K(t)f(u_0+th)dt\Big]\Big[\int t^3K(t)f(u_0+th)dt\Big]a_0'''(u_0) \\
&= \frac{2h^3}{\gamma_n}o(1).
\end{aligned}
$$

In summary, $E[H_n(D_i,D_j)] = \frac{2h^3}{\gamma_n}\tau f(u_0)\eta(u_0)(1+o(1))$, where

$$
\begin{aligned}
\eta(u_0) &= \big[\Sigma'(u_0)f(u_0)+\Sigma(u_0)f'(u_0)\big]\mu_4\mathbf{a}''(u_0)+\frac{1}{3}f(u_0)\Sigma(u_0)\mu_4\mathbf{a}'''(u_0) \\
&\quad +f(u_0)m'(u_0)\big[\mu_4-\mu_2^2\big]a_0''(u_0).
\end{aligned}
$$

This leads to

$$
\sqrt{nh^3}\left[\widehat{\boldsymbol{\alpha}}'(u_0)-\mathbf{a}'(u_0)-\frac{h^4}{2f(u_0)\mu_2}\eta(u_0)\Sigma^{-1}(u_0)+o(h^4)\right]\to N\Big(0,\big[12\tau^2f(u_0)\mu_2^2\big]^{-1}\nu_2\Sigma^{-1}(u_0)\Big)
$$

in distribution. □

Chapter 6

Future Research

In this dissertation, we develop highly efficient, robust and computationally simple statistical methodology and inference procedures for pure nonparametric models, semiparametric partially linear models and varying coefficient models. These new procedures have competitive performance to least squares based methods when the errors come from Normal distribution. And when the errors depart from normality or there are outliers in the data, the new procedures may have much higher efficiencies than least squares based methods. Now we discuss some possible directions for future work.

Hypothesis testing for nonparametric components

After obtaining nonparametric estimates of $\alpha_0(\cdot)$ and $\boldsymbol{\alpha}(\cdot)$ in the model

$$Y = \alpha_0(U) + \mathbf{X}^T\boldsymbol{\alpha}(U) + \mathbf{Z}^T\boldsymbol{\beta} + \epsilon, \tag{6.1}$$

it is natural to ask whether the coefficient functions are actually varying, or whether any covariate is significant, or whether the coefficient functions possess certain parametric forms. So it is of interest to develop hypothesis testing procedures on α. In general, the hypothesis for a specific coefficient function $\alpha_m(\cdot)$ can be written in the form

$$H_0 : \alpha_m(u) \equiv \alpha_m(u;\boldsymbol{\theta}) \qquad \text{vs.} \qquad H_1 : \alpha_m(u) \neq \alpha_m(u;\boldsymbol{\theta}),$$

where $\alpha_m(u;\boldsymbol{\theta})$ is a parametric function of u and $\boldsymbol{\theta}$ is a parameter vector. If one would like to know whether the coefficient function is actually varying, we can test the hypothesis

$$H_0 : \alpha_m(u) \equiv a_m \qquad \text{vs.} \qquad H_1 : \alpha_m(u) \neq a_m,$$

where a_m is a constant. Furthermore, if we want to know if the corresponding covariate is significant of not, we can test

$$H_0 : \alpha_m(u) \equiv 0 \qquad \text{vs.} \qquad H_1 : \alpha_m(u) \neq 0.$$

Because we have developed robust estimators, we are also interested in developing robust test statistics. A complete review of robust hypothesis testing can be found in He (2002). The conventional maximum likelihood ratio test can not be applied in model 6.1, because the nonparametric MLE does not exist for the coefficient functions $\boldsymbol{\alpha}(\cdot)$. We may consider constructing the generalized likelihood ratio test statistic

$$T_{\hat{\boldsymbol{\beta}}} = \ell_n(H_1) - \ell_n(H_0).$$

For tests of $\boldsymbol{\beta}$, we may also consider the Wald-type statistic

$$W_{\hat{\boldsymbol{\beta}}} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

where $\hat{\boldsymbol{\Sigma}}$ is an estimate of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$.

Boundary effects for local rank regression

The design density always has a bounded support in applications. It is well known that the performance of regression smoothers at boundary points usually differs from the performance at interior points. For example, the Watson-Nadaraya

and Gasser-Muller estimators have boundary effects-bias of order $O(h)$ instead of $O(h^2)$ and require boundary modifications (Müller 1988). In Chapter 3, we show that the local CQR estimator does not suffer from the boundary effects and does not require such a modification. In Chapter 5, we present the asymptotic results for local rank estimator at any interior points u_0. The results are very encouraging. So it is worthwhile to investigate the boundary behavior of the local rank estimator as a topic for future research.

Robust semiparametric models for longitudinal data

All the robust models we have built so far are based on the independent error condition. Dependent data, such as longitudinal data or more general functional data, are widely collected in all kinds of scientific studies. Because of the complex covariance structure and the unbalanced nature of longitudinal data, the extensions are challenging and need to be seriously considered.

Longitudinal data emerge dramatically in applications of biology, economics, epidemiology, clinical trials and many other fields. The advantage of a longitudinal study over a cross-sectional study is that the longitudinal study can separate the cohort (subject) and age (time) effects. Longitudinal data are collected from experiments with m subjects and n_i observations in subject i over time. The full data set has the structure

$$\{(\mathbf{x}_{ij}, y_{ij}, t_{ij}), i = 1, \ldots, m, j = 1, \ldots n_i\},$$

where t_{ij} denotes the j^{th} observation time point for the i^{th} subject. When each subject is scheduled to be measured at the same set of times, then resulting data is referred to as a balanced data set. When subjects are each observed at different sets of times or there are skipped observation times, then resulting data is referred

to as an unbalanced data set. Longitudinal data may be highly unbalanced because the data may be collected at irregular and subject-specific time points.

Because the observations from the same subject are correlated with each other, the analysis of longitudinal data should take into account the within subject correlation. Repeated measures analysis of variance can be used to analyze longitudinal or repeated measures data for balanced study design. However, when the data are unbalanced, it is different to apply traditional multivariate regression techniques and some alternative techniques should be used to handle unbalanced data.

Parametric regression models for analyzing longitudinal data have been developed by Laird and Ware (1982) and Liang and Zeger (1986) among others. Diggle et al. (2002) gave a thorough summary of these methods. Although parametric models are very useful, they are used at the risk of introducing modeling bias because they require parametric specification for the baseline mean function of the response variable over time. To free the linear restriction between the outcome variable and the covariates, nonparametric and varying coefficient models were considered by Wild and Yee (1996), Hoover et al. (1998), Lin and Carroll (2000), Fan and Zhang (2000) and Huang et al. (2002) among others. Lin and Carroll (2000) showed that when the standard kernel methods are used, the GEE estimators are typically the most efficient estimator of the nonparametric function even if completely ignoring the within-subject correlation.

In many instances, semiparametric models are more desirable than modeling all the covariates nonparametrically. A semiparametric model for longitudinal data has the form

$$y(t) = \mathbf{x}^T(t)\boldsymbol{\beta} + \eta(t) + \epsilon(t), \tag{6.2}$$

where $y(t)$ is the response variable and $\mathbf{x}(t)$ is the $d \times 1$ covariate vector at time t, $\boldsymbol{\beta}$ are unknown parameters, $\eta(\cdot)$ is an unknown smooth baseline function and $\epsilon(t)$

is a mean 0 stochastic process. This model is especially useful when the effects of treatment $\mathbf{x}$ are of our major interest and the effect of t is a nuisance.

The model (6.2) has been considered by several authors. Zeger and Diggle (1994) proposed an iterative algorithm to estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$ by the backfitting method. They estimated $\eta(\cdot)$ using a kernel method by ignoring the within-subject correlation and estimated $\boldsymbol{\beta}$ using weighted least squares by accounting for the within-subject correlation. Zhang et al. (1998) extended Zeger and Diggle (1994)'s model to a more general class of models termed semiparametric stochastic mixed models. A marginal approach was given by Lin and Ying (2001) in which they estimated $\boldsymbol{\beta}$ under the formation of point processes. Fan and Li (2004) further proposed two new approaches, the difference-based method and the profile least squares method, for estimating the regression coefficients.

Generalized partially linear models (GPLM) for longitudinal data can be formulated as

$$g(\mu(t)) = \mathbf{x}^T(t)\boldsymbol{\beta} + \eta(t), \tag{6.3}$$

where $g(\cdot)$ is known as a link function and $\mu(t) = E[y(t)|\mathbf{x}(t)]$ is the mean of the response variable.

For model (6.3), Severini and Staniswalis (1994) suggest estimating the nonparametric part $\eta(\cdot)$ for fixed $\boldsymbol{\beta}$ using a certain nonparametric method, such as kernel regression with standard bandwidth, and then estimating $\boldsymbol{\beta}$ using the profile method. Lin and Carroll (2001a) generalized the profile-kernel method of Severini and Staniswalis (1994) and proposed a local linear version of the profile generalized estimating equation method for clustered data with a cluster-level nonparametric covariate. Lin and Carroll (2001b) claim that the conventional profile-kernel method fails to yield a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$ if the nonparametric covariate is in observation-level, unless working independence is assumed or $\eta(t)$ is artificially undersmoothed. This result was unexpected.

It is well known that statistical estimation and inference based on least squares are highly sensitive to outliers in the data. Various robust procedures are proposed to make up this deficiency. Welsh and Richardson (1997) reviewed a number of alternatives to robust estimation of mixed models. Richardson (1997) gave some further studies. He et al. (2002) applied M-estimators in semiparametric models to longitudinal data. The authors used regression spline to approximate the nonparametric part and showed that any M-estimation algorithm for the usual linear model can get consistent estimators without specification of the error distribution and covariance structure. In generalized models, Preisser and Qaqish (1999) generalized the GEE procedure to yield parameter estimates and fitted values that are resistant to outliers, introducing the so-called resistant generalized estimating equations (REGEE). The authors used the Mallows type weight or Schweppe type weight in the estimating equations to downweight influential observations or clusters. He et al. (2005) proposed another robust GEE method for GPLM, in which they approximated the nonparametric function by a regression spline and used bounded scores and leverage-based weights in the estimating equations to achieve robustness against outliers. They showed that the regression spline approach avoids the difficulties associated with the profile-kernel method and results in the optimal rate of convergence for estimating both $\boldsymbol{\beta}$ and $\eta(\cdot)$.

All aforementioned works mainly focus on estimation of the baseline function and the regression coefficients. Only a few of them discussed the issues related to model selection. Fan and Li (2004) extended the nonconcave penalized likelihood approaches into semiparametric models for longitudinal data analysis by introducing a new quadratic loss between the observed data and the theoretical model that involves only the unknown parameter $\boldsymbol{\beta}$. The simultaneous selecting variables and estimating coefficients make it feasible to construct confidence intervals for the estimators. Li and Liang (2008) proposed a class of procedures for variable selection in semiparametric models that involve both model selection for nonparametric

components and selection of significant variables for the parametric portion. The authors proposed to select significant variables for the parametric portion using nonconcave penalized quasi-likelihood and established the rate of convergence of the resulting estimate. To select significant variables in the nonparametric component, a semiparametric generalized likelihood ratio test was proposed.

In varying coefficient semiparametric models for longitudinal data

$$y(t) = \alpha_0(t) + \mathbf{x}^T(t)\boldsymbol{\alpha}(t) + \mathbf{z}^T(t)\boldsymbol{\beta} + \epsilon(t), \tag{6.4}$$

we are interested in exploring new robust estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}(t)$ using local smoothing techniques, and conducting variable selection for $\boldsymbol{\beta}$ and also model selection for $\boldsymbol{\alpha}(t)$. Furthermore, we also would like to consider how to make proper statistical inference for $\alpha_j(t)$.

Bibliography

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *Second International Symposium on Information Theory*, pp. 267–281.

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Allen, D. M. (1974), "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction," *Technometrics*, 16, 125–127.

Antoniadis, A. (1997), "Wavelets in Statistics: A Review (with discussion)," *Journal of the Italian Statistical Association*, 6, 97–144.

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *Annals of Statistics*, 24, 2350–2383.

Brumback, B. and Rice, J. (1998), "Smoothing spline models for the analysis of nested and crossed samples of curves," *Journal of the American Statistical Association*, pp. 961–976.

Cai, Z., Fan, J., and Li, R. (2000), "Efficient estimation and inferences for varying-coefficient models," *Journal of the American Statistical Association*, 95, 888–902.

Carroll, R., Fan, J., Gijbels, I., and Wand, M. (1997), "Generalized partially linear single-index models," *Journal of the American Statistical Association*, pp. 477–489.

Chu, C. and Marron, J. (1991), "Choosing a Kernel Regression Estimator," *Statistical Science*, 6, 404–419.

Cleveland, W., Grosse, E., and Shyu, W. (1992), "Local regression models," in *Statistical Models in S*, eds. J. M. Chambers and T. J. Hastie, pp. 309–376, Wadsworth, Pacific Grove.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Craven, P. and Wahba, G. (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numer. Math*, 31, 377–403.

David, H. (1998), "Early sample measures of variability," *Statistical Science*, pp. 368–377.

Diggle, P. J., Heagerty, P. J., Liang, K.-Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data, Second Edition*, Oxford University Press, New York, second edn.

Donoho, D. L. (2000), "High-Dimensional Data Analysis: The Curse and Blessing of Dimensionality," in *Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Centrury"*, Available at http://www-stat.stanford.edu/donoho/Lectures.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Fan, J. (1992), "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *Annals of Statistics*, 21, 196–216.

Fan, J. (1997), "Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis," *Journal of the Italian Statistical Association*, 6, 131–38.

Fan, J. and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, New York.

Fan, J. and Huang, T. (2005), "Profile likelihood inferences on semiparametric varying-coefficient partially linear models," *Bernoulli*, 11, 1031–1057.

Fan, J. and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Li, R. (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723.

Fan, J. and Li, R. (2006a), "An Overview on Nonparametric and Semiparametric Techniques for Longitudinal Data," in *Frontiers in Statistics*, eds. J. Fan and H. Koul, pp. 277–303, Imperial College Press, London.

Fan, J. and Li, R. (2006b), "Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians*, eds. M. Sanz-Sole, J. Soria, J. Varona, and J. Verdera, vol. III, pp. 595–622.

Fan, J. and Zhang, J.-T. (2000), "Two-step Estimation of Functional Linear Models with Applications to Longitudinal Data," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 62, 303–322.

Fan, J. and Zhang, W. (1999), "Statistical estimation in varying coefficient models," *Annals of Statistics*, pp. 1491–1518.

Fan, J. and Zhang, W. (2008), "Statistical methods with varying coefficient models," *Statistics and its interface*, 1, 179.

Fan, J., Hu, T., and Truong, Y. (1994), "Robust non-parametric estimation," *Scandinavian Journal of Statistics*, 21, 433–446.

Foster, D. P. and George, E. I. (1994), "The Risk Inflation Criterion for Multiple Regression," *The Annals of Statistics*, 22, 1947–1975.

Frank, I. E. and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135.

Fu, W. J. (1998), "Penalized regressions: The bridge versus the lasso." *Jounral of Computational and Graphical Statistics*, 7, 397–416.

Gasser, T. and Müller, H.-G. (1984), "Estimating Regression Functions and Their Derivatives by the Kernel Method," *Scandinavian Journal of Statistics*, 11, 171–185.

Geyer, C. (1994), "On the asymptotics of constrained M-estimation," *Ann. Statist*, 22, 1993–2010.

Green, P. and Silverman, B. (1994), *Nonparametric regression and generalized linear models*, Chapman & Hall New York.

Hall, P. and Kang, K. (2001), "Bootstrapping nonparametric density estimators with empirically chosen bandwidths," *Annals of Statistics*, pp. 1443–1468.

Hampel, F. (1968), "Contributions to the Theory of Robust Estimation," Ph.D. thesis, University of California, Berkeley.

Hampel, F. (1975), "Beyond location parameters: robust concepts and methods (with discussion)," in *Proceedings of the 40th Session of the ISI.*

Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hannan, E. J. and Quinn, B. G. (1979), "The Determination of the Order of an Autoregression," *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 190–195.

Härdle, W., Liang, H., and Gao, J. (2000), *Partially Linear Models*, Physica Verlag.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC.

Hastie, T. and Tibshirani, R. (1993), "Varying-coefficient models," *Journal of the Royal Statistical Society. Series B. Methodological*, 55, 757–796.

He, X. and Shi, P. (1996), "Bivariate Tensor-Product B-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58, 162–181.

He, X., Zhu, Z., and Fung, W. (2002), "Estimation in a semiparametric model for longitudinal data with unspecified dependence structure," *Biometrika*, 89, 579–590.

He, X., Fung, W. K., and Zhu, Z. (2005), "Robust Estimation in Generalized Partial Linear Models for Clustered Data," *Journal of the American Statistical Association*, 100, 1176–1184.

Hettmansperger, T. and McKean, J. (1998), *Robust nonparametric statistical methods*, Arnold.

Hjort, N. and Pollard, D. (1993), "Asymptotics for minimisers of convex processes," *Statistical Research Report, University of Oslo*.

Hodges, J. and Lehmann, E. (1956), "The Efficiency of Some Nonparametric Competitors of The t-Test," *Ann. Math. Stat.*, 27, 324–335.

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

Hoover, D., Rice, J., Wu, C., and Yang, L. (1998), "Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data," *Biometrika*, 85, 809–822.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111–128.

Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821.

Huber, P. J. (1981), *Robust Statistic*, John Wiley & Sons.

Hunter, D. R. and Lange, K. (2000), "Quantile Regression Via an MM Algorithm," *Journal of Computational and Graphical Statistics*, 9, 60–77.

Hunter, D. R. and Li, R. (2005), "Variable Selection Using Mm Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Jin, Z., Ying, Z., and Wei, L. (2001), "A simple resampling method by perturbing the minimand," *Biometrika*, 88, 381–390.

Kai, B., Li, R., and Zou, H. (2009a), "Local CQR Smoothing: An Efficient and Safe Alternative to Local Polynomial Regression," *Accepted by Journal of the Royal Statistical Society, Series B*.

Kai, B., Li, R., and Zou, H. (2009b), "New Robust Statistical Procedures for Semiparametric Regression Models," *To be Submitted.*

Kauermann, G. and Tutz, G. (1999), "On model diagnostics using varying coefficient models," *Biometrika*, 86, 119–128.

Kim, M. (2007), "Quantile regression with varying coefficients," *Annals of Statistics*, 35, 92.

Knight, K. (1998), "Limiting Distributions for L_1 Regression Estimators under General Conditions," *The Annals of Statistics*, 26, 755–770.

Knight, K. and Fu, W. (2000), "Asymptotics for lasso-type estimators," *Ann. Statist*, 28, 1356–1378.

Koenker, R. (1984), "A note on L-estimates for linear models," *Stat. and Prob. Letters*, 2, 323–325.

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

Koenker, R. and Bassett, Gilbert, J. (1978), "Regression Quantiles," *Econometrica*, 46, 33–50.

Koenker, R. and Geling, O. (2001), "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis," *Journal of the American Statistical Association*, 96, 458–468.

Koenker, R. and Hallock, K. (2001), "Quantile Regression: An Introduction," *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R., Ng, P., and Portnoy, S. (1994), "Quantile smoothing splines," *Biometrika*, 81, 673–680.

Laird, N. M. and Ware, J. H. (1982), "Random-effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lee, S. (2003), "Efficient Semiparametric Estimation of A Partially Linear Quantile Regression Model," *Econometric Theory*, 19, 1–31.

Leng, C., Lin, Y., and Wahba, G. (2006), "A Note on the Lasso and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284.

Li, R. and Liang, H. (2008), "Variable selection in semiparametric regression modeling," *The Annals of Statistics*, 36, 261–286.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, D. Y. and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data," *Journal of the American Statistical Association*, 96, 103–126.

Lin, X. and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.

Lin, X. and Carroll, R. J. (2001a), "Semiparametric Regression for Clustered Data," *Biometrika*, 88, 1179–1185.

Lin, X. and Carroll, R. J. (2001b), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association*, 96, 1045–1056.

Mack, Y. and Silverman, B. (1982), "Weak and strong uniform consistency of kernel regression estimates," *Probability Theory and Related Fields*, 61, 405–415.

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

Maronna, R., Martin, R., and Yohai, V. (2006), *Robust Statistics: Theory and Practice*, John Wiley & Sons.

Marron, J. S. and Nolan, D. (1988), "Canonical Kernels for Density Estimation," *Statistics & Probability Letters*, 7, 195–199.

McKean, J. (2004), "Robust analysis of linear models," *Statistical Science*, pp. 562–570.

Miller, A. J. (2002), *Subset Selection in Regression*, Chapman & HALL/CRC, New York, second edn.

Müller, H. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, New York.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and its Applications*, 9, 141–142.

Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics*, 33, 1065–1076.

Pollard, D. (1991), "Asymptotics for Least Absolute Deviation Regression Estimators," *Econometric Theory*, 7, 186–199.

Powell, J., Stock, J., and Stoker, T. (1989), "Semiparametric estimation of index coefficients," *Econometrica: Journal of the Econometric Society*, pp. 1403–1430.

Preisser, J. S. and Qaqish, B. F. (1999), "Robust Regression for Clustered Data with Application to Binary Responses," *Biometrics*, 55, 574–579.

Richardson, A. M. (1997), "Bounded Influence Estimation in the Mixed Linear Model," *Journal of the American Statistical Association*, 92, 154–161.

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

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons.

Ruppert, D. and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An effective bandwidth selector for local least squares regression," *J. Amer. Statist. Assoc.*, 90, 1257–1270.

Ruppert, D., Wand, M., and Carroll, R. (2003), *Semiparametric Regression*, Cambridge University Press.

Schwarz, G. (1978), "Estimating the Dimension of a Model." *The Annals of Statistics*, 19, 461–464.

Serfling, R. (1980), "Approximation theorems of mathematical statistics," *New York*.

Severini, T. A. and Staniswalis, J. G. (1994), "Quasi-likelihood Estimation in Semiparametric Models (Corr: V89 P1572)," *Journal of the American Statistical Association*, 89, 501–511.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica*, 7, 221–264.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *Annals*, 5, 595–645.

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053.

Terpstra, J. and McKean, J. (2005), "Rank-based analyses of linear models using R," *Journal of Statistical Software*, 14.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the American Statistical Association*, 58, 267–288.

Wand, M. and Jones, M. (1995), *Kernel Smoothing*, Chapman & Hall/CRC.

Wang, H., Li, R., and Tsai, C. (2007), "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94, 553.

Wang, L., Kai, B., and Li, R. (2009), "Local Rank Inference for Varying Coefficient Models," *Tentatively accepted by Journal of the American Statistical Association.*

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhya*, 26, 359–372.

Welsh, A. H. (1996), "Robust estimation of smooth regression and spread functions and their derivatives," *Statist. Sinica*, 6, 347–366.

Welsh, A. H. and Richardson, A. M. (1997), *Approaches to the robust estimation of mixed models*, pp. 343–385, Elsevier.

Wild, C. J. and Yee, T. W. (1996), "Additive Extensions to Generalized Estimating Equation Methods," *Journal of the Royal Statistical Society, Series B: Methodological*, 58, 711–725.

Wu, C., Chiang, C., and Hoover, D. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data." *Journal of the American Statistical Association*, 93, 1388–1389.

Wu, Y. and Liu, Y. (2009), "Variable Selection in Quantile Regression," *Statistica Sinica*, 19, 801–817.

Xia, Y., Zhang, W., and Tong, H. (2004), "Efficient estimation for semivarying-coefficient models," *Biometrika*, 91, 661–681.

Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press.

Yu, K. and Jones, M. C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, 93, 228–237.

Zeger, S. L. and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.

Zhang, W. and Lee, S. (2000), "Variable bandwidth selection in varying-coefficient models," *Journal of Multivariate Analysis*, 74, 116–134.

Zhang, W., Lee, S., and Song, X. (2002), "Local Polynomial Fitting in Semivarying Coefficient Model," *Journal of Multivariate Analysis*, 82, 166–188.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, 36, 1509–1533.

Zou, H. and Yuan, M. (2008), "Composite Quantile Regression and the Oracle Model Selection Theory," *The Annals of Statistics*, 36, 1108–1126.

Vita

Bo Kai

Education

Ph.D. in Statistics (Expected) August 2009
Department of Statistics, The Pennsylvania State University, USA

M.S. in Statistics May 2008
Department of Statistics, The Pennsylvania State University, USA

M.S. in Probability and Statistics June 2004
Department of Probability and Statistics, Nankai University, China

B.S. in Mathematics June 2001
Department of Mathematics, Nankai University, China

Research Experience

Research Assistant 2007 – Present
The Methodology Center, The Pennsylvania State University, USA
Advisor: Prof. Runze Li

Working on series of topics, aimed at extending nonparametric smoothing techniques, semiparametric estimation and penalized likelihood methods to robust modeling and inference.

Research Assistant 2001 – 2004
Department of Probability and Statistics, Nankai University, China
Advisor: Prof. Shiyi Shen

This research involved probabilistic and semantic analysis of DNA, protein, amino acids sequences, protein backbone's structure modeling and error analysis, etc.

Teaching Experience

Instructor Spring 2008 & Summer 2006
Department of Statistics, The Pennsylvania State University, USA

I taught Stat/Math 414 (Introduction to Probability Theory) and Stat 401 (Experimental Methods).

Teaching Assistant 2004 – 2007
Department of Statistics, The Pennsylvania State University, USA

I taught labs, graded homework, held office hours, proctored exams for a variety of courses from Stat 100 level to Stat 500 level.