UNIVERSITY OF CALIFORNIA
RIVERSIDE

Essays on Nonparametric Based Modal Regression Econometrics

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Tao Wang

June 2022

Dissertation Committee:

Dr. Aman Ullah, Co-Chairperson
Dr. Weixin Yao, Co-Chairperson
Dr. Tae-Hwy Lee

The Dissertation of Tao Wang is approved:

Committee Co-Chairperson

Committee Co-Chairperson

University of California, Riverside

Acknowledgments

I am eternally grateful to my advisors, Dr. Aman Ullah and Dr. Weixin Yao, for their invaluable guidance and unequivocal support throughout my days in the econometrics field to which I aspire. Without their supervision, I would not have divided into the econometrics world and begun the fantastic academic journey. I am especially thankful for all of their time and effort in assisting me in establishing myself as an independent researcher. They also taught me how to cope with the ups and downs of life and inspired me to never give up, which has had a significant positive impact on my academic career.

I would like to extend my sincere thanks to my dissertation committee member, Dr. Tae-hwy Lee, for his support and advice on my research. His enthusiasm for research and concern for students mean more to me than I can express.

I would not have been able to accomplish my Ph.D. without the people in the Department of Economics who supported me during my time here. A particular appreciation goes out to the wonderful administrative assistant, Mr. Gary Kuzas, for his unwavering support. I also thank all of my classmates and friends for their inspiration. Because of them, I am able to strike a good balance between life and research.

Finally, I would like to convey my heartfelt gratitude to my parents and my wife for everything they have done. Their unqualified love, tremendous understanding, and never-ending encouragement have enabled me to progress to where I am now.

To my parents and my wife for all the support.

ABSTRACT OF THE DISSERTATION

Essays on Nonparametric Based Modal Regression Econometrics

by

Tao Wang

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2022
Dr. Aman Ullah, Co-Chairperson
Dr. Weixin Yao, Co-Chairperson

Most research on nonparametric econometrics focuses on mean, median, or quantile regression while there is not too much research about regression methods on the basis of mode value. This dissertation proposes three new models based on modal regression, in which the dependence of the conditional mode of the response variable on the covariates is explored and a kernel based objective function to simplify the computation is employed. In particular, this dissertation is made up of three essays. Chapter 1 provides an overview of the dissertation. Chapter 2 proposes a control function approach to account for endogeneity in a parametric linear triangular simultaneous equations model for modal regression, where the conditional mode of the unobservable error term on explanatory variables is nonzero. To motivate the developed control function method, a dynamic model of rational behavior under uncertainty is introduced, in which the agent maximizes the present discounted value of the stream of future modal utilities, and a modal Euler equation derived from the maximization model that the agent must satisfy in equilibrium is presented. In a general setting that includes nonlinear time series models as a special case, Chapter 3 develops a novel lo-

cal linear estimator of volatility function for nonparametric modal regression applied to the squared residuals from the unknown mean regression, which is particularly useful to serve as a risk indicator for skewed data or financial time series with heavy tails. To reduce the variance of the nonparametric modal volatility estimator, a variance reduction technique is introduced to achieve asymptotic relative efficiency while keeping the asymptotic bias unchanged. Furthermore, to avoid the negative values of volatility, Chapter 3 introduces a local exponential modal estimation. Chapter 4 investigates the estimation and inference of modal regression near the boundary, establishing a theoretical foundation for regression discontinuity designs based on mode value. Under the assumption of mode rank invariance, a novel conditional mode treatment effect in the regression discontinuity designs is proposed, which can be regarded as an attractive complement to the existing mean or quantile treatment effect. The novel mode treatment effect suggested in Chapter 4 has a wide range of applications in economics, statistics, social science, and other related fields, because it can capture the "most likely" effect and be robust to outliers and heavy-tailed distributions. Chapter 5 contains the conclusions. The newly proposed models based on modal regression in this dissertation complement the mean, median, and quantile regressions and provide a better central tendency measure when the data are skewed or heavy-tailed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The mean, median, and mode are three of the most commonly and popularly used location measures and focus on different population characteristics. Each quantity has its own merit and complements each other. Built on the ideas of mean and median, mean regression and median regression have been extensively investigated and popularly used to model the relationship between a dependent variable Y and covariates X. However, research about the regression model built on the concept of mode (called *modal regression*) is rather limited and has not received enough attention that it deserves, partly due to its computational difficulty. Modal regression can supplement mean and median regressions and provide additional useful information that existing regression models might miss, especially for multimodal or skewed datasets. To broaden the scope of existent modal regressions, this dissertation investigates three new models based on mode value.

Chapter 2 proposes a control function approach to account for endogeneity in a parametric linear triangular simultaneous equations model for modal regression, where

the conditional mode of the unobservable error term on explanatory variables is nonzero. We adjust endogeneity with the residuals from the conditional mode decomposition of the endogenous variable as controls in the structural equation, and develop a computationally attractive two-step estimation procedure with the conditional mode independence restriction. Notably, in the first step, we construct the estimated modal residuals from the reduced-form linear modal regression for the endogenous variable; in the second step, we include the reduced-form residual nonparametrically as an additional variable and propose a three-stage estimation method for the resulting semiparametric partially linear modal regression model, which has not been fully investigated in the literature. The consistency and asymptotic properties of the estimators for both the parametric and nonparametric parts are rigorously established under generic regularity conditions, where we demonstrate that the parametric estimator is $(nh^3)^{1/2}$-consistent (n is the sample size and h is a bandwidth) and the estimation of the nonparametric component is oracle. To motivate the developed control function method, we introduce a dynamic model of rational behavior under uncertainty, in which the agent maximizes the present discounted value of the stream of future modal utilities, and develop a modal Euler equation derived from the maximization model that the agent must satisfy in equilibrium. We estimate the modal elasticity of intertemporal substitution directly from the stochastic Euler equation. We in the end construct an adaptive least absolute shrinkage and selection operator technique for selecting instrumental variables and demonstrate the oracle property of the suggested penalized modal regression model. Since mode is identical to mean with symmetric data, several remarks on modal-based control function robust estimation are also addressed for completeness.

In a general setting that includes nonlinear time series models as a special case, Chapter 3 proposes a novel local linear estimator of volatility function for nonparametric modal regression applied to the squared residuals from the unknown mean regression, which is particularly useful for skewed data or financial time series with heavy tails. We show that the proposed modal volatility estimator can be obtained asymptotically as well as if the conditional mean regression function were given, assuming that the observations are from a strictly stationary and absolutely regular process. Under mild regularity conditions, the asymptotic distributions are also proven to be the same as those derived from independent observations, but the convergence rate is slower than that in nonparametric mean regression. Moreover, we put forward a variance reduction technique in terms of modal volatility estimator to attain asymptotic relative efficiency while maintaining the asymptotic bias unchanged. For the purpose of avoiding the drawback of negative estimates, we in the end discuss the extension of the method to local exponential modal estimation and demonstrate that the suggested exponential modal volatility estimator shares exactly the same asymptotic variance as the local linear modal volatility estimator under some mild conditions, but could have a smaller bias.

Chapter 4 investigates the estimation and inference of modal regression near the boundary, establishing a theoretical foundation for regression discontinuity (RD) designs based on mode value. Under the assumption of mode rank invariance, we propose a novel conditional mode treatment effect (CMTE) in the RD designs, which is especially useful for skewed or heavy-tailed data and can be regarded as an attractive complement to the existing mean or quantile treatment effect. For the sake of exposition, we primarily concentrate on

the modal sharp RD design to convey the fundamental concept of CMTE. Analogously to the estimators for the existent treatment effects in the RD designs, we approach the modal regression function as nonparametric, develop a local boundary estimation procedure, and show that the CMTE is identified under moderate assumptions. The consistency and asymptotic properties of the suggested estimator are presented under certain regularity conditions. To construct a trustworthy confidence interval, we develop an efficient bootstrap procedure for practical application depending on undersmoothing. We also discuss several extensions that are of either practical or theoretical importance, including the CMTE in the modal fuzzy RD design.

Chapter 2

Endogeneity in Modal Regression

2.1 Introduction

Modal regression, which seeks to identify the most probable conditional value (mode) of a dependent variable $Y \in R$ given covariates $X \in R^{d_X}$, denoted by $Mode(Y \mid X)$, rather than the mean or quantile used by traditional mean or quantile regression, has received increasing attention in recent econometric practice. It can reveal a novel and intriguing data structure that would otherwise be overlooked by conditional mean or quantile regression without the use of any moment conditions. The Cauchy distribution, for example, is widely recognized for not having a mean or variance. Consequently, the sample mean is a poor estimate that is not consistent, but the mode can be estimated. Even in Bayesian econometric analysis, the mode of the posterior distribution, if skewed, is considered as a Bayesian estimator instead of the mean or median estimator. In comparison to existing mean and quantile regressions, modal regression is more resistant to outliers and some forms of measurement error, is applicable to clustered or inhomogeneous data, is capable of achieving consistent

estimates for truncated data, and can provide shorter prediction intervals when the data are skewed since with the same interval length, the interval around the conditional mode encompasses more samples. Although quantile regression can offer more information about the conditional distribution away from the centre, it cannot directly give the modal estimate when researchers are primarily concerned with the "most likely" effect. These observations suggest that modal regression can be served as a complement to the existing mean or quantile regression. Notable works for estimating modal regression with multivariate data by imposing certain modal structures and a unique global mode assumption on $Mode(Y \mid X)$ include Lee (1989, 1993), Kemp and Santos Silva (2012), Yao and Li (2014), Chen et al. (2016), Yao and Xiang (2016), Krief (2017), Ota et al. (2019), Kemp et al. (2020), Feng et al. (2020), Zhang et al. (2020), Ullah et al. (2021, 2022), among others.[1] Nevertheless, to the best of our knowledge, there has not been any attempt to investigate the presence of endogeneity in modal regression by *permitting the conditional mode value of the unobservable error term on the explanatory variable to be nonzero*. It has been implicitly assumed by all of the research in mode estimation that there is no endogeneity in models, which is unnecessarily strong in practice and restricts the breadth of empirical applications of modal regression to a few instances. A natural question then arises is how to identify modal coefficients in the presence of such an endogeneity issue.

Endogeneity, resulting from measurement error, individual choice, or market equilibrium, lies at the heart of many problems in econometrics and statistics, manifesting itself

[1]For univariate X ($d_X = 1$), we can apply the nonparametric kernel density estimation method to estimate mode, i.e., $Mode(Y \mid X) = \arg\max_Y f_{Y|X}(Y \mid X)$, where $f_{Y|X}(Y \mid X)$ is the continuous conditional density of Y given X. With a given X, it will be the same as maximizing the joint density of Y and X. However, such a method is difficult to apply when the dimension of covariates is large owing to the "curse of dimensionality". Additionally, the kernel density estimation method cannot directly give an estimate of the marginal effect unless we take the finite difference approximations. Thus, along the lines of the mean or quantile regression, lots of research proposes estimating modal parameters with commonly imposed restrictions on the function form of $Mode(Y \mid X)$.

in various instances, such as socioeconomic variables of education-wage and supply-demand. In mean regression, endogeneity can be interpreted as either the nonindependence between the explanatory variables and the error term or the nonzero value of the conditional expectation of the error term given covariates. However, the conceptual foundation of endogeneity in modal regression is presently unsolved in the literature since modal estimation does not require any moment constraints. Analogous to quantile regression, we in this paper *interpret endogeneity in modal regression as the nonzero value of the conditional mode of error term given covariates*. We point out that if the moments of the error term and explanatory variables exist, we could also utilize the conventional definition of endogeneity in mean regression to define endogeneity in modal regression. Generally, when endogeneity is present in modal regression, the model will be sufficiently different from the one without endogeneity to necessitate separate treatment, because the standard kernel-based estimation method (defined in Section 2.3) will potentially produce biased and inconsistent estimators. This negative result motivates us to fill the literature gap by allowing for the possibility of endogeneity in modal regression and systematically studying its estimation procedure and asymptotic behavior under weak conditions. To this end, we extend the control function approach (Smith and Blundell, 1986; Newey et al., 1999; Blundell and Powell, 2003), which is essentially different from the traditional instrumental variable approach that forms moment conditions for estimation, to a semiparametric partially linear modal regression version of the triangular simultaneous equations model in order to deal with endogeneity. Notice that the similar control function approach used to correct for endogeneity in mean and quantile nonparametric structural models has been adopted by Newey et al. (1999), Ma and Koenker

(2006), Li et al. (2007), Su and Ullah (2008), Imbens and Newey (2009), Kim and Petrin (2011), Chernozhukov et al. (2015), and references cited therein with the mean or quantile independence restriction.

In comparison to the previous literature, we primarily make four important contributions. *The first* is to parametrically incorporate endogenous regressors into the modal regression and develop a control function estimation method to account for the endogeneity of the regressors in the original structural equation. We do not restrict the functional form of the control function to avoid any potential misspecifications, and novelly introduce a mode independence condition (clarified in Section 2.2) to identify the model and retrieve the parameters of interest. On this basis, we discuss briefly the extension of the proposed endogeneity framework to nonparametric simultaneous model regressions as well. *The second* is to establish a three-stage estimation method for a pseudo semiparametric partially linear modal regression after including the estimated residual from the reduced form equation as an additional variable into the structural equation. We thus contribute significantly to the large and still growing literature on mode estimation,[2] since there is no previous research systematically investigating partially linear modal regression on estimation and asymptotic behavior of both parametric and nonparametric estimators based on local linear approximation. *The third* is to propose a penalized modal regression model with an adaptive least

[2]The partially linear model, $Y_i = Z_i^T\gamma + m(V_i) + \epsilon_i$ (see model settings in Section 2.2 for the meaning of each part), was initially introduced by Engle et al. (1986) to study the relationship between electricity sales and temperature based on mean regression, and has attracted much attention from econometricians in both theory and empirical applications since then; see Heckman (1986), Robinson (1988), Bhattacharya and Zhao (1997), Krief (2017) to mention only a few. The partially linear modal regression is of interest for several reasons. First, it can capture the nonlinear relationship between regressor and dependent variable while avoiding the "curse of dimensionality" through the presence of a linear function. As the nonparametric part liberates the model from strict structural assumptions, the estimate of γ is less impacted by model bias. Second, it enables the parametric and nonparametric components to exist simultaneously in the model, nesting both the linear modal regression with $V_i = 0$ and the nonparametric modal regression with $Z_i = 0$.

absolute shrinkage and selection operator (LASSO) to identify the relevant instrumental variables, in which we shall show that irrelevant instruments will be estimated as zero as if we knew they were. In empirical applications, it is common for researchers to collect a large number of instrumental variables in order to improve the precision of estimators. However, to our limited knowledge, there still lacks study on instrumental variable selection in penalized modal regression with adaptive LASSO. *The fourth* is to construct a modal Euler equation derived from a dynamic model of rational behavior under uncertainty, in which the agent maximizes the present discounted value of the stream of future modal utilities. Such a modal Euler equation can be used to supplement the existing expected utility models by revealing the distinguishing features of the data, as well as capturing the "most likely" effect. Empirically, we utilize the suggested control function approach to account for endogeneity to estimate the modal elasticity of intertemporal substitution (EIS) directly from the stochastic Euler equation. According to the aforementioned contributions, the newly proposed model in this paper highlights the attractiveness of modal regression and opens up a wide range of potential applications for empirical work. Since the focus of this paper is on modal regression (asymmetric data), we do not treat the case where modal regression line is identical to mean regression line as the main analysis (symmetric data); see Figure 2.1. Nonetheless, to illustrate the effectiveness of modal regression, we make several remarks throughout the paper to show that with symmetrically distributed data, the proposed estimation procedure can be regarded as a modal-based robust control function methodology that achieves mean estimators against outliers or aberrant observations without sacrificing efficiency.

The content of this paper is organized as follows. In Section 2.2, we primarily introduce model settings with endogeneity and outline the estimation framework using the mode independence condition. In Section 2.3, we focus on developing a two-step estimation procedure relying on local linear approximation to estimate a semiparametric partially linear modal regression, as well as providing a modified modal expectation-maximization (MEM) algorithm to simplify computations. In Section 2.4, we establish large sample properties of the resulting estimators in different steps/stages under suitable conditions and discuss bandwidth choice in practice utilizing asymptotic theorems. Several remarks on modal-based robust estimators with symmetric data are provided as well. The results of Monte Carlo simulations and applications to two real datasets of Rational Behavior under Modal Utility Maximization and Colonial Origins of Comparative Development are reported in Section 2.5 to show the necessity of correcting endogeneity in modal regression. The adaptive LASSO method for selecting relevant instruments for the proposed penalized modal regression model is presented in Section 2.6. Conclusions and discussions are given in Section 2.7. The additional results, including a Modal Asset Pricing Model and the numerical results for the Return to Schooling dataset and modal-based robust estimation, are deferred to the appendix, along with all technical proofs.

2.2 Econometric Framework

The modal regression model considered for the independent and identically distributed $(i.i.d.)$ observations $\{Y_i, X_i, Z_i\}_{i=1}^n$ from the random vector (Y, X, Z) has the following triangular system with the parametric form

$$\begin{cases} Y_i = X_i\beta + Z_{1,i}^T\gamma + U_i \text{ (structural equation)}, \\ \\ X_i = \alpha + Z_i^T\pi + V_i \text{ (reduced form equation)}, \end{cases} \tag{2.1}$$

where $Y_i \in R$ is a real valued continuously distributed random scalar, $X_i \in R$ is an endogenous explanatory variable, $Z_i = (Z_{1,i}^T, Z_{2,i}^T)^T \in R^{dZ}$ is an observed vector of exogenous explanatory variables in which $Z_{1,i} \in R^{dZ_1}$ and $Z_{2,i} \in R^{dZ_2}$, $Z_{2,i}$ is a vector of excluded instruments, U_i is the unobservable error term of the structural equation, V_i is the unobservable error term of the reduced form equation, which is interpreted as the deviation of X_i from its conditional mode $Mode(X_i \mid Z_i)$, β and γ are 1×1 and $d_{Z_1} \times 1$ unknown structural parameters of interest, respectively, and α and π are 1×1 and $d_Z \times 1$ unknown parameters, severally.[3] We use T to denote the transpose of a matrix or vector. The setup in (2.1) is similar to the endogenous quantile regression considered in Li et al. (2007), with the exception that we concentrate on modal regression. It should be noted that U_i and V_i are allowed to be statistically dependent on each other according to the setting. As equations in (2.1) are investigated under modal regression, we impose the following conditions

$$Mode(V_i \mid Z_i) = 0 \text{ (almost surely)}$$

because of the exogeneity of Z_i and

$$Mode(U_i \mid X_i, Z_i) \neq 0 \text{ (almost surely)}$$

according to the endogeneity of X_i. For identification, we do not include any constants in the structural equation, and assume the standard rank condition that the dimension

[3] As the first paper dealing with endogeneity in modal regression, we restrict attention to the parametric form for the purpose of exposition. However, the proposed control function technique can be extended to other semiparametric/nonparametric models, which is straightforward but will raise a considerably complicated identification and estimation problem; see Remark 2.2.4.

d_{Z_2} of $Z_{2,i}$ is equal to (or greater than) one and there exists at least one nonzero modal coefficient for $Z_{2,i}$ (we further explain the identification for modal regression in Section 2.3). For simplicity of exposition, we focus on the univariate case of X_i. However, the proposed estimation procedure and asymptotic results in this paper can be easily extended to the multivariate case, but with more complicated notations involved.

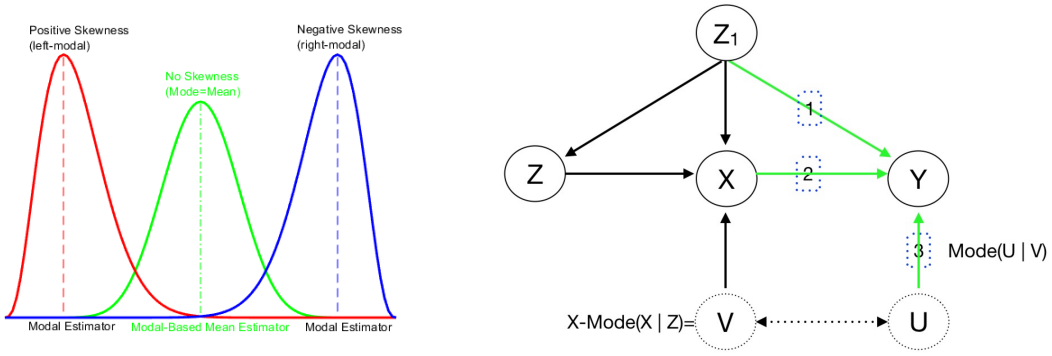

Figure 2.1: Data Distribution and Directed Acyclic Graph

To identify and estimate the modal parameters β and γ, based on the conditions illustrated above, we further assume that the following equations are satisfied almost surely

$$\begin{cases} Mode(Y_i \mid X_i, Z_i, V_i) = X_i\beta + Z_{1,i}^T\gamma + Mode(U_i \mid X_i, Z_i, V_i) \\ \qquad\qquad\qquad\qquad = X_i\beta + Z_{1,i}^T\gamma + Mode(U_i \mid \alpha + Z_i^T\pi + V_i, Z_i, V_i) \\ \qquad\qquad\qquad\qquad = X_i\beta + Z_{1,i}^T\gamma + Mode(U_i \mid V_i, Z_i) \\ \qquad\qquad\qquad\qquad = X_i\beta + Z_{1,i}^T\gamma + Mode(U_i \mid V_i), \\ Mode(X_i \mid Z_i) = \alpha + Z_i^T\pi \end{cases} \tag{2.2}$$

with the exclusion restriction of a mode independence of U_i on Z_i conditional on V_i, indicating that V_i is sufficient in evaluating the conditional mode $Mode(U_i \mid V_i)$ such that

Z_i does not provide any extra information. This mode restriction suffices if U_i has an equal mode for all values of Z_i conditional on the error term V_i, which is analogous to the usual orthogonality condition for a linear model considered in mean or quantile regression ($U_i \mid X_i, Z_i \sim U_i \mid V_i, Z_i \sim U_i \mid V_i$ in which $\sim$ indicates the equality of conditional distributions) (Newey et al., 1999; Blundell and Powell, 2004; Su and Ullah, 2008), but weaker than the independence between U_i and (X_i, Z_i) conditional on V_i. The triangular structure can be explained more intuitively in Figure 2.1 through a directly acyclic graph, which clearly shows that there are three pathways affecting dependent variable Y, while the effect of U through Path 3 is indirectly influenced by variable V. Without accounting for the causal effect of V on U, the modal estimators will be biased and inconsistent. Observe that the crucial condition for estimating, as shown in Figure 2.1, is that Z affects X but does not directly impact U. The specific motivating examples of (2.2) include estimation of return to education and production functions. For example, Card (2001) discussed the control function approach in the measurement of the causal effect of education on labor market earnings, and argued that in certain instances, it is reasonable to assume that individual ability (U_i) is a function of the schooling residual (V_i) .

Remark 2.2.1 *Compared to the distributional independence restriction, the mode independence assumption* $Mode(U_i \mid V_i, Z_i) = Mode(U_i \mid V_i)$ *is not very restrictive and is a natural extension of the existing mean or quantile independence condition. Especially, Newey et al. (1999) employed series approximations with the conditional mean independence to present a two-step nonparametric estimator for a triangular simultaneous equation model; Li et al. (2007) imposed the conditional quantile independence with series approximations for solv-*

ing the endogeneity in quantile regression models; and Su and Ullah (2008) made use of the conditional mean independence to propose a local polynomial estimator for the nonparametric simultaneous equations model. We emphasize that the imposed mode independence assumption is intended to simplify computations and illustrations because it can alleviate the "curse of dimensionality" to some extent. We can release this assumption and instead allow the conditional mode of the structural error to depend on both the reduced residuals and the instruments, which is a promising direction for future research but beyond the scope of the present paper.

Define $Mode(U_i \mid V_i) = m(V_i)$ as a real-valued unknown function of V_i on R that maps the reduced form error into the structural equation error, we have

$$\begin{cases} Y_i = X_i\beta + Z_{1,i}^T\gamma + m(V_i) + \underbrace{U_i - m(V_i)}_{\text{new error term}}, \\ \\ Mode(Y_i \mid X_i, Z_{1,i}, V_i) = X_i\beta + Z_{1,i}^T\gamma + m(V_i), \end{cases} \tag{2.3}$$

which suggests that the parameters of interest can be recovered through a *semiparametric partially linear modal regression* that has not been extensively considered in the literature of modal regression. To avoid introducing model misspecifications, we do not impose any structural assumptions on the function form of $m(\cdot)$ (but it needs to satisfy certain smoothness properties). As shown in (2.3), we correct for modal endogeneity by including the estimate of V_i as an additional variable, which can be thought of as a variant of the control function approach that handles endogeneity issue as an omitted variable problem. Therefore, the new error term, $U_i - m(V_i)$, is orthogonal to $(X_i, Z_{1,i}, V_i)$ in mode sense by construction, implying that the conditional mode is equal to zero.

Remark 2.2.2 *The instrument Z_i is served to remove the exogenous variations away from X_i in mode sense. As mode does not have the additive property in general, we cannot apply the traditional two-stage least squares (2SLS) estimation method for modal regression. However, analogous to quantile regressions with endogeneity, we can also use the "fitted value" approach to recover the parameters of interest by imposing constraints on the mode of the reduced-form error. To be more precise, we substitute X_i with $\alpha + Z_i^T\pi + V_i$ to obtain $Y_i = [\alpha + Z_i^T\pi]\beta + Z_{1,i}^T\gamma + (U_i + \beta V_i)$. With the assumption that the conditional mode of the composite error term $U_i + \beta V_i$ is independent of Z_i, i.e., $Mode(U_i + \beta V_i \mid Z_i) = Mode(U_i + \beta V_i)$, we are able to consistently estimate β and γ and establish the appropriate asymptotic properties, which can be treated as an alternative manner to address endogeneity in modal regression. Although this approach is attractive due to the simplicity of calculation, it imposes too strong assumptions on the error term. We note that the assumption of independence between the composite error term and the instruments might be challenging to maintain in practice, particularly when the reduced form error term is not independent of the instruments (heteroskedasticity); see the relevant discussion in Blundell and Powell (2004) and Li et al. (2007).*

If the realizations of $\{V_i\}_{i=1}^n$ were observable, we could directly estimate the semi-parametric modal regression (2.3) with the kernel-based objective functions (see Section 2.3 for more information). However, they are not observable in practice and must be substituted with consistent estimates. To that end, we propose a fairly simple but efficient two-step estimation procedure for recovering the values of parameters of interest. In particular, in the *first step*, we construct the estimated modal residual $\hat{V}_i$ from the reduced

form equation using linear modal regression of the endogenous variable X_i on Z_i. In the *second step*, we correct the endogeneity bias by performing semiparametric partially linear modal regression of Y_i on X_i, $Z_{1,i}$ and $\hat{V}_i$ based on local linear approximation,[4] which is of independent interest, but neither the asymptotic nor its finite sample properties have been thoroughly investigated in the literature.

Due to the existence of different convergence rates of parametric and nonparametric components, the modal estimators in the second step can be formatted with a three-stage estimation method to effectively rake the partially linear structure into account. Particularly, in the *first stage*, the local linear approximation is adopted to transform the original partially linear modal regression to a local linear modal regression. In the *second stage*, the parametric coefficients are estimated by linear modal regression after plugging in the estimator of the nonparametric part in the first stage. *Finally*, the parametric estimators from the second stage are plugged into the original model to obtain the nonparametric modal regression, and the estimators are obtained once more employing local linear approximation. The estimation of the modal regression coefficients in different steps/stages can be easily achieved by virtue of a modified modal expectation-maximization (MEM) algorithm introduced in Yao et al. (2012) and Yao (2013). We derive the limiting distributions of the proposed estimators for both parametric and nonparametric components under mild conditions. Perhaps not surprisingly, the asymptotic theorems for the proposed modal estimators

[4]In order to construct consistent estimators of the unknown parameters, extra care should be taken, especially in light of the dominance of the parametric component of the model. The consistency of parametric estimators in the first step is critical not just in and of itself, but also for the identification of modal coefficients in the structural equation. Under certain mild bandwidth conditions, there should be minimal difference between the model (in the second step) investigated in this paper and the partially linear modal regression after replacing the unobservable $\{V_i\}_{i=1}^n$ with their consistent estimates. We demonstrate that our estimators have the oracle property provided that the preliminary estimator converges sufficiently fast.

stay the same as those for the oracle case where V_i and other components were known, provided the bandwidths from different steps/stages are chosen in an appropriate way to avoid bias effects from previous step/stages. A similar oracle property for the modal estimator has been observed in the fixed effects modal regression for panel data investigated by Ullah et al. (2021). Especially, we argue that the parametric components can be estimated at the usual parametric modal convergence rate, and conclude that the final stage nonparametric component estimators have asymptotic bias and variance equivalent to those of the nonparametric local linear modal estimators under some primitive conditions specified in Section 2.4. We further discuss the choice of bandwidths for the newly proposed model in practice according to the asymptotic results. To demonstrate the finite sample performance of the resultant modal estimators, we present several numerical results, including Monte Carlo experiments and empirical data analyses.[5]

Remark 2.2.3 *To estimate the partially linear mean regression model, Robinson (1988) proposed a two-step estimation method, in which the first step obtains a consistent estimator of the unknown conditional mean function, and the second step estimates a simple linear mean regression to recover the parameters by concentrating out the unknown function. Such a method is convenient and powerful for dealing with semiparametric models. However, it is not applicable to partially linear modal regression model because mode does not have the additive property (i.e.,* $Mode(U_i + V_i) \neq Mode(U_i) + Mode(V_i)$*) in general, unless data are subjected to a strict symmetric distribution, where mode is identical to mean. Nonetheless,*

[5]We also present a modal asset pricing model in the appendix, where we solve the standard intertemporal problem of a consumer-investor agent and consider a two-period economy with two assets. We then dispose modal Euler equations derived from the maximization models that the agent must satisfy in equilibrium. Such a model is considered from modal regression with endogeneity and can be estimated with the suggested control function method, indicating the broad applicability of the proposed model.

modal-based robust estimation in Remark 2.3.8 will be preferred in this case (symmetric data) because of the faster convergence rate.

Furthermore, in empirical applications, researchers may have a large number of instrumental variables but are unsure which ones to include in the analysis. It is undesirable to retain irrelevant instrumental variables in the model since this may lead to decreased modeling accuracy. As a result, a theoretically optimal method for instrumental variable selection in the proposed modal regression model is necessary. In the regularization framework, many different types of penalties introduced in the machine learning community belonging to the group of shrinkage methodologies have been utilized to achieve variable selection taking the form of "loss function + penalty", but not too much attention has been paid to the two-step control function model based on mode value. We then in the last section of this paper concentrate on the first step equation for a regularization setting, where we employ the adaptive LASSO method (Zou, 2006) to select relevant instruments with probability tending to one (i.e., sparsity) and simultaneously estimate the nonzero modal coefficients.[6] We shall show that the irrelevant instruments are estimated to be zero as if they were known. The asymptotic normality of the adaptive LASSO modal estimator is established as well (i.e., oracle property in the sense introduced by Fan and Li (2001)).

Remark 2.2.4 (Nonparametric Simultaneous Modal Regressions) *We in this paper concentrate on the parametric form of modal regression for easy illustration. However, the requirement of a pre-determined functional form can increase the risk of model misspec-*

[6]We realize that Fan (2012) suggested an instrumental variable selection procedure for the traditional 2SLS estimation based on the adaptive LASSO method with a fixed number of variables. We primarily adopt such a procedure as well to the proposed modal regression to construct a regularization framework, which extends the usefulness of the adaptive LASSO beyond variable selection for mean or quantile regression.

ification and lead to invalid estimates. In practice, it may be more realistic to investigate nonparametric modal regression with endogeneity. To show the applicability of the developed econometric framework, we release the strict parametric assumption, consider a slight extension of the model, and investigate the following equations

$$\begin{cases} Y_i = g(X_i, Z_{1,i}) + U_i \ (\textit{structural equation}), \\ X_i = h(Z_i) + V_i \ (\textit{reduced form equation}), \end{cases}$$

where $g(\cdot)$ and $h(\cdot)$ are real-valued (non-constant) functions satisfying certain smoothness properties. The other model settings are identical to those in the paper. A similar equation under the content of mean regression has been investigated by Newey et al. (1999) and Su and Ullah (2008), demonstrating that $g(\cdot)$ is identified up to an additive constant if there is no functional relationship between (X, Z_1) and V. We are interested in estimating $g(\cdot)$ and its derivatives based on modal regression. With the conditional mode independence condition, we have

$$Mode(Y_i \mid X_i, Z_i, U_i) = g(X_i, Z_{1,i}) + Mode(U_i \mid V_i),$$

which presents an additive structure and can be solved by the three-step estimation procedure described below: (1) Produce a consistent estimate of $h(Z_i)$ by performing local linear modal regression of X_i on Z_i. Denote the estimate as $\hat{h}(Z_i)$ and calculate the estimated residual $\hat{V}_i$, where $\hat{V}_i = X_i - \hat{h}(Z_i)$; (2) Obtain a consistent estimator of $m(x, z_1, v)$, denoted as $\hat{m}(x, z_1, v)$, by conducting local linear modal regression of Y_i on X_i, $Z_{1,i}$, and $\hat{V}_i$; (3) Estimate $g(x, z_1)$ consistently up to an additive constant by $\hat{g}(x, z_1) = \int \hat{m}(x, z_1, v) dQ(v)$, where $Q(\cdot)$ is a deterministic weighting function with $\int dQ(v) = 1$. The asymptotic properties can be established using the same arguments as in this paper, where we can show that

the replacement of the unobserved residuals has no effect on the asymptotic properties of the resulting estimators. Future research may provide relevant findings by examining this expansion model in more depth.

2.3 Two-Step Modal Estimation

We in this section propose a two-step estimation procedure to satisfactorily recover the parameters of interest in (2.3), where the second step can be converted into a computationally feasible three-stage estimation method, and describe the algorithm to numerically estimate the models. Before presenting the suggested estimation procedure, we define the modal estimator as follows.

Definition 1 *Given kernel function $K(\cdot)$, bandwidth h, and the unique global mode assumption for $f_{Y|X}(Y \mid X)$, the modal estimator of θ with respect to modal function $m(X, \theta)$ is defined as*

$$\hat{\theta} = \arg\max_{\theta} E[L_\theta(X, Y)], \text{ where } L_\theta(X, Y) = \frac{1}{h} K\left(\frac{Y - m(X, \theta)}{h}\right)$$

and θ belongs to a compact parameter space Θ.

The above definition is understandable from the machine learning perspective and kernel density estimation. According to Feng et al. (2020), the modal estimator θ can be defined as $\hat{\theta} = \arg\max_\theta \int_X f_{Y|X}(m(X, \theta) \mid X) df_X(X)$, where $f_X(\cdot)$ is the marginal density of X. Given empirical observations, we can transform the density maximization problem over some hypothesis spaces into a task of maximizing the kernel density function and effectively achieve Definition 1. Let $g_\varepsilon(\varepsilon)$ be the continuous density function of $\varepsilon = Y - Mode(Y \mid$

$X) = Y - m(X, \theta)$. Under the maintained assumption of h such that $h \to 0$ with sample size increasing, we have

$$\begin{aligned} \sup_{\varepsilon \in R} |g_\varepsilon(\varepsilon) - \int K(w) g_\varepsilon(\varepsilon + wh) dw| &\leq \sup_{\varepsilon \in R} \int |g_\varepsilon(\varepsilon) - g_\varepsilon(\varepsilon + wh)| K(w) dw \\ &\leq \sup_{\varepsilon \in R} \int |g_\varepsilon^{(1)}(\varepsilon) wh| K(w) dw = |g_\varepsilon^{(1)}(\varepsilon)| h \int |w| K(w) dw \to 0, \end{aligned} \tag{2.4}$$

where $g_\varepsilon^{(1)}(\varepsilon)$ denotes the first derivative of $g_\varepsilon(\varepsilon)$. Hence, there exists a modal parameter θ that can maximize the density of ε, which establishes the underlying modal estimation mechanism. Completely different from the bandwidth in nonparametric estimation determining the smoothness of the function, the bandwidth h in modal regression controls the estimation of mode and the balance between robust estimate (h is treated as a constant) and mode estimate (h depends on sample size).

In accordance with the above definition, we develop a two-step estimation procedure to obtain modal estimates. The **first step** is the construction of estimated residuals $\{\hat{V}_i\}_{i=1}^n$ in the reduced form equation using linear modal regression of X_i on Z_i. Specifically, we maximize the following global kernel-based objective function[7]

$$Q_n(\alpha, \pi) = \frac{1}{nh} \sum_{i=1}^{n} \phi \left(\frac{X_i - \alpha - Z_i^T \pi}{h} \right), \tag{2.5}$$

where $\phi(\cdot) : R \to R$ is a nonnegatively symmetric kernel, and $h = h(n) \to 0$ as $n \to \infty$ is a sequence of positive bandwidth utilized in this step that depends on sample size n. To prevent notation confusion, we suppress the n for all bandwidths used in this paper. As stated in Yao and Li (2014) and Ullah et al. (2021), the choice of kernel function in modal regression has less impact on the asymptotic behavior of estimators compared to the choice

[7]When $\pi = 0$, the objective function (2.5) becomes $\frac{1}{nh} \sum_{i=1}^{n} \phi \left(\frac{X_i - \alpha}{h} \right)$, which is the kernel estimate of the density function of X_i at $X_i = \alpha$. Therefore, the maximizer of the preceding equation will be the mode value of X_i. With $n \to \infty$ and $h \to 0$, it will converge to the mode of distribution of X_i under certain conditions based on Definition 1, which is the fundamental concept of modal regression coefficient estimation.

of bandwidth. We thus use $\phi(\cdot)$ as a normal kernel to form a closed-form expression in the M-Step of the following MEM Algorithm 1.[8] The general conditions imposed on the kernel function are discussed in Section 2.4. The solutions of (2.5), represented by $\hat{\alpha}$ and $\hat{\pi}$, stand for the first step estimators. We denote the estimator of the unknown error term V_i as $\hat{V}_i = X_i - \hat{\alpha} - Z_i^T\hat{\pi}$.

Remark 2.3.5 (Identification) *Before delving into the details of estimating, we give a remark on the identification issue, which is essential for understanding the proposed modal estimator and deriving the limiting distribution theory. In modal regression models, a necessary (but not sufficient) condition for identification is that the number of population orthogonality conditions is at least as large as the number of model parameters, i.e.,*

$$E\left[\frac{1}{h^3}Z_i\phi\left(\frac{X_i-\alpha-Z_i^T\pi}{h}\right)\left(X_i-\alpha-Z_i^T\pi\right)|_{\alpha=\alpha_0,\pi=\pi_0}\right]=0,$$

where α_0 and π_0 are the true values of the parameters. Then, α_0 and π_0 are locally identified if there exists a neighborhood of α_0 and π_0 within which only α_0 and π_0 satisfy (2.5)*. If there are multiple solutions to these moment conditions, the parameters cannot be identified without other restrictions. Furthermore, consistent with the sufficient condition in Chen et al. (2014) for local identification in parametric models, we can achieve local identification for modal regression if the partial derivative matrix of the left-hand side of the above*

[8] We further illustrate the convenience of using normal kernel function here. Taking the first derivative of the aforementioned kernel-based objective function with respect to π, we can obtain

$$\frac{1}{nh^3}\sum_{i=1}^{n}Z_i\phi\left(\frac{X_i-\alpha-Z_i^T\pi}{h}\right)\left(X_i-\alpha-Z_i^T\pi\right)|_{\alpha=\hat{\alpha},\pi=\hat{\pi}}=0.$$

The above equation can be represented by the population version in Remark 2.3.5, which is the moment condition in standard modal regression estimation. It is clear that the maximization of the modal regression objective function is essentially a weighted least square problem. We can then use an iterative procedure to obtain estimates, which builds the underlying mechanism of the MEM algorithm.

moment condition with respect to the α and π parameters has full rank. The estimation steps/procedures listed below are all subjected to the same set of arguments.

The **second step** is the estimation of semiparametric partially linear modal regression of Y_i on X_i, $Z_{1,i}$ and $\hat{V}_i$ by imposing the conditional mode independence restriction and treating $\hat{V}$ as a nuisance parameter. The systematic investigation of this model is novel in modal regression, which is not explicitly available in the literature. Thus, there is an independent interest in establishing estimation procedure and asymptotic theorems. Because there exist both parametric and nonparametric components simultaneously in the model, which should be estimated with *modal* parametric and nonparametric rates of convergence, respectively, we propose the following three-stage estimation method to achieve the optimal convergence rates under the assumption that the unknown function $m(\cdot)$ has a continuous second derivative.

In the **first stage**, we apply local linear technique to approximate $m(\hat{V}_i)$ for $\hat{V}_i$ in the neighborhood of v by substituting V_i with the residual from the first step and assuming continuity and differentiability of the unknown function $m(\cdot)$

$$m(\hat{V}_i) \approx m(v) + m^{(1)}(v)(\hat{V}_i - v) \equiv \alpha_1 + \alpha_2(\hat{V}_i - v),$$

in which $m^{(1)}(v)$ indicates the first derivative of $m(v)$, "$\approx$" denotes the approximation by ignoring higher orders, "$\equiv$" means "is defined as", $\alpha_1 = m(v)$, and $\alpha_2 = m^{(1)}(v)$. We let $\{\tilde{\beta}, \tilde{\gamma}, \tilde{\alpha}_1, \tilde{\alpha}_2\}$ be the maximizers of the following local kernel-based objective function

$$Q_n(\beta, \gamma, \alpha_1, \alpha_2) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} \phi \left(\frac{Y_i - X_i\beta - Z_{1,i}^T\gamma - \alpha_1 - \alpha_2(\hat{V}_i - v)}{h_1} \right) K \left(\frac{\hat{V}_i - v}{h_2} \right), \tag{2.6}$$

where $K(\cdot) : R \to R$ is a nonnegatively symmetric kernel, $h_1 = h_1(n) \to 0$ and $h_2 = h_2(n) \to 0$ are two sequences of bandwidths tending to zero as n increases, and h_2 controls the degree of smoothing as usual in nonparametric regression. For simplicity of calculation, we also choose $K(\cdot)$ as a normal kernel in numerical analysis. We then have the initial estimators of the parameters denoted as $\{\tilde{\beta}, \tilde{\gamma}, \tilde{m}(v), \tilde{m}^{(1)}(v)\}$.

Because we only use data in a local neighborhood of v to estimate the global parameters β and γ, the initial estimators $\tilde{\beta}$ and $\tilde{\gamma}$ do not have global convergence rates, as in conventional semiparametric mean regression. Treating $\tilde{m}(\cdot)$ and $\hat{V}_i$ as nuisance parameters, we can further improve the convergence rates of the parametric component estimators using all data in the **second stage** by plugging the initial estimator $\tilde{m}(\hat{V}_i)$ from the first step and maximizing the following global kernel-based objective function

$$Q_n(\beta, \gamma) = \frac{1}{nh_3} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \tilde{m}(\hat{V}_i) - X_i\beta - Z_{1,i}^T\gamma}{h_3} \right), \tag{2.7}$$

where $h_3 = h_3(n)$ is a sequence of bandwidths that tend to zero as the sample size n approaches infinity. We denote the estimators of β and γ from (2.7) as $\hat{\beta}$ and $\hat{\gamma}$.

Remark 2.3.6 (Global Mean) *In addition to the kernel-based objective function, we can alternatively apply the following global mean method to obtain the final estimators of β and γ in the second stage by taking advantage of the full sample information*

$$\hat{\beta} = Mean(\tilde{\beta}(\hat{V}_i)) = \int \tilde{\beta}(\hat{V}_i) dW(v), \text{ and } \hat{\gamma} = Mean(\tilde{\gamma}(\hat{V}_i)) = \int \tilde{\gamma}(\hat{V}_i) dW(v),$$

where $W(\cdot)$ is a deterministic weighting function with $\int dW(v) = 1$. We can then follow the same proving procedures as in this paper to demonstrate that the estimators are $\sqrt{nh_1^3}$-

consistent and asymptotically normal under certain regularity conditions. The mechanism of such a mean method is comparable to that of average marginal effect or derivative estimation.

In the **last stage**, we improve the efficiency of the estimator of the nonparametric part by plugging in the previous parametric estimators. Since the parametric component is estimated with a modal parametric convergence rate in the second stage, which is faster than the fastest possible rate of convergence for the modal nonparametric component, it is feasible to estimate the nonparametric part as asymptotically efficiently as if the parametric part were known. We thereupon maximize the following local kernel-based objective function in the same way that we do in fully nonparametric modal regression

$$Q_n(\alpha_1, \alpha_2) = \frac{1}{nh_4h_5} \sum_{i=1}^{n} \phi \left(\frac{Y_i - X_i\hat{\beta} - Z_i^T\hat{\gamma} - \alpha_1 - \alpha_2(\hat{V}_i - v)}{h_4} \right) K \left(\frac{\hat{V}_i - v}{h_5} \right), \qquad (2.8)$$

where $h_4 = h_4(n)$ and $h_5 = h_5(n)$ are two sequences of positive numbers tending to zero, named bandwidths, as $n \to \infty$. Consistent with nonparametric estimation, the bandwidth h_5 controls the smoothness of the estimated function. We then have the final estimators denoted as $\{\hat{m}(v), \hat{m}^{(1)}(v)\}$, which are expected to be more efficient than the initial estimators since we do not need to account for the uncertainty of estimating the parametric component.

Remark 2.3.7 (B-Splines) *We adopt the local linear approximation method because of its attractive properties, such as high statistical efficiency, automatic boundary effect corrections, and design adaptation (Fan and Gijbels, 1996). However, in addition to the proposed two-step estimation procedure, popular spline methods* (*such as B-spline*) *with good approximating properties can also be applied to estimate the modal coefficients. Espe-*

cially, let $B(\hat{V}) = (B_1(\hat{V}), \cdots, B_q(\hat{V}))^T$ *denote B-spline basis functions of order* l*, where* $q = N + l + 1$ *and* N *is the number of interior knots. Then,* $m(\hat{V})$ *can be approximated by* $m(\hat{V}) \approx B(\hat{V})^T\lambda$*, where* λ *is a* $q \times 1$ *vector of unknown parameters. This implies that we are able to obtain the estimators by maximizing the global kernel-based objective function shown below*

$$\frac{1}{nh_b}\sum_{i=1}^{n}\phi\left(\frac{Y_i - X_i\beta - Z_{1,i}^T\gamma - B(\hat{V})^T\lambda}{h_b}\right),$$

where $h_b = h_b(n) \to 0$ *is a bandwidth that depends on* n*. Thus, we can avoid estimating non-parametric modal regression with the use of B-splines. Following the same proof procedures for Theorem 2.4.6, it can be shown that the estimators of* β*,* γ*, and* λ *are* $\sqrt{nh_b^3}$*-consistent and asymptotically normal under mild conditions.*

The proposed modal estimators do not admit analytic expressions and require feasible implementation through a numerical algorithm. To obtain the numerical solutions of the aforementioned estimators, we develop a modified MEM Algorithm 1 based on Yao et al. (2012) and Yao (2013) that decomposes the optimization into E-Step and M-Step. We primarily show the algorithms for (2.5) and (2.6), whereas (2.7) and (2.8) can be quantitatively solved using the appropriate MEM algorithm with minor modifications as well. It is worth noting that maximizing the kernel-based objective function is equivalent to maximizing the log value of the related function. The monotone ascending property of the proposed MEM algorithm can then be established along the lines of the classical EM algorithm by applying Jensen's inequality, which guarantees the convergence and stability of the algorithm. What is more, owing to the use of shrinking bandwidths, the modified MEM algorithm may be

Algorithm 1 MEM Algorithm

Equation (2.5)

E-Step. Calculate the weight $w\left(i \mid \alpha^{(g)}, \pi^{(g)}\right), i = 1, \cdots, n$ with the preliminary estimates of the modal parameters as

$$w\left(i \mid \alpha^{(g)}, \pi^{(g)}\right) = \frac{\phi\left(\frac{X_i - \alpha^{(g)} - Z_i^T \pi^{(g)}}{h}\right)}{\sum_{i=1}^{n} \phi\left(\frac{X_i - \alpha^{(g)} - Z_i^T \pi^{(g)}}{h}\right)}.$$

M-Step. Update $\left(\alpha^{(g+1)}, \pi^{(g+1)}\right)$ with the weight calculated in the E-Step

$$\begin{aligned}\left(\alpha^{(g+1)}, \pi^{(g+1)}\right) &= \arg\max_{\alpha,\pi} \sum_{i=1}^{n} \left\{ w\left(i \mid \alpha^{(g)}, \pi^{(g)}\right) \log \frac{1}{h} \phi\left(\frac{X_i - \alpha - Z_i^T \pi}{h}\right) \right\} \\ &= (Z^{*T} W_Z Z^*)^{-1} Z^{*T} W_Z X,\end{aligned}$$

where g is the iteration indicator, $Z^* = (Z_1^*, \cdots, Z_n^*)^T$ with $Z_i^* = (1 \;\; Z_i^T)^T$, $X = (X_1, \cdots, X_n)^T$, and W_Z is an $n \times n$ diagonal matrix with diagonal elements $\{w\left(i \mid \alpha^{(g)}, \pi^{(g)}\right)\}_{i=1}^n$.

Iterate. Given the initial values (e.g., mean regression estimates), iterate E-Step and M-Step repeatedly until a stopping criteria is satisfied, i.e., $\|\pi^{(g+1)} - \pi^{(g)}\| < 10^{-5}$.

Equation (2.6)

E-Step. Define $\kappa = (\beta, \gamma, \alpha_1, \alpha_2)$. Calculate the weight $w\left(i \mid \kappa^{(g)}\right), i = 1, \cdots, n$ with the preliminary estimates of the modal parameters as

$$w\left(i \mid \kappa^{(g)}\right) = \frac{\phi\left(\frac{Y_i - X_i\beta^{(g)} - Z_{1,i}^T \gamma^{(g)} - \alpha_1^{(g)} - \alpha_2^{(g)}(\hat{V}_i - v)}{h_1}\right) K\left(\frac{\hat{V}_i - v}{h_2}\right)}{\sum_{i=1}^{n} \phi\left(\frac{Y_i - X_i\beta^{(g)} - Z_{1,i}^T \gamma^{(g)} - \alpha_1^{(g)} - \alpha_2^{(g)}(\hat{V}_i - v)}{h_1}\right) K\left(\frac{\hat{V}_i - v}{h_2}\right)}.$$

M-Step. Update $\kappa^{(g+1)}$ with the weight calculated in the E-Step

$$\kappa^{(g+1)} = \arg\max_{\kappa} \sum_{i=1}^{n} \left\{ w\left(i \mid \kappa^{(g)}\right) \log \frac{1}{h_1} \phi\left(\frac{Y_i - X_i\beta - Z_{1,i}^T \gamma - \alpha_1 - \alpha_2(\hat{V}_i - v)}{h_1}\right) \right\}$$

Algorithm 1 MEM Algorithm

$$= (X^{*T} W_X X^*)^{-1} X^{*T} W_X Y,$$

where $X^* = (X_1^*, \cdots, X_n^*)^T$ with $X_i^* = (X_i, Z_{1,i}^T, 1, \hat{V}_i - v)$, $Y = (Y_1, \cdots, Y_n)^T$, and W_X is an $n \times n$ diagonal matrix with diagonal elements $\{w(i \mid \kappa^{(g)})\}_{i=1}^n$.

Iterate. Given the initial values (e.g., local linear mean regression estimates), iterate E-Step and M-Step repeatedly until a stopping criteria is satisfied, i.e., $\|\kappa^{(g+1)} - \kappa^{(g)}\| < 10^{-5}$.

stuck at the local optima. To avoid the potential local maximum and achieve the global favorable one, it is advisable to try different initial values in practice to select the best optimal estimate by comparing the values of the target function (Ullah et al., 2021, 2022).[9]

Remark 2.3.8 (Modal-Based Robust Control Function) *We in this paper investigate the newly proposed modal regressions utilizing kernel-based objective functions augmented with shrinking bandwidths, where we assume that the error distribution is skewed to enable the mode to differ from the mean. It is observed that with the focus on the mean regression version of* (2.1) *such that* $E(V_i \mid Z_i) = 0$ *and* $E(U_i \mid X_i, Z_i) \neq 0$*, the proposed estimation procedure in this paper can still be applied to account for outliers or aberrant observations but with the conditions that bandwidths* h*,* h_1*, and* h_3 *are treated as constants* (*do not depend on sample size* n *and can determine the degree of robustness and efficiency*)*; see Yao et al. (2012). Under suitable conditions, we can obtain more robust and efficient estimators compared to mean estimators by choosing appropriate tuning parameters when the data*

[9]When the MEM algorithm has been trapped in a local optimal area, a rather naive strategy would be to keep iterating in the expectation that the algorithm will eventually locate the global maximum after a large number of iterations. However, it is preferable to start with estimates obtained by other estimation techniques, such as mean estimation, quantile estimation, or any other robust estimations.

have a heavy-tailed error distribution or outliers. In addition, modal-based estimation will be as asymptotically efficient as mean estimation when there are no outliers or the error is normally distributed.[10]

The fundamental principle for modal-based robust control function estimation is that given symmetric data, the modal regression line is identical to the mean regression line

$$\begin{cases} Mode(Y_i \mid X_i, Z_i, V_i) = E(Y_i \mid X_i, Z_i, V_i) \\ \qquad\qquad\qquad\qquad = X_i\beta + Z_{1,i}^T\gamma + E(U_i \mid V_i) \\ \qquad\qquad\qquad\qquad = X_i\beta + Z_{1,i}^T\gamma + Mode(U_i \mid V_i), \\ Mode(X_i \mid Z_i) = E(X_i \mid Z_i) = \alpha + Z_i^T\pi, \end{cases}$$

where the estimation procedure and kernel-based objective functions are exactly the same as those previously described. Particularly, the estimators are obtained by maximizing the corresponding kernel-based objective functions, with the bandwidths related to error terms treated as constants. Such a modal-based estimation is capable of dealing with data containing outliers or heavy-tailed distributions under the content of the control function, which has not previously been investigated in the literature. The convergence rates and asymptotic normality of the proposed modal-based robust estimators for both the parametric and nonparametric parts are established without assuming any parametric form on the error distribution. The asymptotic results are shown to be completely different from those of modal estimators, and the convergence rates are the same as those of mean regression; see Remarks 2.4.11, 2.4.12, and 2.4.15 for theoretical results and Appendix A for simulation results.

[10] If the noise follows a normal distribution, the least square estimator will be the most efficient estimator of the regression coefficient. However, when the errors are heavy-tailed or contain outliers, the efficiency of the least square estimator will be severely reduced. Although M-estimation and quantile estimation can be applied to achieve robustness, they will lose efficiency when the data are from a normal distribution.

2.4 Asymptotic Properties

We in this section provide a full characterization of the asymptotic behavior of the proposed estimators in different steps/stages. To begin, we shall list certain notations that will be utilized throughout the rest of the section. We define $m^{(2)}(v) = \partial^2 [m(v)] / \partial v^2$, $\mu_j = \int w^j K(w) \mathrm{d}w$, $v_j = \int w^j K^2(w) \mathrm{d}w$ for $j = 0, 1, 2, 3$, and use $g^{(c)}(\cdot)$ to denote the cth derivative of density function $g(\cdot)$. For any vector or matrix A, let $\|A\| = [\mathrm{trace}(A^T A)]^{1/2}$ be the Euclidean norm. We call that $T_n(x) = T(x) + o_p(s_n)$ (or $O_p(s_n)$) uniformly for $x \in \mathcal{X}$ if $\sup_{x \in \mathcal{X}} |T_n(x) - T(x)| = o_p(s_n)$ (or $O_p(s_n)$), and use "$\xrightarrow{d}$" and "$\xrightarrow{p}$" to represent convergence in distribution and probability, respectively. We define a function $f(n) = O(1)$ if there exist some non-zero constants c and N such that $f(n)/c \to 1$ for $n \geq N$. To facilitate the derive of the consistency and asymptotic theorems for the proposed modal estimators in a general framework, we impose the following regularity conditions.

C1 (Parameter Space) The true values of parameters α_0, π_0, β_0, and γ_0 are in the interior of the known compact parameter space, which is a subset of $R^1 \times R^{d_Z} \times R^1 \times R^{d_{Z_1}}$.

C2 (Identification) The dimension of Z (d_Z) is larger than that of Z_1 (d_{Z_1}).

C3 (Smoothness) Define $S(V)$ as a function space in which $m(\cdot) \in S(V)$ if $m : V \to R$, then $m(\cdot)$ has at least a continuous second derivative on an open set containing the point v.

C4 (Kernel Function) The kernel functions $\phi(\cdot) : R \to R$ and $K(\cdot) : R \to R$ are non-negatively symmetric density functions with bounded support and integrate to one. Furthermore, $\int t^2 K(t) dt < \infty$ and $\int t^2 K^2(t) dt < \infty$.

C5 (Conditional Density I) The conditional density function of V given Z denoted by $g_V(\cdot \mid Z) : R \times R^{d_Z} \to R$ is greater than zero and continuous at V for all V and Z. Furthermore, $g_V(\cdot \mid Z)$ is assumed to have the fourth continuous derivative and global mode at zero, i.e., $g_V(\cdot \mid Z) < g_V(0 \mid Z)$ for all $V \neq 0$ and Z. Also, $0 < \inf g_V(\cdot \mid Z) \leq \sup g_V(\cdot \mid Z) < \infty$.

C6 (Conditional Density II) For a fixed point v, $f_V(v) > 0$, where $f_V(\cdot)$ is the marginal density of V that is continuous at v, and $g_\epsilon(\cdot \mid X, Z, V) > 0$ is continuous at ϵ for all X, Z, V, where $\epsilon = Y - (X\beta + Z_1^T\gamma + m(V))$ and $g_\epsilon(\cdot \mid X, Z, V)$ is the conditional density function of ϵ. Furthermore, $g_\epsilon(\cdot \mid X, Z, V)$ is assumed to have the fourth continuous derivative and global mode at zero, i.e., $g_\epsilon(\cdot \mid X, Z, V) < g_\epsilon(0 \mid X, Z, V)$ for all $\epsilon \neq 0$ and X, Z, V. Both $f_V(v)$ and $g_\epsilon(\cdot \mid X, Z, V)$ are bounded away from infinity.

C7 (Moment) There exists a constant $s > 2$ such that $E(\|Z\|^{2s}) < \infty$ and $E(|X|^{2s}) < \infty$. The matrices J, Γ, J_X, and Γ_2 defined in the following theorems are negative definite.

The conditions listed above are relatively mild and can be satisfied in a variety of practical situations; see the similar conditions used in Kemp and Santos Silva (2012), Yao and Li (2014), and Ullah et al. (2021, 2022). C1 is an ordinary regularity condition on parameters that is generally easy to verify. Moreover, compactness is not restrictive in microeconometric applications and can be relaxed at lengthy arguments. C2 is the necessary condition for identification, which is the same as the one used in Li et al. (2007). It states that Z and Z_1 may share some common components, but at least one non-overlapping component must be present. C3 is a commonly used condition on the smoothness of the unknown function in local linear estimation. It controls the precision in the approximation

as the second derivative of $m(\cdot)$ impacts the bias asymptotically. It has been observed that a higher-order bias can be achieved if we impose more restrictive conditions on the smoothness of the function $m(\cdot)$. The bounded support in C4 imposed on kernel functions is for the brevity of proofs and may be eased with certain integrability restrictions on the tail of the kernel functions; for example, the normal kernel function is allowed. In this paper, we choose the normal kernel function for $\phi(\cdot)$ in theoretical analysis. Thus, we do not include all of the regularity conditions for the general kennel functions here. The details of the kernel conditions can be found in Kemp and Santos Silva (2012). C5 and C6 imply certain smoothness of distributions in the neighborhood of zero, which is necessary for identification. More specifically, C5 is required for the first step estimation, while C6 is the regularity condition for the second step estimation. Both of them imply that the conditional density of the error term has a well defined global mode at zero. It is to be conceded that this assumption is utilized for illustrative purposes. When the population is not homogeneous (i.e., clustered/inhomogeneous data), the proposed estimation procedure can also be applied in a multimode environment, where the various modal solutions will be found by starting from multiple initial values. C7 is the classic rank condition for ensuring the existence of the asymptotic mean and variance for the proposed modal estimators by placing restrictions on the moments of covariates. Unlike modal-based estimation in Remark 2.3.8, we do not need to impose any moments on the error terms for modal regression.

Remark 2.4.9 *Bandwidths are critical parameters in the proposed estimation procedure for reducing the bias from the previous step/stages to a sufficiently small order, so that the bias can be neglected asymptomatically and the oracle property is achieved. This under-*

smoothing (goes to zero faster relative to the usual optimal bandwidth choice) is standard in the semiparametric literature when the first stage estimates are employed in a second stage parametric or nonparametric estimation. All conditions related to bandwidth sequences in different steps/stages are specified for each of the theorems stated below. It is worth noting that the i.i.d. assumption for the data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ *in this paper is shared with many prior analyses on control function literature. Nevertheless, it is inessential and can be relaxed to allow for some forms of stationary time series dependence, such as* α*-mixing in Su and Ullah (2008), without affecting asymptotic results by imposing some restricted conditions.*

We then discuss the asymptotic properties of the proposed estimation procedure. The main results are to show that the estimators in different steps/stages are asymptotically equivalent to infeasible estimators (without knowing the true values of some components). As a result, the proposed procedure resembles many other kernel-based multi-stage nonparametric procedures in that the first stage estimators do not contribute to the asymptotic property of the final stage estimators. The limiting distributions for the first step estimators are illustrated as follows.

Theorem 2.4.1 *Under the regularity conditions C1-C5 and C7, with probability approaching one, as* $n \to \infty, h \to 0$*, and* $nh^5 \to \infty$*, there exists a consistent maximizer* $\hat{\theta} = (\hat{\alpha}, \hat{\pi}^T)^T$ *of* (2.5) *such that*

$$\|\hat{\theta} - \theta_0\| = O_p\left(\left(nh^3\right)^{-1/2} + h^2\right),$$

where $\theta_0 = (\alpha_0, \pi_0^T)^T$ *represents the true parameter vector.*

Theorem 2.4.2 *With* $nh^7 = O(1)$ *and* $Z^* = [1\ Z^T]^T$*, under the same conditions as Theorem 2.4.1, the estimator satisfying the consistency result in Theorem 2.4.1 has the following asymptotic result*

$$\sqrt{nh^3}\left(\hat{\theta}-\theta_0-\frac{h^2}{2}J^{-1}M(1+o_p(1))\right)\xrightarrow{d}\mathcal{N}\left(0,\int t^2\phi^2(t)dtJ^{-1}LJ^{-1}\right).$$

Furthermore, with the assumption that $nh^7\to 0$*, we have*

$$\sqrt{nh^3}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,\int t^2\phi^2(t)dtJ^{-1}LJ^{-1}\right),$$

where $J=E(Z^*Z^{*T}g_V^{(2)}(0\mid Z))$*,* $L=E(Z^*Z^{*T}g_V(0\mid Z))$*, and* $M=E(Z^*g_V^{(3)}(0\mid Z))$*.*

The proofs of Theorems 2.4.1 and 2.4.2 are similar to those of Yao and Li (2014), which investigated linear modal regression, so we omit them for brevity. Theorems 2.4.1 shows that the convergence rate of the first step estimators can be divided into two components. The first component $\left(nh^3\right)^{-1/2}$ reflects the convergence rate of the variance term, while the second component h^2 represents the convergence rate of the bias term. In comparison to mean estimation, modal estimation introduces an additional bias component due to the capture of mode through kernel estimating. Theorem 2.4.2 indicates that the mean squared error (MSE) optimal smoothing h is proportional to $n^{-1/7}$, and that there is a trade-off between the convergence rate and the asymptotic bias in the asymptotic normality of modal estimators.[11] By undersmoothing ($lim_{n\to\infty}\sqrt{nh^7}\to 0$), we can successfully eliminate the asymptotic bias at the expense of a slower convergence speed. With the MSE-optimal bandwidth, the modal estimators in the first step have a limiting non-centered normal distribution with the convergence rate $n^{-1/4}$, which is slower than that of mean estimators due to the usage of a small portion of total observations around mode (controlled by bandwidth h). Such a slower convergence rate is the price we need to pay for estimating the mode.

[11]The exact expression of the asymptotically optimal bandwidth by minimizing MSE is $\hat{h}_{MSE}=[3v_2tr(J^{-1}LJ^{-1})/(M^TJ^{-1}J^{-1}M)]^{1/7}n^{-1/7}$. However, such an expression is less useful in practice since it is tedious and difficult to estimate components in the expression to achieve the estimate of bandwidth.

Remark 2.4.10 *In order to conduct large sample statistical inference, it is necessary to have consistent estimators for the asymptotic variance-covariance components* $J^{-1}L(J^{-1})^T$*, which can be estimated by a kernel estimator, denoted as* $\hat{J}^{-1}\hat{L}(\hat{J}^{-1})^T$*. By undersmoothing, the approximate* $(1-\alpha)$ *confidence interval for* θ_0 *can be obtained as*

$$\left\{\hat{\theta}-t_{\frac{\alpha}{2}}\left((nh^3)^{-1}\int t^2\phi^2(t)dt\hat{J}^{-1}\hat{L}(\hat{J}^{-1})^T\right)^{1/2},\hat{\theta}+t_{\frac{\alpha}{2}}\left((nh^3)^{-1}\int t^2\phi^2(t)dt\hat{J}^{-1}\hat{L}(\hat{J}^{-1})^T\right)^{1/2}\right\},$$

where $t_{\alpha/2}$ *denotes the upper* $\alpha/2$ *quantile of the standard normal distribution. However, this is not particularly convenient in practice owing to the complex structure of the components and the requirement to nonparametrically estimate conditional densities with additional smoothing parameters. We practically can instead apply the bootstrap method to draw* S *sets of* n *observations with replacement from* $\{(X_i,Z_i)\}_{i=1}^n$*, say* $\{(X_{si},Z_{si})\}_{i=1}^n$*, to obtain a bootstrapped estimator* $\hat{\theta}^*$ *and carry out analytical inference subsequently. This type of bootstrap procedure is generally consistent. Under regularity conditions, the asymptotic distribution of* $\sqrt{nh^3}[\hat{\theta}-\theta_0]$ *can be approximated by the limiting distribution of* $\sqrt{nh^3}[\hat{\theta}^*-\theta_0]$*; see Zhang et al. (2020).*

Remark 2.4.11 (Modal-Based Robust Estimator (First Step)) *We assume that bandwidth* h *is a constant number/tuning parameter that is independent of sample size* n *and that the error term* V_i *is symmetrically distributed. With the additional assumptions that* $E\{\phi_h^{(1)}(V)\}=0$*,* $E(\phi_h^{(1)}(V)^2)$ *is finite for any* $h>0$*, and there exists a constant* $C>0$ *such that* $E\{\sup_{X:|X-V|<C}|\phi_h^{(3)}(V)|\}<\infty$ *(used to regulate the magnitude of the remainder in the Taylor expansion), the asymptotic theorem for the modal-based robust estimator* $\hat{\theta}_{robust}$ *will be*[12]

[12]With constant bandwidths, the kernel-based objective functions are special M-type robust regressions.

$$\sqrt{n}\left(\hat{\theta}_{robust}-\theta_0\right)\xrightarrow{d}N\left\{0,\left(E(\phi_h^{(2)}(V))\right)^{-2}E(\phi_h^{(1)}(V))^2\left\{Cov(Z^*)\right\}^{-1}\right\},$$

where $\hat{\theta}_{robust}$ *is the local maximizer of* (2.5)*,* $\phi_h(\cdot)=\phi(\cdot/h)/h$*, and* $\phi_h^{(c)}(\cdot)$ *is the cth derivative of* $\phi_h(\cdot)$*. Thus, the asymptotic variance of* $\hat{\theta}_{robust}$ *depends on the tuning parameter* h*. The performance of the modal-based robust estimator can then be better than or at least as good as the least square estimator by appropriately choosing* h*. The asymptotic relative efficiency of the modal-based robust estimator over the least square estimator is* $Var(V)[E(\phi_h^{(2)}(V))]^2E[(\phi_h^{(1)}(V))^2]^{-1}$*. The optimal tuning parameter should be chosen as* $h_{opt}=\arg\max_h[E(\phi_h^{(2)}(V))]^2E[(\phi_h^{(1)}(V)^2)]^{-1}$*, which is solely dependent on the derivatives of* $\phi(\cdot)$*. In practice, we can combine with the grid search method to select tuning parameter by numerically calculating the components in the aforementioned asymptotic relative efficiency expression.*

To appreciate the effect of the first step estimation on the second step, we rewrite (2.6) as the following equation

$$\frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{Y_i-X_i\beta-Z_{1,i}^T\gamma-m(V_i)-(m(\hat{V}_i)-m(V_i))}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right),\tag{2.9}$$

where $m(\hat{V}_i)-m(V_i)=m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)$, $\bar{V}_i$ is between $\hat{V}_i$ and V_i according to the first-order Taylor expansion, and $m^{(1)}(\bar{V}_i)$ is the first derivative of $m(\cdot)$ with regard to V_i evaluated

The proof of the asymptotic result could be as simple as in M-type regression. Particularly, we can prove the $\sqrt{n}$-consistency result by following the procedures for proving Theorem 2.4.5. After that, we have

$$\begin{aligned}0&=\frac{1}{n}\sum_{i=1}^{n}Z_i^*\phi_h^{(1)}\left(V_i-Z_i^{*T}(\hat{\theta}_{robust}-\theta_0)\right)\\&=\frac{1}{n}\sum_{i=1}^{n}Z_i^*\left\{\phi_h^{(1)}(V_i)-\phi_h^{(2)}(V_i)Z_i^{*T}(\hat{\theta}_{robust}-\theta_0)+\frac{1}{2}\phi_h^{(3)}(V_i)\left(Z_i^{*T}(\hat{\theta}_{robust}-\theta_0)\right)^2+o_p(1)\right\},\end{aligned}$$

where the third component is also $o_p(1)$. Since the bandwidth h is a constant, we can derive the theorem directly from the central limit theorem and Slutsky's theorem.

at $\bar{V}_i$. Even though the associated asymptotic results are available in the literature of both linear and nonparametric modal regressions, none of them are directly applicable here because we need to take into account the additional bias factor introduced by the previous step, which is shown in (2.9). However, given fairly mild bandwidth conditions, it can be demonstrated that the bias from the first step is asymptotically disregarded and does not affect the convergence rate of the second step estimators. The following theorems establish that the proposed first stage estimators converge at the optimal local linear modal regression rate and have a limiting non-centered normal distribution.

Theorem 2.4.3 *Under the regularity conditions C1-C7, with probability approaching one, as* $n \to \infty, h/h_2 \to 0, h_1 \to 0, h_2 \to 0$, $h_2^2/h_1 \to 0$, *and* $nh_2h_1^5 \to \infty$, *there exist consistent maximizers* $(\tilde{\eta}, \tilde{m}(v), h_2\tilde{m}^{(1)}(v))$ *of* (2.6) *such that*

i. $|\tilde{m}(v) - m(v)| = O_p\left(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2\right)$,

ii. $|h_2(\tilde{m}^{(1)}(v) - m^{(1)}(v))| = O_p\left(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2\right)$,

iii. $\|\tilde{\eta} - \eta_0\| = O_p\left(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2\right)$,

where $\tilde{\eta} = (\tilde{\beta}, \tilde{\gamma}^T)^T$, *and* $\eta_0 = (\beta_0, \gamma_0^T)^T$ *is the true parameter vector.*

Theorem 2.4.4 *With* $nh_2^5h_1^3 = O(1)$, $nh_2h_1^7 = O(1)$, *and* $Z_X = [X\ Z_1^T]^T$, *under the same conditions as Theorem 2.4.3, the estimators satisfying the consistency results in Theorem 2.4.3 have the following asymptotic result*

$$\sqrt{nh_2h_1^3}\Bigg[\begin{pmatrix} \tilde{m}(v) - m(v) \\ h_2(\tilde{m}^{(1)}(v) - m^{(1)}(v)) \\ \tilde{\eta} - \eta_0 \end{pmatrix} - \Gamma^{-1}\left(\frac{h_2^2}{2}m^{(2)}(v)E\begin{pmatrix} \mu_2 \\ \mu_3 \\ \mu_2 Z_X \end{pmatrix} - \frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0 \mid X, Z, V = v)}{g_\epsilon^{(2)}(0 \mid X, Z, V = v)}\right.$$

$$E\left(\begin{array}{c}\mu_0\\ \mu_1\\ \mu_0 Z_X\end{array}\right)(1+o_p(1))\Bigg] \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid X, Z, V=v)\int t^2\phi^2(t)dt}{f_V(v)g_\epsilon^{(2)}(0 \mid X, Z, V=v)^2}\Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

If we allow $nh_2^5h_1^3 \to 0$ *and* $nh_2h_1^7 \to 0$*, the asymptotic theorem becomes*

$$\sqrt{nh_2h_1^3}\left(\begin{array}{c}\tilde{m}(v)-m(v)\\ h_2(\tilde{m}^{(1)}(v)-m^{(1)}(v))\\ \tilde{\eta}-\eta_0\end{array}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid X, Z, V=v)\int t^2\phi^2(t)dt}{f_V(v)g_\epsilon^{(2)}(0 \mid X, Z, V=v)^2}\Gamma^{-1}\Sigma\Gamma^{-1}\right),$$

where $\Gamma = E\begin{pmatrix}\mu_0 & \mu_1 & \mu_0 Z_X^T\\ \mu_1 & \mu_2 & \mu_1 Z_X^T\\ \mu_0 Z_X & \mu_1 Z_X & \mu_0 Z_X Z_X^T\end{pmatrix}$ *and* $\Sigma = E\begin{pmatrix}v_0 & v_1 & v_0 Z_X^T\\ v_1 & v_2 & v_1 Z_X^T\\ v_0 Z_X & v_1 Z_X & v_0 Z_X Z_X^T\end{pmatrix}$.

We have a sort of oracle property here in that the asymptotic properties of the first stage estimators are the same as in the feasible case where the residuals $\{V_i\}_{i=1}^n$ are observed. The consistency of the first step estimators in Theorem 2.4.1 implies that the estimated residuals satisfy $|\hat{V}_i - V_i| = O_p((nh^3)^{-1/2} + h^2) = O_p(n^{-2/7})$, which actually converges to zero faster than $h_1^2 = O(n^{-2/8})$ and $h_2^2 = O(n^{-2/8})$, and is the underlying reason why the effect of $\hat{V}_i$ on the asymptotic results for the first stage estimators is negligible asymptotically. These observations indicate that the condition $h/h_2 \to 0$ is reasonable considering from the perspective of MSE-optimal bandwidth rates. A similar oracle property has been observed in the fixed effects modal regression for panel data investigated by Ullah et al. (2021), where they proposed a pseudo-demodeing two-step method for estimating modal coefficients. Compared to the linear modal estimators in the first step, as we only use data in the neighborhood of v, the estimator $\tilde{\eta}$ is $\sqrt{nh_2h_1^3}$-consistent with an extra

bias caused by the estimation of nonparametric component. It is noticed that the bias will converge to zero with undersmoothing ($lim_{n\to\infty}\sqrt{nh_2h_1^7} \to 0$ and $lim_{n\to\infty}\sqrt{nh_2^5h_1^3} \to 0$).

Remark 2.4.12 (Modal-Based Robust Estimators (First Stage)) *Similar to the results in Remark 2.4.11, we can build the asymptotic theorem for the modal-based robust estimators with the assumption that the data are symmetrically distributed. To achieve robustness and efficiency, we treat bandwidth h_1 as a tuning parameter (constant). Then, if $h_2 \to 0$ as $n \to \infty$ ($nh_2 \to \infty$), with conditions C1-C4 and the additional assumptions that $E(\phi_{h_1}^{(1)}(\epsilon) \mid X, Z, V) = 0$ and $E(\phi_{h_1}^{(2)}(\epsilon)^2 \mid X, Z, V)$, $E(\phi_{h_1}^{(1)}(\epsilon)^3 \mid X, Z, V)$, and $E(\phi_{h_1}^{(3)}(\epsilon) \mid X, Z, V)$ are continuous with respect to (X, Z, V), we have*[13]

$$\sqrt{nh_2}\left[\begin{pmatrix} \tilde{m}_{robust}(v) - m(v) \\ h_2(\tilde{m}^{(1)}_{robust}(v) - m^{(1)}(v)) \\ \tilde{\eta}_{robust} - \eta_0 \end{pmatrix} - \Gamma^{-1}\frac{h_2^2}{2}m^{(2)}(v)E\begin{pmatrix} \mu_2 \\ \mu_3 \\ \mu_2 Z_X \end{pmatrix}\right]$$
$$\xrightarrow{d} \mathcal{N}\left(0, \left\{G(h_1)/F^2(h_1)\right\}\Gamma^{-1}\Sigma\Gamma^{-1}\right),$$

where $\tilde{m}_{robust}(v)$, $\tilde{m}^{(1)}_{robust}(v)$, and $\tilde{\eta}_{robust}$ are the modal-based robust estimators from (2.6), $F(h_1) = E(\phi_{h_1}^{(2)}(\epsilon) \mid X, Z, V = v)$, and $G(h_1) = E(\phi_{h_1}^{(1)}(\epsilon)^2 \mid X, Z, V = v)$. Consistent with traditional nonparametric mean estimation, $\sqrt{nh_2}$ consistency is achieved. The asymptotic bias term is the same as in the local linear mean estimation because of the faster convergence rate of the estimator in the first step, while the ratio of the asymptotic variance of the modal-based robust estimators to those of the local linear mean estimators is given by $Var(\epsilon \mid X, Z, V = v)G(h_1)/F^2(h_1)$. Following Yao et al. (2012), we can demonstrate that the infimum of the above ratio is equal to one for all $h_1 > 0$, implying that the performance of

[13] As the convergence rate of the first step estimator is faster, the asymptotic theorem can be proved by combining the results in Yao et al. (2012), Zhang et al. (2013), and the proof of Theorem 2.4.4 by treating bandwidth h_1 as a constant. We leave it out for the sake of brevity.

the modal-based estimation is better than (*the error distribution has heavy tails*) *or at least as good as* (*the error follows a normal distribution*) *the local linear mean estimation. With undersmoothing* ($lim_{n\to\infty}\sqrt{nh_2^5} \to 0$)*, the bias can be asymptotically ignored.*[14]

To investigate the property of the second stage estimators, we rewrite (2.7) as

$$\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i - m(V_i) - X_i\beta - Z_{1,i}^T\gamma + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right), \tag{2.10}$$

which shows that there exist two extra terms in the true objective function, namely, $m(V_i) - \tilde{m}(V_i) = O_p(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2) = O_p(n^{-2/8})$ and $\tilde{m}(\hat{V}_i) - \tilde{m}(V_i) = \tilde{m}^{(1)}(\bar{V}_i)(\hat{V}_i - V_i) = O_p(\left(nh^3\right)^{-1/2} + h^2) = O_p(n^{-2/7})$, needing to be taken into account, in which $\tilde{m}^{(1)}(\bar{V}_i)$ is the first derivative of $\tilde{m}(\cdot)$ with respect to V_i assessed at $\bar{V}_i$. With mild regularity conditions on bandwidths, these two additional components can converge sufficiently fast to be disregarded asymptotically. Built on these, we then formally establish the following asymptotic results for the second stage estimators.

Theorem 2.4.5 *Under the regularity conditions C1-C7 and the additional bandwidth conditions* $h/h_2 \to 0$, $h_1/h_3 \to 0$, *and* $h_2/h_3 \to 0$, *with probability approaching one, as* $n \to \infty, h_3 \to 0$, *and* $nh_3^5 \to \infty$, *there exists a consistent maximizer* $\hat{\eta} = (\hat{\beta}, \hat{\gamma}^T)^T$ *of* (2.7) *such that*

$$\|\hat{\eta} - \eta_0\| = O_p\left(\left(nh_3^3\right)^{-1/2} + h_3^2\right).$$

Theorem 2.4.6 *With* $nh_3^7 = O(1)$, *under the same conditions as Theorem 2.4.5, the estimator satisfying the consistency result in Theorem 2.4.5 has the following asymptotic result*

$$\sqrt{nh_3^3}\left(\hat{\eta} - \eta_0 - \frac{h_3^2}{2}J_X^{-1}M_X(1 + o_p(1))\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dt J_X^{-1}L_X J_X^{-1}\right).$$

[14] Practically, h_1 can be chosen in the same way (maximizing the asymptotic relative efficiency) shown in Remark 2.4.11, while h_2 can be obtained by minimizing the MSE of the modal-based estimators such that $h_{2,opt} = (Var(\epsilon \mid X, Z, V = v)G(h_1)/F^2(h_1))^{-1/5}h_{2,mean}$, where $h_{2,mean}$ is the asymptotic optimal bandwidth for the local linear mean estimation. With the requirement of undersmoothing, we can further let $h_{2,opt}^* = h_{2,opt} \times n^{-2/15}$.

Furthermore, under the assumption that $nh_3^7 \to 0$*, we have*

$$\sqrt{nh_3^3}\,(\hat{\eta} - \eta_0) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dt J_X^{-1} L_X J_X^{-1}\right),$$

where $J_X = E(Z_X Z_X^T g_\epsilon^{(2)}(0 \mid X, Z))$*,* $L_X = E(Z_X Z_X^T g_\epsilon(0 \mid X, Z))$*, and* $M_X = E(Z_X g_\epsilon^{(3)}(0 \mid X, Z))$*.*

The preceding two theorems indicate that although there is an endogeneity issue and V_i has to be estimated preliminarily, we can achieve the optimal convergence rate of parametric components as in the usual linear modal regression under appropriate conditions. This finding is in agreement with the classical result for the partially linear mean regression model. The conditions $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$ indicate that $h_1 \to 0$ and $h_2 \to 0$ are faster than $h_3 \to 0$ as $n \to 0$, which is required to guarantee that the influence of the bias term in the first stage is not carried over to the second stage. We emphasize that the MSE-optimal rate of bandwidth h_3 is $n^{-1/7}$, which obviously converges faster than the MSE-optimal rate $n^{-1/8}$ for h_1 and h_2. To reconcile this contradiction, we need to impose a restrictive condition on the bandwidths h_1 and h_2 in order to achieve the oracle property (see the following discussions on bandwidth selection in practice). As with the previous estimators, if we further impose the undersmoothing condition ($lim_{n\to\infty}\sqrt{nh_3^5} \to 0$), the estimator $\hat{\eta}$ is shown to be asymptotically normal centered at the true value priced at a slower convergence rate.

Remark 2.4.13 *The results of Theorem 2.4.6 resemble many other kernel-based multi-stage nonparametric procedures, in which the first stage estimators do not contribute to the asymptotic variance of the current stage estimators due to undersmoothing. However, this*

is not the case in most parametric estimation problems; see Newey et al. (1999) and the subsequent results for modal-based robust estimation in Remark 2.4.14, where the first stage estimators do contribute to the asymptotic variance of the second stage estimators. On the other hand, if we do not impose bandwidth constraints to ensure that the estimators from the first stage converge faster, the convergence rate of $\hat{\eta}$ will be dominated by $\tilde{m}(v)$ and $\tilde{\eta}$, which is slower than $\sqrt{nh_3^3}$ and depends on bandwidths h_1 and h_2.

Remark 2.4.14 (Modal-Based Robust Estimators (Second Stage)) *Because only data in a local neighborhood of v are used to estimate the parametric parameters, after we obtain the modal-based robust estimates from the first stage estimation, we can treat the non-parametric part as known and estimate the parametric part with a constant h_3 to improve the convergence rate. Undersmoothing, like in modal regression, is necessary to asymptotically ignore the bias from the previous stage and achieve $\sqrt{n}$-consistency and asymptotic normality. Define*

$$\Gamma_n = -\sum_{i=1}^{n} \phi_{h_3}^{(2)}(\epsilon) Z_{X,i}(m(V_i) - \tilde{m}(V_i)) \xrightarrow{p} \Gamma_1.$$

Following the results in Remark 2.4.11, if $nh_3^4 \to 0$ and $nh_3^2/\log(1/h_3) \to \infty$ as $n \to \infty$, it can be demonstrated that

$$\sqrt{n}\,(\hat{\eta}_{robust} - \eta_0) \xrightarrow{d} N\big\{0, \big(E\big\{\phi_{h_3}^{(2)}(\epsilon) \mid X, Z\big\}\big)^{-2} Var\big\{(Z_X \phi_{h_3}^{(1)}(\epsilon) - \Gamma_1) \mid X, Z\big\}\big\},$$

where $\hat{\eta}_{robust}$ is the modal-based robust estimator from (2.7). The comments for Remark 2.4.11, including bandwidth choice, are also applied here. In contrast to modal regression, the previous stage estimation now contributes to the asymptotic variance of $\hat{\eta}_{robust}$, reflected by term Γ_1. Similar results have been presented in Kai et al. (2011) for composite quantile regression.

Analogous to the previous stages, we can rewrite (2.8) as

$$\frac{1}{nh_4h_5}\sum_{i=1}^{n}\phi\left(\frac{Y_i - X_i\hat{\beta} - Z_{1,i}^T\hat{\gamma} - m(V_i) - (m(\hat{V}_i) - m(V_i))}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right), \tag{2.11}$$

which indicates that the bias effect $(m(\hat{V}_i) - m(V_i))$ from the first stage still needs to be accounted for, but can be asymptotically ignored as expected by imposing special conditions on bandwidths to ensure $m(\hat{V}_i)$ converges to the truth at a rate sufficiently fast. Taking the estimation bias from the second stage into account, we have the following results for the third stage estimators.

Theorem 2.4.7 *Under the regularity conditions C1-C7 and the additional bandwidth conditions* $h/h_5 \to 0$ *and* $h_3/h_5 \to 0$*, with probability approaching one, as* $n \to \infty, h_4 \to 0, h_5 \to 0$*,* $h_5^2/h_4 \to 0$*, and* $nh_5h_4^5 \to \infty$*, there exist consistent maximizers* $(\hat{m}(v), h_5\hat{m}^{(1)}(v))$ *of* (2.8) *such that*

i. $|\hat{m}(v) - m(v)| = O_p\left(\left(nh_5h_4^3\right)^{-1/2} + h_4^2 + h_5^2\right)$,

ii. $|h_5(\hat{m}^{(1)}(v) - m^{(1)}(v))| = O_p\left(\left(nh_5h_4^3\right)^{-1/2} + h_4^2 + h_5^2\right)$.

Theorem 2.4.8 *With* $nh_5^5h_4^3 = O(1)$ *and* $nh_5h_4^7 = O(1)$*, under the same conditions as Theorem 2.4.7, the estimators satisfying the consistency results in Theorem 2.4.7 have the following asymptotic result*

$$\sqrt{nh_5h_4^3}\Bigg[\begin{pmatrix}\hat{m}(v) - m(v)\\ h_5(\hat{m}^{(1)}(v) - m^{(1)}(v))\end{pmatrix} - \Gamma_2^{-1}\left(\frac{h_5^2}{2}m^{(2)}(v)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix} - \frac{h_4^2}{2}\frac{g_\epsilon^{(3)}(0 \mid V = v)}{g_\epsilon^{(2)}(0 \mid V = v)}\begin{pmatrix}\mu_0\\ \mu_1\end{pmatrix}\right)(1 + o_p(1))\Bigg] \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid V = v)\int t^2\phi^2(t)dt}{f_V(v)g_\epsilon^{(2)}(0 \mid V = v)^2}\Gamma_2^{-1}\Sigma_2\Gamma_2^{-1}\right).$$

If we allow $nh_5^5h_4^3 \to 0$ *and* $nh_5h_4^7 \to 0$*, the asymptotic theorem becomes*

$$\sqrt{nh_5h_4^3}\begin{pmatrix} \hat{m}(v) - m(v) \\ h_5(\hat{m}^{(1)}(v) - m^{(1)}(v)) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid V = v)\int t^2\phi^2(t)dt}{f_V(v)g_\epsilon^{(2)}(0 \mid V = v)^2}\Gamma_2^{-1}\Sigma_2\Gamma_2^{-1}\right),$$

where $\Gamma_2 = \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}$ *and* $\Sigma_2 = \begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}$*. With symmetric kernel* $K(\cdot)$*,* $\hat{m}(v)$ *and* $\hat{m}^{(1)}(v)$ *are asymptotically independent.*

Theorems 2.4.7 and 2.4.8 show that under appropriate conditions, the final estimators of the nonparametric component are asymptotically equivalent to the oracle case, where the true values of the parametric parts were known. The conditions $h/h_5 \to 0$ and $h_3/h_5 \to 0$ indicate that $h \to 0$ and $h_3 \to 0$ are faster than $h_5 \to 0$ as $n \to 0$, which are used to ensure the asymptotically negligible effect of the second stage estimation on the third stage. It is noticed that the MSE-optimal rate for bandwidths h_4 and h_5 is $n^{-1/8}$. Thus, in practice $h/h_5 \to 0$ and $h_3/h_5 \to 0$ are generally satisfied. Similar to the previous discussion, the bias will vanish with undersmoothing ($lim_{n\to\infty}\sqrt{nh_5h_4^7} \to 0$ and $lim_{n\to\infty}\sqrt{nh_5^5h_4^3} \to 0$). Theorem 2.4.8 shows that the third stage estimators have the similar formality of asymptotic theorems as the first stage estimators (having the same convergence rate), whereas they should be expected to have smaller asymptotic variances due to the use of known information of parametric components. As a result, the efficiency is improved.

Remark 2.4.15 (Modal-Based Robust Estimators (Third Stage)) *After obtaining the* $\sqrt{n}$*-consistency for the parametric part with modal-based robust estimation, we can carry out local linear estimation with the objective function* (2.8) *to update the estimator for the nonparametric part. If we treat* h_4 *as a constant and* $h_5 \to 0$ *as* $n \to \infty$ ($nh_5 \to \infty$),

with the similar conditions in Remark 2.4.12, the asymptotic distributions of $\hat{m}_{robust}(v)$ *and* $\hat{m}^{(1)}_{robust}(v)$ *estimated from* (2.8) *are*

$$\sqrt{nh_5}\left[\begin{pmatrix}\hat{m}_{robust}(v)-m(v)\\ h_2(\hat{m}^{(1)}_{robust}(v)-m^{(1)}(v))\end{pmatrix}-\Gamma_2^{-1}\frac{h_5^2}{2}m^{(2)}(v)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}\right]\xrightarrow{d}\mathcal{N}\left(0,\frac{G(h_4)}{F^2(h_4)}\Gamma_2^{-1}\Sigma_2\Gamma_2^{-1}\right).$$

As the convergence rate in the second stage is faster, there is no effect of $\hat{\eta}_{robust}$ *on estimators in the third stage. The comments for Remark 2.4.12 are also applicable here, demonstrating that the major advantage of modal-based estimation over mean estimation is the competitive asymptotic efficiency. Similar to the results in modal regression, the third stage modal-based robust estimators are more efficient than the corresponding first stage estimators, since the uncertainty in the parametric part does not need to be taken into consideration.*

One practical issue concerning the implementation of the proposed estimation procedure is the selection of bandwidths. We note that there appear to be few results available in the modal regression literature for data-driven bandwidth selection with optimal properties. Yao and Li (2014) and Ullah et al. (2021) suggested a plug-in method for bandwidth choice based on minimizing the asymptotic MSE of modal estimators, which may not be suitable for this paper due to the special conditions imposed on bandwidths. Precisely, the above asymptotic theorems indicate that the bandwidths in different steps/stages are required to satisfy different conditions to ensure that the asymptotic bias of the estimators in previous step/stages converges to zero at a faster rate than in the current estimation step/stage. Meanwhile, as the mode is capturing the "most likely" points, which is different from mean estimation, the conventional cross-validation method based on the MSE criterion is inapplicable. We thus propose a simple bandwidth selection procedure that combines the

optimal bandwidth rates (reflecting in the following power numbers of bandwidths) and the undersmoothing requirement.

It is worth noting that bandwidths h_2 and h_5 serve the same purpose (i.e., controlling smoothness) as they do in nonparametric estimation. For simplicity, we follow a rule of thumb to set

$$h_2 = 1.06\hat{\sigma}(\hat{V}_i)n^{-0.15} \text{ and } h_5 = 1.06\hat{\sigma}(\hat{V}_i)n^{-0.13},$$

where $\hat{\sigma}(\hat{V}_i)$ is the standard deviation of variable $\hat{V}_i$, and 0.15 and 0.13 are from the MSE-optimal convergence rates and undersmoothing requirement. For bandwidths h, h_1, h_3, and h_4, they play much important roles in estimation and can determine the number of estimated modes. We work with the undersmoothing assumption on the bandwidths following Kemp and Santos Silva (2012) to apply the grid search method to select a number of potential bandwidths. Specifically, we obtain the mean regression residual first, and then select 50 bandwidth values ranging from 50MAD to $0.5\text{MAD}n^{-\gamma_{h_j}}$ ($\gamma_h = 0.16$, $\gamma_{h_1} = 0.15$, $\gamma_{h_3} = 0.143$, $\gamma_{h_4} = 0.13$), in which MAD is the median value of the absolute deviation of the mean regression residual from the corresponding median value and γ_{h_j} is from the MSE-optimal convergence rates and undersmoothing requirement. In empirical applications, we choose bandwidths as

$$h_j = 1.6MADn^{-\gamma_{h_j}}.$$

It is important to note that while the above bandwidth selection method may not offer global optimal estimates, it does provide a straightforward method for selecting bandwidths in numerical analysis that has been shown to perform effectively. The issue of how to choose optimal bandwidths with endogeneity in modal regression is an interesting one that merits more investigation.

Remark 2.4.16 *One way to consider the cross-validation method for modal estimation is based on the fact that with the same interval length, the interval around the conditional mode should cover more samples. Taking first step estimation as an example, we can then maximize the kernel-based objective function below to obtain a data-driven bandwidth*

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(\frac{Y_{-i} - \hat{\alpha}_{-i} - Z_{-i}^T \hat{\pi}_{-i}}{\bar{h}}\right),$$

where $-i$ represents "without the observation indexed by i" and $\hat{\alpha}_{-i}$ and $\hat{\pi}_{-i}$ are the corresponding modal estimators. Practically, we can choose $\bar{h}$ as $0.05\max|Y_i - Y_j|$, $i, j = 1, \cdots, n$. The theoretical property of such modal cross-validation deserves further study in the future.

2.5 Numerical Examples

To further illustrate the newly developed estimation procedure in dealing with endogeneity in modal regression and to support the theoretical developments for the proposed estimators, we carry out two Monte Carlo simulations and three real data analyses, one of which is presented in Appendix A. In all numerical studies, we deploy the normal kernel function defined as $\frac{1}{\sqrt{2\pi}} exp(-\frac{u^2}{2})$ for $\phi(\cdot)$ and $K(\cdot)$ and choose bandwidths using the approach described above. To elucidate the resultant modal-based robust estimators, we also conduct a Monte Carlo experiment with different error distributions, as shown in Appendix A.

2.5.1 Monte Carlo Experiments

Two simulation experiments are given in this part to illustrate the finite sample performance of the proposed estimators and are used to examine whether empirical evidence can be established in support of asymptotic normality. The first Monte Carlo experiment has a

nonparametric control function. The second set of Monte Carlo experiments uses the parametric control function but with different degrees of endogeneity. We use DGP to represent the data generating process in this subsection. To display the behavior of the developed estimators, we consider sample sizes of $n \in \{200, 400, 600, 1000\}$. In all simulations, a total of M =200 simulation replications are conducted, and the data are *i.i.d.* draws in each replication. For the sake of comparison, we also run the naive linear modal regression on the structural equation directly. For each simulation, we concentrate on the coefficients β and γ and compute the average values of estimates, the standard errors (SEs), and the MSEs of all estimators considered in order to compare and evaluate the performance of the proposed estimators, where

$$\text{MSE}(\hat{\beta}) = \frac{1}{M}\sum_{l=1}^{M}(\hat{\beta}_l - \beta)^2 \text{ and } \text{MSE}(\hat{\gamma}) = \frac{1}{M}\sum_{l=1}^{M}(\hat{\gamma}_l - \gamma)^2$$

in which $\hat{\beta}_l$ and $\hat{\gamma}_l$ are the lth estimators, and β and γ are the true values.

DGP 1 At first, we generate data according to the following model that satisfies the conditional mode independence assumption

$$\begin{cases} Y_i = X_i\beta + Z_{1,i}\gamma + U_i, \\ U_i = V_i + 4exp(-(V_i - 1)^2) + 0.5[\tilde{U}_i - 1], \\ X_i = \alpha + Z_{1,i}\pi_1 + Z_{2,i}\pi_2 + V_i, \end{cases}$$

where $Z_{1,i}$, $Z_{2,i}$, and V_i are drawn from the following multivariate normal distribution

$$\begin{pmatrix} Z_{1,i} \\ Z_{2,i} \\ V_i \end{pmatrix} \sim \text{i.i.d.}\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right),$$

and $\tilde{U}_i$ is drawn from a skewed distribution $0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ with $E(\tilde{U}) = 0$ and $Mode(\tilde{U}) = 1$ (Yao and Li, 2014; Ullah et al., 2021). Without loss of generality, we only consider the case of one instrumental variable $Z_{2,i}$ and denote the set of all instruments as $Z_i = [Z_{1,i}\ Z_{2,i}]^T$. The regressor X_i is correlated with the error term U_i through V_i, and the instrumental variable set Z_i is correlated with X_i but not with U_i to correct for endogeneity in modal regression. The parameter values are set as $(\beta, \gamma, \alpha, \pi_1, \pi_2) = (1, 1, 1, 1, 1)$. We then have $Mode(V_i \mid Z_i) = 0$, $Mode(\tilde{U}_i \mid Z_i) = 1$, and the control function[15]

$$Mode(U_i \mid V_i, Z_i) = V_i + 4exp(-(V_i - 1)^2).$$

Table 2.1: Results of Simulations—DGP 1

	Two-Step Estimation				Naive Estimation			
Sample Size	β (SE)	MSE(β)	γ (SE)	MSE(γ)	β (SE)	MSE(β)	γ (SE)	MSE(γ)
n=200	0.9067 (0.4580)	0.2174	1.0525 (0.5961)	0.3563	2.1171 (0.2419)	1.3061	-0.1066 (0.4891)	1.4626
n=400	0.9778 (0.3155)	0.0996	1.0443 (0.4236)	0.1805	2.1172 (0.2098)	1.2919	-0.1109 (0.4416)	1.4281
n=600	0.9502 (0.2339)	0.0569	1.0203 (0.3563)	0.1267	2.1307 (0.1605)	1.3043	-0.1141 (0.3516)	1.3641
n=1000	0.9742 (0.1756)	0.0314	1.0313 (0.2446)	0.0605	2.1757 (0.1757)	1.4130	-0.1599 (0.3645)	1.4776
True Value	$\beta = 1$		$\gamma = 1$					

The estimation results for β and γ are shown in Table 2.1, from which we can see that the proposed estimation procedure work well for all sample sizes considered. The linear modal regression estimators without addressing endogeneity (naive estimation) are inconsistent with larger positive biases for β and negative biases for γ for all sample sizes (columns 6-9 in Table 2.1), whereas the proposed two-step estimators (columns 2-5 in Table 2.1) can approximate the true values of parameters with reasonable biases. As expected, with the sample size increasing, the biases for the naive estimators do not shrink toward zero

[15]According to the simulation setting, we can obtain the mean control function $E(U_i \mid V_i, Z_i) = V_i + 4exp(-(V_i-1)^2) - 0.5$, which indicates the difference in control function between mean and modal regressions when the data are skewed. In this section we do not compare the performance of modal regression to that of mean regression. However, the interested readers are referred to Yao and Li (2014), Yao and Xiang (2016), and Ullah et al. (2021, 2022) for more simulation examples about the comparisons of these two models in terms of prediction performance.

and remain substantial even when n is large, while the proposed estimators rapidly converge to the true values of parameters and the standard errors shrink quickly. In addition, as a result of undersmoothing, the variance of the new estimator dominates in MSE. Although it is hard to gauge the convergence rate of the proposed estimators with undersmoothing from the reported results, their MSEs decrease steadily with increasing sample size in all cases.[16] All of the results are in line with the asymptotic properties, implying that the proposed estimators are indeed consistent.

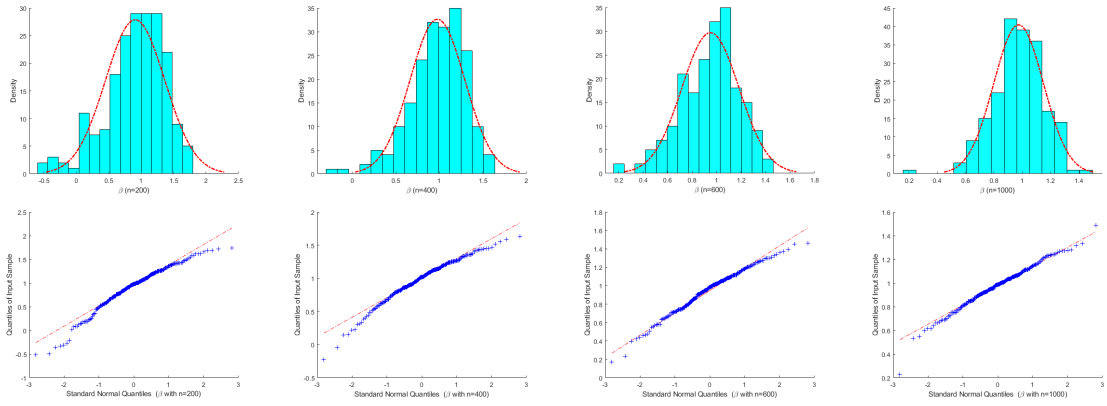

Figure 2.2: Histograms and QQ Plots for Estimates (β)—DGP 1

To further illustrate the asymptotic behavior of the proposed estimators, we provide a visual comparison to a normal distribution by displaying the corresponding histograms and quantile-quantile (QQ) plots for the simulated estimates of β and γ in Figures 2.2 and 2.3, respectively. The plots are all in accord with the theoretical results that the p-rposed estimators are asymptotically normally distributed. The histograms for the sample

[16] We emphasize that although the finite sample performance of the proposed estimators is relatively good according to the simulation results, the lack of data-driven choice of bandwidths could be a disadvantage of the proposed estimation procedure, and shall be explored in a future study. In addition, Table 2.1 shows that there exists some small bias in the proposed modal estimators. Such an issue can be addressed in more depth using the bootstrap methodology for bias correction.

estimates are centered at the true values of the population parameters. As the sample size n increases, the points in the QQ plots match up along a straight line more, which indicates that the asymptotic normal approximation becomes more precise for these two estimators.

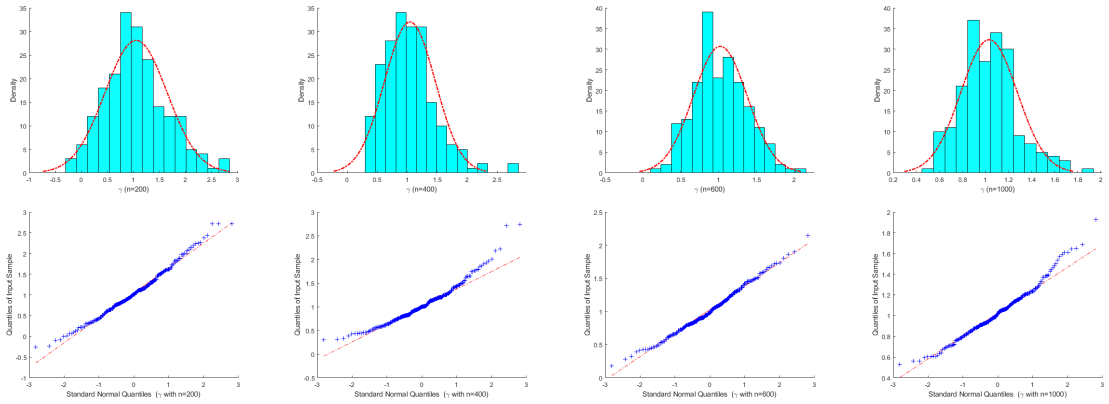

Figure 2.3: Histograms and QQ Plots for Estimates (γ)—DGP 1

DGP 2 In this setting, we conduct simulation experiments with different degrees of endogeneity and generate data according to the model described below

$$\begin{cases} Y_i = X_i\beta + Z_{1,i}\gamma + U_i, \\ X_i = \alpha + Z_{1,i}\pi_1 + Z_{2,i}\pi_2 + V_i, \ i = 1, \cdots, n, \end{cases}$$

where we set the parameter values as $(\beta, \gamma, \alpha, \pi_1, \pi_2) = (3, 2, 1, 1, 1)$ and draw $Z_{1,i}$, $Z_{2,i}$, V_i, and U_i from the multivariate normal distribution with zero mean, unit variance, and correlation coefficient ρ for V_i and U_i

$$\begin{pmatrix} Z_{1,i} \\ Z_{2,i} \\ V_i \\ U_i \end{pmatrix} \sim \text{i.i.d.} N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix} \right),$$

in which ρ =0.2, 0.5, and 0.8 represent weak, middle, and strong endogeneity, respectively. The different magnitude of the endogeneity strength is comparable to the setting in Su and Ullah (2008). With normal distribution, it is then easy to verify that $Mode(V_i \mid Z_i) = 0$, $Mode(U_i \mid Z_i) = 0$, and the control function $Mode(U_i \mid V_i, Z_i) = \rho V_i$, which satisfies the conditional mode independence restriction.[17]

Table 2.2: Results of Simulations—DGP 2

	Two-Step Estimation				Naive Estimation			
Sample Size	β (SE)	MSE(β)	γ (SE)	MSE(γ)	β (SE)	MSE(β)	γ (SE)	MSE(γ)
				$\rho = 0.2$				
n=200	2.9935 (0.1515)	0.0229	1.9908 (0.1981)	0.0391	3.0594 (0.0959)	0.0127	1.9310 (0.1833)	0.0382
n=400	3.0083 (0.0941)	0.0089	1.9945 (0.1432)	0.0204	3.0680 (0.0674)	0.0091	1.9322 (0.1310)	0.0217
n=600	2.9976 (0.0824)	0.0068	1.9957 (0.1227)	0.0150	3.0617 (0.0572)	0.0071	1.9287 (0.1135)	0.0179
n=1000	2.9960 (0.0681)	0.0046	1.9956 (0.1077)	0.0116	3.0684 (0.0470)	0.0069	1.9222 (0.0933)	0.0147
				$\rho = 0.5$				
n=200	2.9974 (0.1367)	0.0186	2.0024 (0.1936)	0.0373	3.1777 (0.0835)	0.0385	1.8173 (0.1855)	0.0676
n=400	2.9958 (0.0995)	0.0099	2.0049 (0.1494)	0.0222	3.1647 (0.0677)	0.0317	1.8333 (0.1255)	0.0435
n=600	2.9890 (0.0851)	0.0073	2.0020 (0.1240)	0.0153	3.1654 (0.0527)	0.0301	1.8401 (0.1042)	0.0364
n=1000	2.9935 (0.0702)	0.0049	2.0116 (0.0965)	0.0094	3.1638 (0.0466)	0.0290	1.8426 (0.0828)	0.0316
				$\rho = 0.8$				
n=200	2.9512 (0.1813)	0.0351	2.0346 (0.2452)	0.0610	3.2595 (0.0705)	0.0723	1.7399 (0.1519)	0.0906
n=400	2.9858 (0.1142)	0.0132	2.0098 (0.1517)	0.0230	3.2629 (0.0542)	0.0720	1.7308 (0.1174)	0.0862
n=600	2.9976 (0.0828)	0.0068	1.9968 (0.1209)	0.0145	3.2617 (0.0393)	0.0700	1.7393 (0.0948)	0.0769
n=1000	2.9944 (0.0658)	0.0043	2.0079 (0.0984)	0.0097	3.2567 (0.0338)	0.0671	1.7398 (0.0696)	0.0725
True Value	$\beta = 3$		$\gamma = 2$					

Table 2.2 provides finite sample results for estimates of β and γ with different values of ρ. The similar conclusions as those in DGP 1 can be drawn from the results. Compared to naive linear modal regression without taking endogeneity into account, the proposed estimation procedure has very nice finite sample properties even in small samples when endogeneity is relatively strong. In particular, we note that as sample size increases,

[17]Different from DGP 1, according to the simulation setting in DGP 2, we can obtain the mean control function $E(U_i \mid V_i, Z_i) = \rho U_i$ as well, which indicates the same control function for mean and modal regressions given a symmetric dataset. However, since we are focusing on modal estimation with shrinkage bandwidths, we do not compare the results to those of mean regression. The additional results with regard to modal-based robust estimation are shown in Appendix A.

both the bias and MSE of the proposed estimators decrease. However, when ρ is small, indicating weak endogeneity in modal regression, the developed estimators provide better performance than the naive estimators in terms of bias but worse performance in terms of MSE with a small sample size ($n = 200$). It has been observed that the naive estimators are biased in all cases and the magnitude of bias becomes larger as ρ increases, which is consistent with the reality that endogeneity leads to biased estimates. The results highlight that in the absence of a general guideline for testing endogeneity in modal regression, it may have benefit to apply the two-step estimation procedure suggested in this paper with relatively large data to avoid any potential misspecification stemming from endogeneity in practice.

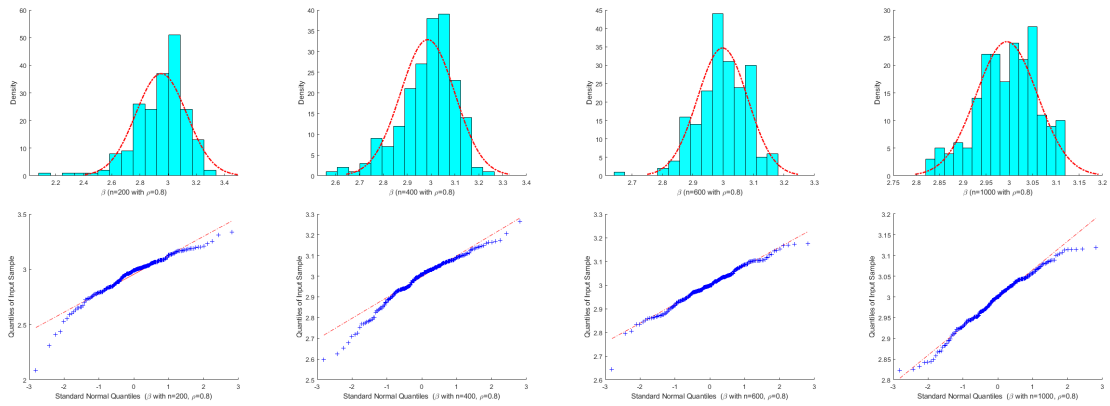

Figure 2.4: Histograms and QQ Plots for Estimates (β with $\rho = 0.8$)—DGP 2

We display histograms and QQ plots in Figures 2.4 and 2.5 in the same manner as in DGP 1. Due to space limitations, we only list the plots for the case $\rho = 0.8$. The appearances of the plots for the other two cases are similar, which are given in Appendix A. Similar to the findings in DGP 1, the distributions appear to be symmetric around the

true value in all designs. It is apparent that the figures indicate the asymptotic normality of the proposed estimators, which is compatible with the theoretical results presented in Section 2.4.

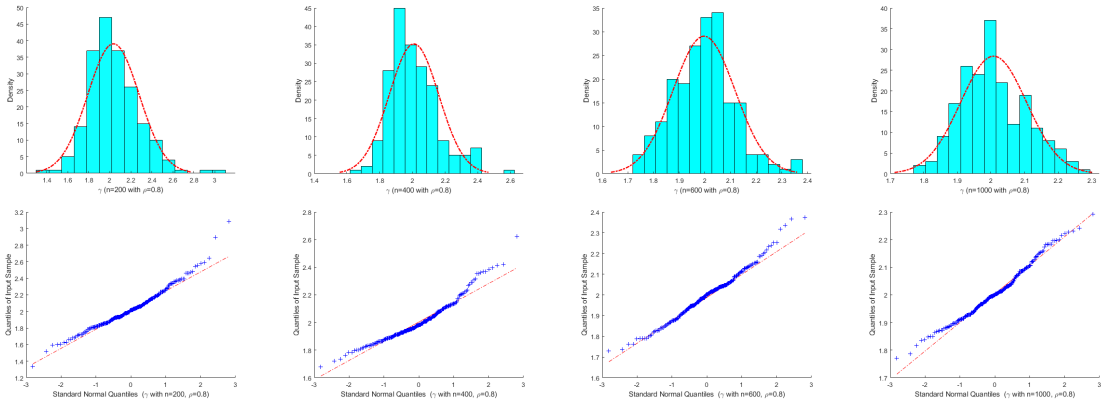

Figure 2.5: Histograms and QQ Plots for Estimates (γ with $\rho = 0.8$)—DGP 2

2.5.2 Empirical Analyses

We present three applications of the estimation results derived in Section 2.3. The first application is to utilize the proposed estimation method to estimate EIS resulting from modal utility maximization, where the agent maximizes the present discounted value of the stream of future modal utilities rather than the traditional expected utilities. We in the second example apply the proposed modal regression to study the effect of institutions that protect property rights on economic performance, which was originally investigated in Acemoglu et al. (2001). The third application reports the results of a simple analysis of return to schooling, which makes use of samples from Card (1995) to illustrate potential differences between the proposed model and the existing regression models (shown in Appendix A). These three examples demonstrate the bias in modal regression under endogeneity and show that the

variables we focus on are economically relevant based on mode value after correcting endogeneity. The empirical analyses also show that the proposed model can reveal certain interesting data structures that existing regressions may ignore.

I. Rational Behavior under Modal Utility Maximization

For the purpose of fixing the idea of control function method proposed in this paper, we initiate the use of modal preferences in a dynamic economic setting by developing a dynamic model of rational behavior under uncertainty, in which the agent maximizes the present discounted value of the stream of future mode utilities, and deriving a modal Euler equation from the maximization model. In other words, the agent prefers a modal utility preference rather than the expected utility as is often assumed. We then utilize the data from Yogo (2004) to estimate the EIS from mode value, which is a parameter of central importance in macroeconomics and finance measuring the responsiveness of the consumption growth rate to the real interest rate; see Campbell (2003) for more information. The resulting Euler equation from the modal utility maximization model is attractive because it can complement the existing expected utility and quantile utility models to reveal distinguishing features of the data and capture the "most likely" effect. Therefore, the modal utility preference investigated here provides the microeconomic foundations for modal regression.

It is reasonable in practice to assume modal preferences as the agent may be more concerned with maximization of the "most likely" utility, instead of the expected utility. Particularly, the agent will prefer X over Y ($X \succeq Y$) iff $Mode(U(X)) \geq Mode(U(Y))$ and vice versa. It is important to emphasize that the modal preference is independent of the utility function, as for any continuous and strictly increasing function $U(\cdot)$,

$$X \succeq Y \Longleftrightarrow Mode[U(X)]Mode[U(Y)] \Longleftrightarrow U\left(Mode[X]\right)U\left(Mode[Y]\right) \Longleftrightarrow Mode[X]Mode[Y]. \tag{2.12}$$

Thus, in contrast to the expected utility maximizer that the concavity of $U(\cdot)$ implies risk aversion, the function itself does not have any impact on the risk attitude of a modal maximizer. However, when the utility function has more than one argument, it is not suitable to apply the above invariance property to get rid of $U(\cdot)$.

On the basis of the previous discussions, we focus on a model that describes the behavior of economic agents associated with the decision on intertemporal consumption and savings across an infinity horizon economy, following the model framework in de Castro et al. (2019) and de Castro and Galvao (2019). Define C_t as the amount of consumption goods that an individual consumes in period t. The consumer owns x_t units of the risky asset with price $p(d_t)$, which pays dividend of d_t, at the beginning of the period t. For the sake of simplicity, the price of the consumption good is normalized to one. With wealth $\left[d_t + p\left(d_t\right)\right] x_t$, the consumer decides how many units of the risky asset x_{t+1} to save for the next period and the consumption C_t. Then, the budget constraint is

$$C_t + p\left(d_t\right) x_{t+1} \leq \left[d_t + p\left(d_t\right)\right] x_t, \tag{2.13}$$

where the positivity restriction $C_t, x_{t+1} \geq 0$ should be satisfied as well. Different from the mean- or quantile-maximization, we retain the standard additive separability in time and assume that the consumer maximizes

$$Mode\left[\sum_{t=0}^{\infty} \beta^t U\left(C_t\right) \mid \Omega_0\right] \tag{2.14}$$

over an infinity horizon economy, where $\beta \in (0,1)$ is the discount factor, $U : R_+ \mapsto R$ is the utility function, and Ω_0 is the information set at time $t = 0$. Following the procedures

in the mean and quantile Euler equations to apply the recursive substitution, the consumer maximizes the following modal objective function with budget constraint

$$\begin{aligned}
&U(C_0) + Mode\left[\beta U(C_1) + Mode\left[\beta^2 U(C_2) + Mode\left[\beta U(C_3) + \cdots \mid \Omega_2\right] \mid \Omega_1\right] \mid \Omega_0\right] \\
=&Mode\left[Mode\left[Mode\left[U(C_0) + \beta U(C_1) + \beta^2 U(C_2) + \beta^3 U(C_3) + \cdots \mid \Omega_2\right] \mid \Omega_1\right] \mid \Omega_0\right] \\
=&Mode^{\infty}\left[\sum_{t=0}^{\infty} \beta^t U(C_t)\right].
\end{aligned} \tag{2.15}$$

In contrast to the mean Euler equation, we cannot substitute $Mode^{\infty}$ with $Mode$ because the law of iteration and linearity does not hold for mode. However, the above modal expression is similar to the quantile expression in de Castro and Galvao (2019), where one can show that the value function exists and is differentiable under regularity conditions.[18]

Remark 2.5.17 (Additive Property) *The law of iteration does not apply to mode, indicating that for any two σ-algebras $\Omega_t \subset \Omega_{t+1}$, in general, $Mode[Mode(X \mid \Omega_{t+1}) \mid \Omega_t] \neq Mode(X \mid \Omega_t)$. Nevertheless, we can derive the following condition under which the additive property is satisfied. Given the random variables W_1 and W_2, assume there exists a random variable G and continuous and increasing functions g_1 and g_2 such that $W_1 = g_1(G)$ and $W_2 = g_2(G)$. Then,*

$$Mode(W_1 + W_2) = Mode(W_1) + Mode(W_2).$$

To demonstrate this, we define $r(G) = g_1(G) + g_2(G)$. By virtue of the invariance property of mode ($Mode(g_1(G)) = g_1(Mode(G))$ with a strictly increasing and continuous function g_1), we have $Mode(W_1 + W_2) = Mode(g_1(G) + g_2(G)) = Mode(r(G)) = r(Mode(G)) =$

[18]Combining with the value function, we have $v(x_0, d_0) = Mode^T\left[\sum_{t=0}^{T-1} \beta^t U(C_t) + \beta^T v(x_t, d_T)\right]$. As $T \to \infty$, $\beta^T v(x_t, d_T) \to 0$ when v is bounded. Such theoretical developments and derivations are of independent interest. We will address these in detail in another ongoing research.

$g_1(Mode(G)) + g_2(Mode(G)) = Mode(g_1(G)) + Mode(g_2(G)) = Mode(W_1) + Mode(W_2)$. *Such an additive property is extremely useful for deriving the modal Euler equation; see Remark 2.5.18.*

Compared to the mean and quantile value functions, we have

$$v\left(x_t, d_t\right) = \max_{x_{t+1} \geq 0} \left\{U\left(\left[d_t + p\left(d_t\right)\right] x_t - p\left(d_t\right) x_{t+1}\right) + \beta Mode\left[v\left(x_{t+1}, d_{t+1}\right) \mid \Omega_t\right]\right\}, \quad (2.16)$$

where $v(\cdot)$ is a value function and the mode is taken conditional on the time t information available to the econometrician. It indicates that the value function at time t is equal to the utility of consumption at time t plus the discounted value of the mode of the value function at time $t+1$. In equilibrium, the holdings are $x_t = 1$ for all t. We then obtain

$$-p\left(d_t\right) U^{(1)}\left(C_t\right) + \beta Mode[U^{(1)}\left(C_{t+1}\right)\left(d_{t+1} + p\left(d_{t+1}\right) \mid \Omega_t\right] = 0, \quad (2.17)$$

where $U^{(1)}(t)$ represents the first derivative of U with respect to t.

Remark 2.5.18 *The above result is built on the equation that*

$$\frac{\partial Mode}{\partial x}[v(x,d)] = Mode\left[\frac{\partial v}{\partial x}(x,d)\right],$$

which holds true if $\frac{\partial v(x,d)}{\partial d} \geq 0$ *and* $\frac{\partial^2 v(x,d)}{\partial x \partial d} \geq 0$. *These two required equations imply that* $v(x,d)$ *and* $v(x',d) - v(x,d)$ *are increasing in* d *for* $x' > x$. *According to the result in Remark 2.5.17, we can write* $Mode(v(x+\delta,d)) = Mode(v(x+\delta,d) - v(x,d) + v(x,d)) = Mode(v(x+\delta,d) - v(x,d)) + Mode(v(x,d))$ *for a sufficiently small* δ. *Then, following the definition of derivative considering from limitation, it can be demonstrated that* $Mode(v(x,d))$ *is differentiable and the derivative is* $Mode\left[\frac{\partial v}{\partial x}(x,d)\right]$.

By defining the asset's return as $1 + r_{t+1} \equiv \frac{d_{t+1}+p(d_{t+1})}{p(d_t)}$, we simplify the Euler equation to

$$Mode\left[\beta\left(1+r_{t+1}\right)\frac{U^{(1)}\left(C_{t+1}\right)}{U^{(1)}\left(C_t\right)}-1 \mid \Omega_t\right]=0, \tag{2.18}$$

which is a modal regression but with the endogeneity issue. Given a constant relative risk aversion utility function $U(C) = C^{1-\gamma}/(1-\gamma)$ and instruments Z_t chosen from Ω_t, the Euler equation can be rewritten as

$$Mode\left[\beta\left(1+r_{t+1}\right)\left(C_{t+1}/C_t\right)^{-\gamma}-1 \mid Z_t\right]=0, \tag{2.19}$$

which is a conditional mode restriction in the form of our econometric modal regression models with instrument variables.

Consistent with the quantile Euler equation, the modal Euler equation can be log-linearized with no approximation error as well. It is observed that for a random variable W with a unique global mode, $Mode(ln(W)) = ln(Mode(W))$ as $ln(.)$ is strictly increasing and continuous ("invariance" with respect to monotonic transformation).[19] Define $\varepsilon_{t+1} = \beta\left(1+r_{t+1}\right)\left(C_{t+1}/C_t\right)^{-\gamma}$. We have $Mode(\varepsilon_{t+1} \mid \Omega_t) = 1$ and

$$ln\left(1+r_{t+1}\right)=\gamma ln\left(C_{t+1}/C_t\right)-ln(\beta)+ln\left(\varepsilon_{t+1}\right). \tag{2.20}$$

The above equation indicates that

$$Mode(ln\left(1+r_{t+1}\right)-\gamma ln\left(C_{t+1}/C_t\right)+ln(\beta) \mid \Omega_t)=0, \tag{2.21}$$

where the parameter $1/\gamma$ is the standard measure of EIS implicit in the utility function.[20] Note that the EIS in modal regression shares the same interpretation as in mean regression, but considered from a mode viewpoint.

[19] As $E(ln(W))$ is not necessarily equal to $ln(E(W))$, the log-linearization will bring higher-order terms to the equation when we act as if the natural logarithm could be interchanged with $E(\cdot)$.

[20] The reciprocal of γ is the coefficient of relative risk aversion under power utility. Based on this equation, we can also estimate the discount factor from a mode perspective.

Numerical Results: We use the aggregate level quarterly data from Yogo (2004) to estimate EIS for Australia, Canada, France, Germany, Italy, Japan, Netherlands, Sweden, Switzerland, the United Kingdom, and the United States. The primary sources of international data are Morgan Stanley Capital International and the International Financial Statistics of the International Monetary Fund. For the dataset, the real interest rate is constructed using a proxy for the nominal short-term interesting rate, and real consumption growth is the first difference in log real consumption per capita. The instruments for the endogenous regressor consumption are twice lagged measures of real consumption growth, nominal interest rate, inflation, and a log dividend-price ratio for equities.

Table 2.3: Estimates of the 1/EIS using the Interest Rate as Dependent Variable

Country	Sample Period	Two-Step Modal	Naive Linear Modal	Mean-2SLS
Australia	1970.3-1998.4	0.4647	0.1856	0.4906
Canada	1970.3-1999.1	-1.0070	-0.1661	-1.0374
France	1970.3-1998.3	-4.3545	-0.0562	-3.1177
Germany	1979.1-1998.3	-0.7838	-0.0213	-1.0541
Italy	1971.4-1998.1	-2.4578	-0.4153	-3.3401
Japan	1970.3-1998.4	-0.8375	-0.0054	-0.1841
Netherlands	1977.3-1998.4	-0.2740	-0.0732	-0.5260
Sweden	1970.3-1999.2	-0.1411	0.0357	-0.0956
Switzerland	1976.2-1998.4	-1.4561	-0.0807	-1.5637
United Kingdom	1970.3-1999.1	0.2760	0.2464	1.0604
United States	1947.3-1998.4	0.7170	0.4218	0.6833

The estimated results are shown in Tables 2.3-2.4, where we also report the naive linear modal estimates and the mean-2SLS estimates for comparison. Because we are primarily concerned with the magnitude of estimates, we do not provide standard errors. However, they can be simply obtained by bootstrap method if researchers intend to conduct inference. It is observed that both modal and mean regressions can be utilized to

capture the fact that $1/\gamma$ is relatively small. The estimates of γ from modal regression are different from those from mean regression for most countries, which can be attributed to the skewed datasets (Figure 2.6).

Table 2.4: Estimates of the EIS using the Interest Rate as Covariate

Country	Sample Period	Two-Step Modal	Naive Linear Modal	Mean-2SLS
Australia	1970.3-1998.4	0.0357	0.1209	0.0453
Canada	1970.3-1999.1	-0.3557	-0.1748	-0.3046
France	1970.3-1998.3	-0.2297	-0.1519	-0.0813
Germany	1979.1-1998.3	-0.1853	-0.0831	-0.4195
Italy	1971.4-1998.1	-0.1145	-0.1328	-0.0709
Japan	1970.3-1998.4	-0.2887	-0.0093	-0.0388
Netherlands	1977.3-1998.4	0.0600	-0.1805	-0.1481
Sweden	1970.3-1999.2	-0.0167	0.0116	-0.0018
Switzerland	1976.2-1998.4	-0.4845	-0.1688	-0.4883
United Kingdom	1970.3-1999.1	0.0481	0.1851	0.1666
United States	1947.3-1998.4	0.0145	0.1606	0.0597

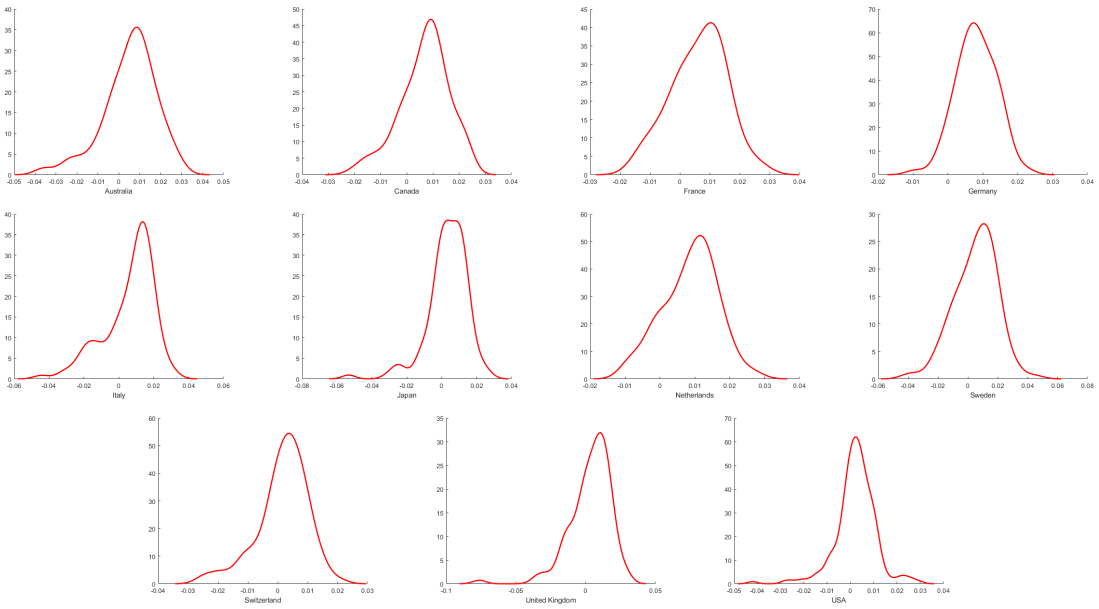

Figure 2.6: Empirical Distribution of the Real Interest Data

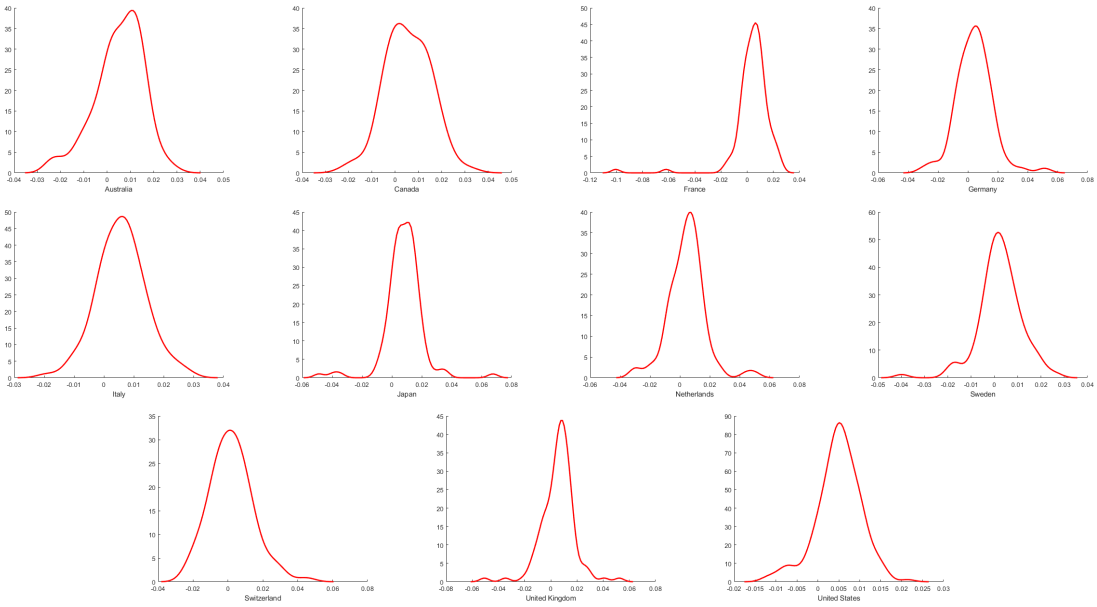

Figure 2.7: Empirical Distribution of the Real Consumption Growth

Concentrating on the EIS estimates (Table 2.4), we can see that the naive linear modal estimator is biased, yielding values that are completely different from those of the proposed estimator. For most countries, the developed modal estimates have the same signs but different magnitudes as the mean estimates. When the data are nearly symmetrically distributed (Figure 2.7), the differences between the proposed estimation and the mean-2SLS estimation are small. However, the results show some substantial differences with the skewed data, such as for France and the United Kingdom. Moreover, the modal estimate of EIS for Netherlands differs significantly from the mean estimate. Particularly, modal estimation gives a positive result, whereas mean estimation provides a negative one. Generally, the value of EIS equal to one is of economic interest since it implies that the consumer's optimal consumption choice is a constant fraction of wealth. However, no country's EIS

is close to one when the modal Euler equation is estimated.[21] All of these findings reveal that the proposed modal regression with endogeneity can be served as an important tool for studying economic behavior. We note that the estimates of EIS for Italy based on the developed estimation method and the naive estimation method are quite similar. This may imply that there is no modal endogenity issue in the Italy data, which urges the development of an endogeneity test for modal regression.

II. Colonial Origins of Comparative Development

There exists a large number of literature emphasizing the role of institutions and property rights in modern economic development. According to Acemoglu et al. (2001), countries with better institutions, more secure property rights, and less distortionary policies will utilize physical and human capital more efficiently to achieve a greater level of income. They estimated the impact of institutions on economic performance through investigating institutional differences among countries colonized by European. The main data they used include the mortality rates of soldiers, bishops, and sailors stationed in the colonies between the seventeenth and nineteenth centuries (em), the GDP per capita in 1995 (pgp), the average protection against expropriation risk between 1985 and 1995 ($avexpr$), and the latitude value (lat), where $avexpr$ is the index from Political Risk Services that be used as a proxy for institutions. To account for the endogeneity in $avexpr$, they used the mortality rates expected by the first European settlers in the colonies as an instrument for current institutions in these countries. We in this subsection shall reinvestigate their results from

[21] We only interpret the results from the difference between the modal and mean estimates, and do not pay much attention to negative EIS values. Note that a negative EIS indicates convex utility, thus the estimate is likely a statistical artifact. Havranek (2015) provided a meta-analysis of the literature in estimating EIS and concluded that "the literature shows strong selective reporting: researchers discard negative and insignificant estimates too often, which pulls the mean estimate up by about 0.5."

mode value by applying the proposed estimation procedure, where the main linear modal regression is formulated as

$$\log(pgp) = avexpr\beta + Z_i^T\gamma + U_i, \tag{2.22}$$

in which Z_i is a vector of control variables. The coefficient of interest is β, which measures the modal effect of institutions on income per capita.

The estimated results are shown in Table 2.5, where standard errors are in parentheses (bootstrap is applied with 200 replications to obtain modal regression standard errors).[22] For comparison, we report the results obtained from the proposed model in this paper, the naive modal regression, and the mean regression with 2SLS estimation. Each set of columns shows a different specification, with covariates and alternative samples that were presented in Acemoglu et al. (2001). It can be seen that the estimates from naive modal regression are obviously biased and the correction for endogeneity induces a large change in the coefficients. Compared to the results of mean regression, in most cases, the proposed modal regression indicates that there is a significantly stronger impact of institutions on income per capita. The difference between modal and mean estimates is not very large, which is due in part to the nearly symmetric data. This argument is also supported by the robustness results in the appendix.

Although the proposed modal regression also demonstrates the large effect of institutions on economic performance, there are some differences between the results of modal regression and those of mean regression in terms of other variables. Particularly, different

[22]The asymptotic limit explicitly defined in Section 2.4 cannot be directly applied in practice to calculate the variance of the modal estimator due to the existence of numerous unknown quantities. Even though these unknown terms can be estimated by the corresponding kernel estimators, we do not advocate this approach for statistical inference, which requires the introduction of additional tuning parameters. Note that, we can also make use of Bayesian inference techniques to approximate the distribution of a modal estimator.

Table 2.5: Regression with Endogeneity of Log GDP Per Capita

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
				Two-Step Modal					
Avexpr	1.0658	1.0774	1.3377	1.2741	0.5573	0.5505	1.2823	1.4948	1.0350
	(0.0207)	(0.0248)	(0.0345)	(0.0332)	(0.0222)	(0.0246)	(0.0385)	(0.0492)	(0.0134)
Lat		-0.7763		0.8154		0.0614		-1.6060	
		(0.1428)		(0.1622)		(0.1612)		(0.1530)	
Asia dummy							-1.2255	-1.4600	
							(0.0507)	(0.0562)	
Africa dummy							-0.3110	-0.2534	
							(0.0383)	(0.0408)	
"Other" dummy							-1.6909	-1.8631	
							(0.0986)	(0.1012)	
				Naive Linear Modal					
Avexpr	0.4962	0.4542	0.4689	0.4570	0.4632	0.4495	0.4050	0.3880	0.4475
	(0.0094)	(0.0092)	(0.0104)	(0.0107)	(0.0126)	(0.0133)	(0.0078)	(0.0076)	(0.0096)
Lat		1.4175		1.5936		0.3756		0.7958	
		(0.1047)		(0.1244)		(0.1265)		(0.0805)	
Asia dummy							-0.5223	-0.4755	
							(0.0301)	(0.0310)	
Africa dummy							-0.8630	-0.8330	
							(0.0227)	(0.0198)	
"Other" dummy							0.2838	0.1705	
							(0.0502)	(0.0479)	
				Mean-2SLS					
Avexpr	0.9443	0.9957	1.2812	1.2118	0.5780	0.5757	0.9822	1.1071	0.9808
	(0.1565)	(0.2217)	(0.3585)	(0.3543)	(0.0981)	(0.1173)	(0.2995)	(0.4636)	(0.1709)
Lat		-0.6472		0.9385		0.0383		-1.1782	
		(1.3351)		(1.4631)		(0.8352)		(1.7554)	
Asia dummy							-0.9242	-1.0471	
							(0.4003)	(0.5246)	
Africa dummy							-0.4643	-0.4373	
							(0.3580)	(0.4242)	
"Other" dummy							-0.9405	-0.9904	
							(0.8480)	(0.9980)	

Note: Model 1 and Model 2 are for base samples. Model 3 and Model 4 are for base samples without Neo-Europes (the USA, Canada, Australia, and New Zealand). Model 5 and Model 6 are for base samples without Africa. Model 7 and Model 8 are for base samples containing continent dummies. Model 9 is for base samples with log output per worker as the dependent variable.

from the results of mean regression, the coefficients for latitude in Model 2, Model 4, and Model 8 are significant at the 5% significance level, which is consistent with the results of many previous studies which found that latitude has a significant determinant of economic performance, but contracted to the results in Acemoglu et al. (2001). The coefficient for latitude in Model 6 is no longer significant after excluding all African countries from the sample. This suggests that considering the "most likely" effect, latitude does not have a significant effect on economic performance for countries other than those located in Africa after we take the effect of institutions into account. In addition, for Model 7 and Model 8, when we add continent dummies to the modal regressions, it changes the estimated effect of institutions, and the dummies are jointly significant at the 5% significance level. This observation is in contrast to what Acemoglu et al. (2001) concluded in their paper, indicating that based on mode effect the reason African countries are poorer is partly due to cultural or geographic factors. The above results are also supported by the robustness check in the appendix.

2.6 Penalized Modal Regression

In the first step of the proposed estimation procedure, we may have a large set of instrumental variables to use in practice and face the dimensionality curse of many instruments. To circumvent this difficulty, a theoretically optimal method for instrumental variable selection for the proposed modal regression is necessary. We in this section provide an adaptive LASSO (Zou, 2006) method to cull the weak instrumental variables to obtain more robust results. The proposed procedure can automatically eliminate the irrelevant instrumental

variables by setting the corresponding coefficients to zero, while simultaneously estimating the nonzero modal coefficients by solving the kernel-based objective function. The proposed regularization procedure is not only applicable to addressing the high-dimensional challenge in the first step, but can also be used to estimate important covariate effects and select key variables in the second step. For reasons of clarity of presentation, we concentrate on the first step estimation in this section and assume that the number of variables is fixed.

It is straightforward to generalize the adaptive LASSO method to our case of modal regression with the assumption that (2.5) is sparse,[23] where the solutions of nonpenalized modal regression are used as weights. The model framework is then specified as

$$Q(\theta) = -\frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{X_i - Z_i^{*T}\theta}{h}\right) + \lambda_n\|\hat{w}\circ\theta\|, \tag{2.23}$$

where λ_n is a nonnegative regularization parameter used to control the shrinkage of parameter estimation, $\hat{w}$ is a vector of nonnegative data-dependent weights, and $\hat{w}\circ\theta = \sum_{j=1}^{d_Z+1}\hat{w}_j|\theta_j|$ with $\circ$ denoting the Hadamard product. The resulting penalized modal estimator is denoted as $\hat{\theta}^P$. We use a subscript n to denote the dependency of λ_n on the number of observations. The values chosen for $\{\hat{w}_j\}_{j=1}^{d_Z+1}$ are crucial for guaranteeing the optimality of the solution, which is set to be $\hat{w}_j = 1/|\hat{\theta}_j|^\gamma$ for some appropriately chosen $0 < \gamma < 2$ (see Remark 2.6.21). In practice, in order to leave the intercept unpenalized, one can set $\hat{w}_1 = 0$. As long as $\hat{\theta}_j \neq 0$ for every j, the function given by (2.23) is well defined and strictly convex, allowing it to identify and estimate the nonzero effects of the instruments to obtain the predicted residual values.

[23]The proposed penalized objective function can be easily generalized to handle nonparametric modal regression in Remark 2.3.7 via basis expansion. We choose adaptive LASSO as it can possess the following three properties in the estimator: unbiasedness, sparsity, and continuity by penalizing the coefficients of different covariates at a different level by using adaptive weights. However, other penalty functions (i.e., smoothly clipped absolute deviation and minimax concave penalty) can also be applied here.

Remark 2.6.19 *The adaptive LASSO can be viewed as a generalization of the L_1 or LASSO penalty (Tibshirani, 1996). The LASSO modal estimator is obtained by the above objective function when $\gamma = 0$. Nevertheless, the LASSO penalty is known to over-penalized large coefficients and tends to be biased without possessing the oracle penalty. The adaptive LASSO, on the other hand, can reduce estimation bias by allowing for different weights to be used for different variables. Such flexibility in turn produces a relatively higher penalty for zero coefficients and a lower penalty for nonzero coefficients.*

Remark 2.6.20 *In adaptive LASSO, the weight of the zero parameter approaches infinity, while the weight of the nonzero parameter goes to a constant. We follow the tradition in mean regression to choose $\hat{\theta}$ for adaptive weight in applications and set $\gamma = 1$. Furthermore, due to the consistency of $\hat{\theta}^P$, the term $|\theta_j|/|\hat{\theta}_j|$ converges to $I(\theta_j \neq 0)$ in probability as $n \to \infty$. As a result, in the asymptotic sense, the suggested adaptive LASSO procedure can be considered as an automated implementation of best-subset selection.*

To prove the asymptotic properties of the penalized modal regression, we assume that the dataset $\{(Z_i^*, X_i)\}$ consists of n observations from the following linear model

$$X_i = Z_i^{*T}\theta + V_i = Z_{i1}^{*T}\theta_1 + Z_{i2}^{*T}\theta_2 + V_i \tag{2.24}$$

without loss of generality, where $Z_i^* = (Z_{i1}^{*T}, Z_{i2}^{*T})^T$, $\theta = (\theta_1^T, \theta_2^T)^T$, $Z_{i1}^* \in R^s$, $Z_{i2}^* \in R^{d_Z+1-s}$, and the true modal regression coefficients are $\theta_2 = \theta_{20} = 0$ and $\theta_1 = \theta_{10}$ with each component being nonzero. This means that the first s instrumental variables are relevant while the remaining $d_Z + 1 - s$ are noisy instrumental variables. We then present the asymptotic properties of the adaptive LASSO modal estimator, where we show that with fixed d_Z the

proposed method can lead to consistent instrumental variable selection, and the resulting estimator can achieve the optimal modal convergence rate when the tuning parameter is appropriately chosen. The results can be generalized to a large class of penalties as well.

Theorem 2.6.9 *Under the same conditions in Theorem 2.4.2, let* $\delta_n = h^2 + (nh^3)^{-1/2}$, $\delta_n^{-1}\lambda_n \to 0$, *and* $\delta_n^{-(\gamma-1)}\lambda_n \to \infty$, *the penalized modal estimation can correctly identify all zero elements, that is*

$$P(\hat{\theta}_2^P = 0) \to 1.$$

With fixed d_Z, the above theorem indicates that the proposed modal estimation possesses sparsity with properly chosen λ_n, that is, the true set of relevant instruments can be identified for the endogenous variable with probability tending to one. It is observed that the original adaptive LASSO requires $n^{(\gamma-1)/2}\lambda_n \to \infty$ to satisfy the sparsity property, while the proposed penalized modal estimation requires a heavier penalty $\delta_n^{-(\gamma-1)}\lambda_n \to \infty$ due to the use of mode.

Theorem 2.6.10 *With* $nh^7 = O(1)$, *under the same conditions as Theorem 2.6.9, the penalized modal estimator has the following asymptotic result*

$$\sqrt{nh^3}\left(\hat{\theta}_1^P - \theta_{10} - \frac{h^2}{2}J_1^{-1}M_1(1+o_p(1))\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dt J_1^{-1}L_1J_1^{-1}\right).$$

Furthermore, under the assumption that $nh^7 \to 0$, *we have*

$$\sqrt{nh^3}\left(\hat{\theta}_1^P - \theta_{10}\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dt J_1^{-1}L_1J_1^{-1}\right),$$

where J_1, L_1, *and* M_1 *are the* $s \times s$ *submatrices of* J, M, L *corresponding to the nonzero components of* θ.

Theorem 2.6.10 indicates that the proposed instrumental variable selection procedure enjoys the oracle property in the sense of Fan and Li (2001), that is, regardless of the choice of the shrinkage tuning parameter, the variable selection procedure is consistent and the estimators of coefficients achieve the optimal convergence rate as if the subset of true zero coefficients were already known. The proposed penalized estimator also satisfies asymptotic normality. We thus extend the oracle property of the adaptive LASSO penalty to the context of penalized modal regression, which reduces model complexity while improving model accuracy.

Remark 2.6.21 *Because the proof of the preceding theorem requires the root-nh^3 consistency of $\hat{\theta}$, it is worth noting that any root-nh^3 consistent estimator of θ_0 can be applied as the adaptive weight $\hat{w}_j$ without changing the asymptotic properties of the adaptive LASSO modal solution.*

The choice of tuning parameter λ_n is critical for variable selection, where one often proceeds by finding estimators that match a range of corresponding values and then identifying the preferred estimator using some criteria, such as various information-based criteria with maximum likelihood penalized estimators. Notice that cross validation is a common approach but is known to frequently result in overfitting, whereas Akaike information criterion (AIC)-based methods are not consistent for model selection since they may select irrelevant variables as $n \to \infty$ (Wang et al., 2007). As is typical in high-dimensional sparse modeling, we select λ_n by a Bayesian information criterion (BIC)-type procedure

$$\lambda_{n,opt} = \arg\min_{\lambda_n} BIC(\lambda_n) = -\frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{X_i - Z_i^{*T}\hat{\theta}^P}{h}\right) + \frac{\log(nh^3)}{nh^3}df_{\lambda_n} \tag{2.25}$$

considering changes in the criterion for the cases of overfitting or underfitting, where df_{λ_n} is the degrees of freedom of the fitted model (the number of nonzero coefficients of $\hat{\theta}$ for modal regression). Compared to the BIC in mean regression, we have modified the corresponding results in the penalized modal regression framework. The first term of (2.25) can be treated as an "artificial" likelihood since it exhibits certain essential properties of a parametric log-likelihood, while the second term reflects the convergence rate of the modal estimator, where the effective sample size is nh^3.

Remark 2.6.22 (**Consistency of BIC**) *To demonstrate the consistency of the BIC selection, that is, the probability of the selected model being equal to the true model asymptotically approaches one, we can follow Wang et al. (2007) to study the BIC corresponding to estimators that fail to select all of the significant variables and estimators that select too many variables. More specifically, suppose S_T denote the true model, S_λ indicate the set of the indices of the covariates selection by the penalized modal regression with tuning parameter λ, $\Omega_- = \{\lambda : S_\lambda \not\supset S_T\}$ denote the underfitted models, and $\Omega_+ = \{\lambda : S_\lambda \ / S_T\}$ represent the overfitted models. Then, one can verify that $P(\inf_{\lambda\in\Omega_-\cup\Omega_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$ under mild conditions. This means that we cannot asymptotically choose a λ that identifies an overfitted or underfitted model. Construct a sequence of reference tuning parameters $\lambda_n = \log(nh^3)/\sqrt{nh^3}$ (i.e., $\lambda_n \to 0$ and $\sqrt{nh^3}\lambda_n \to \infty$). Because the penalty modal estimator $\hat{\theta}^P_{\lambda_n}$ is exactly the same as the oracle estimator, it follows immediately that $P(BIC_{\lambda_n} = BIC_{S_T}) \to 1$. As a result, we have $P(S_{\lambda_{n,opt}} = S_T) \to 1$, indicating that if the true model is contained within the set of candidate models, it can be guaranteed to be selected by the proposed BIC criteria.*

Remark 2.6.23 (**Post-Selection Estimators**) *Belloni and Chernozhukov (2013) proposed an OLS post-LASSO estimator and showed that it outperforms the LASSO estimator in reducing asymptotic risks associated with high-dimensional sparse models. Inspired by this, after we obtain the penalized modal estimators, we can utilize them as the variable selection operators in the first step and revert back to the parametric modal regression to produce residual estimates, where we define the modal post-adaptive LASSO (BIC) estimator as*

$$\hat{\theta}^{BIC} = \arg \max_{\theta \in S_{\lambda_{n,opt}}} \frac{1}{nh} \sum_{i=1}^{n} \phi \left(\frac{X_i - Z_i^{*T}\theta}{h} \right).$$

We emphasize that the aforementioned estimator does not outperform the penalized estimator asymptotically, but the finite sample performances can be different.

To numerically solve the proposed penalized modal regression, we present a modified shooting algorithm based on Zhang and Lu (2007), where we optimize over one component of the unknown parameter vector, fixing all other components. We approximate the objective function using the Newton-Raphson update through an iterative least square procedure; see Algorithm 1. Define $Q(\theta) = \sum_{i=1}^{n} \phi_h(X_i - Z_i^*\theta)$, $\dot{Q}_j(\theta) = \partial Q(\theta)/\partial\theta_j$, and write θ as $(\theta_j, (\theta^{-j})^T)^T$, where θ^{-j} is the d_Z-dimensional vector consisting of all θ_j's other than θ_j. The modified shooting algorithm is then initialized by taking $\hat{\theta}^P = \hat{\theta}$ and letting $\lambda_{n,j} = \lambda_n/|\hat{\theta}_j|$. With diagonal kernel weight W_Z associated with $w(i \mid \theta^{(g)})$ (Algorithm 1), the gth iterative stage follows

$$\hat{\theta}_j^P = \begin{cases} \frac{\lambda_j - Q_0}{\left(Z_i^{*j}\right)^T W_Z Z_i^{*j}} & \text{if} \quad Q_0 > \lambda_{n,j} \\ \frac{-\lambda_j - Q_0}{\left(Z_i^{*j}\right)^T W_Z Z_i^{*j}} & \text{if} \quad Q_0 < -\lambda_{n,j} \\ 0 & \text{if} \quad |Q_0|\lambda_{n,j}, \end{cases} \tag{2.26}$$

where for each $j = 1, \cdots, d_Z + 1$, we set $Q_0 = \dot{Q}_j(0, \hat{\theta}_{g-1}^{-j,P})$. The penalized modal estimator can be obtained by iteratively solving the above equations. Following Fu (1998), it can be shown that the modified shooting algorithm is guaranteed to converge to the global maximizer.

Remark 2.6.24 ($\mathbf{d_Z > n}$ **Setting**) *The setting considered in this section is that the number of modal regression parameters d_Z is fixed, under which the regression parameter is sparse in the sense that many of its elements are zero. In the future, it would be interesting to investigate the case, where d_Z grows at some rate of n, i,e, $d_Z = O(n^a)$, $a > 1$. With growing d_Z, sparseness generally refers to the proportion of zero parameters. One may follow the results in Huang et al. (2008) to carefully derive conditions such that the modal estimator is oracle in the sense of having the same large sample properties as an estimator in which the zero components of the modal parameter were known a priori. Particularly, when the number of variables exceeds the number of observations, the modal estimator is not consistent and cannot be used in constructing weights $\hat{\omega}_j$. We can propose to take $\hat{\omega}_j = \min(|\hat{\hat{\theta}}_j|^{-\gamma}, n^{1/2})$ with $\hat{\hat{\theta}}_j$ being an estimator of θ_{0j} consistent with the rate $a_n \to 0$ (e.g., LASSO modal estimator). Furthermore, with the $d_Z > n$ setting, it is necessary to choose $k_n > \log(nh^3)$ to obtain model selection consistency with BIC.*

2.7 Concluding Remarks

The present paper, to the best of our knowledge, is the first work that analyzes the endogeneity issue in modal regression and systematically studies its statistical properties with the conditional mode independence restriction. In particular, we introduce a computation-

ally efficient two-step estimation procedure based on control function to estimate parametric modal regression with endogeneity, followed by a three-stage estimation method for semi-parametric partially linear modal regression with the estimated modal residual from the reduced form equation in the second step. We derive the asymptotic properties for both the parametric and nonparametric components, and show that a general linear modal regression convergence rate can be obtained for the parametric component. Under reasonable conditions, the estimation of the nonparametric component is oracle. We numerically estimate the proposed model by virtue of a modified MEM algorithm. With two Monte Carlo experiments and three empirical applications, we find the good finite sample performance of the proposed estimation procedure for addressing endogeneity in modal regression, which further indicates the importance of endogeneity correction and the practical value of the proposed estimators. The modal Euler equation derived from modal utility maximization is particularly interesting since it provides consumers with a new utility preference. The results in the Colonial Origins of Comparative Development example based on mode value reveal some differences compared to the results in Acemoglu et al. (2001), demonstrating the necessity of considering modal regression to complement the existing mean or quantile regression. We also discuss several potential model extensions in the paper, including modal-based robust estimation to achieve robustness and efficiency for symmetric data. To practically select the relevant instrumental variables, we develop an adaptive LASSO method for the proposed modal regression.

This paper provides a number of promising areas for future work. For easy illustration, we restrict the independence between the instruments and the mode value of the

structural error conditional on the reduced residual. However, this assumption may not hold in some economic settings and potentially rule out any additive functional relationship in the model; see the models of demand or supply in Kim and Petrin (2011). We can release this assumption and instead allow the conditional mode of the structural error to depend on both the reduced residual and the instruments, i.e., $Mode(U_i \mid V_i, Z_i) = m(V_i, Z_i)$. For such a nonseparable case, we may need more complicated conditions for identification and estimation. It is also of particular interest to propose a test for analyzing whether regressors are endogenous in modal regression. Different from the traditional endogeneity test in nonparametric mean or quantile regression, we can focus on testing the control function in such a way that whether $Mode(U_i \mid V_i) = 0$ almost surely. In terms of the model setting in this paper, under the null hypothesis such that there is no endogeneity in modal regression, the unknown function $m(V_i)$ will be zero, i.e., $P(Mode(U_i \mid V_i) = 0) = 1$, while under the alternative hypothesis, we have $P(Mode(U_i \mid V_i) = 0) < 1$. All of these will be kept for future research.

Chapter 3

Modal Volatility Function with Variance Reduction

3.1 Introduction

Conditional volatility estimation is of interest in its own right, which is crucial for statistical inference and plays as the key intermediate step in the estimation of economic or financial quantities in practice, such as the analysis of growth curves, asset pricing practice, or applications where the second moment is treated as a proxy for risk. During the last four decades, there has been a substantial amount of literature concerning the estimation of the volatility function in discrete time by taking the conditional variance derived from the squared residuals as a latent variable without imposing any restrictive assumptions on a parametric model. To quote a few of them, Pagan and Ullah (1988) proposed a Nadaraya-Watson estimator of conditional variance to capture the relationship between volatility and

economic factors; Fan and Yao (1998) suggested a local linear method for estimating the conditional variance of a two-dimensional strictly stationary and absolutely regular process; Ziegelmann (2002) used a nonparametric local exponential method to estimate the volatility function to ensure nonnegativity; Yu and Jones (2004) introduced a likelihood-based local linear estimation of the conditional variance function to incorporate the positivity of variance; Ziegelmann (2008) developed a least-absolute-deviations estimator of the conditional variance function; and Mishra et al. (2010) proposed a combined semiparametric estimator to include the parametric and nonparametric estimators of the conditional variance. For other literature, we refer to the review paper written by Su et al. (2012) and the references cited therein. However, all of the papers mentioned above are considered from either the mean or median (quantile) regression estimator. When there are several outliers in the data or the data is skewed, resulting in non-normally estimated standardized residuals, which is a common characteristic of financial time series data, traditional nonparametric mean estimators may lose robustness or have misspecification. Although the median (quantile) estimator is resistant to outliers, it cannot directly exhibit how the "most likely" value of volatility is affected by data (the median estimator will also lose efficiency with normally distributed data). For example, in the stock market, an investor may like to know more about mode risk as opposed to mean or median risk. In addition, with fat-tailed distributions, it is not guaranteed that certain moments of error will exist when errors are leptokurtic. Large values might easily be due to the fat-tailed nature of the data and should not be attributed entirely to increases in variance. As a consequence, it is important to construct a volatility estimator that automatically adapts to skewed/tailed data.

To reveal the whole characteristics of the volatility of data and complement the existing mean or median volatility estimator, we propose a novel modal estimator for the volatility function in a nonparametric heteroskedastic regression model by directly imposing model assumptions on the conditional mode of volatility given covariate, which is not the conditional variance. In the usual mean regression model, the volatility/variance function is treated as the variance such that $\sigma^2(X_t) = E\{(\sigma(X_t)\varepsilon_t - E(\sigma(X_t)\varepsilon_t)) \mid X_t\}^2$ (see model (3.1) for the meaning of each term) with conditions $E(\varepsilon_t \mid X_t) = 0$ and $E(\varepsilon_t^2 \mid X_t) = 1$. Analogous to this, we define *modal volatility function* in this paper as

$$\sigma^2(X_t) = Mode\{(\sigma(X_t)\varepsilon_t - E(\sigma(X_t)\varepsilon_t)) \mid X_t\}^2$$

with the conditions $E(\varepsilon_t \mid X_t) = 0$ and $Mode(\varepsilon_t^2 \mid X_t) = 1$, where $Mode(\cdot \mid \cdot)$ denotes the conditional mode value. This modal volatility, which measures the *mode* risk associated with mean prediction, is of interest since it provides a plausible linkage between risk in mode sense and expected return on financial assets. Such a modal volatility function is not identical to the traditional variance function under heavy-tailed/skewed data circumstances, especially in the case of infinite variance, where the mean estimate does not exist (i.e., Cauchy distribution). Instead, it is a more general "scale" measure, which is the primary reason we refer to it as the volatility function; see Section 3.2 for more details.

In comparison to the mean, the mode is a significant numerical feature of the dataset when the data have outliers or heavy-tailed/skewed distributions. It has the essential virtue of being resistant to distributional assumptions and making no prior conditions about the symmetry of the innovation process. Such properties are particularly appealing for financial applications, because it is well-accepted that financial time series data, such as

portfolio returns and log returns, always exhibit a heavy-tailed and asymmetrical marginal distribution, making mode a suitable indicator for releasing its feature. According to Ullah et al. (2021, 2022), modal regression can be applicable to truncated or skewed data, present new feature selection tools, provide better point prediction and shorter prediction intervals, and reveal heterogeneous and clustering structures among the data (allowing the existence of local modes). Furthermore, when the data are symmetrically distributed, where the modal regression line is identical to the mean regression line, modal regression can overcome the shortcoming of lack of robustness of mean regression to achieve robust estimators by adjusting the bandwidth values;[1] see the related discussions in Remark 3.2.33. These attractive features make the study of modal regression considerably important. Recently, there has been an increasing interest in applying modal regression to investigate data characteristics; see Lee (1989, 1993), Kemp and Santos Silva (2012), Yao and Li (2014), Chen et al. (2016), Yao and Xiang (2016), Zhou and Huang (2016), Krief (2017), Chen (2018), Li and Huang (2019), Ota et al. (2019), Kemp et al. (2020), Ullah et al. (2021, 2022), among others. Built on the aforementioned work, we are interested in applying nonparametric modal regression on the volatility function for dependent samples to understand the local variability of data from the perspective of the "most likely" value, where the exact parametric forms of the mean regression function and volatility function are not predefined but both are assumed

[1]We in this paper concentrate on asymmetric data in order to capture the "most likely" effect. However, it is well-known that the mode is identical to the mean for symmetric data, and the mode is resistant to outliers and heavy-tailed distributions. Thus, compared to the mean estimator, the modal-based estimator can achieve robustness and efficiency. We also discuss such a case in the paper and show that the presented modal-based volatility estimator for dependent data has the same asymptomatic bias and variance as the corresponding estimator for independent data under regularity conditions. Such an equivalence is important because it allows extensions of efficiency arguments along the lines of those of Yao et al. (2012) to our *modal-based robust volatility estimator*. Specifically, the developed modal-based volatility estimator could be more efficient than the mean volatility estimator if the data contain outliers or have a heavy-tailed distribution. The two estimators would share almost the same efficiency if the data indeed have a normal distribution. This should be considered as a big advantage compared to other robust (median and maximum likelihood-type) estimators, which will sacrifice efficiency with normally distributed data.

to be smooth.[2] To our best knowledge, there is no literature that investigates the volatility of data from the mode perspective for dependent samples.

Specifically, we aim at applying local linear modal regression to estimate the volatility function under stationary α-mixing dependent samples and settle theoretical properties rigorously. The majority of volatility function research has concentrated on the case of independent and identically distributed (i.i.d.) data, which is not necessarily valid in empirical applications. There are numerous economic analysis problems involving high-dimensional data or information network data, in which the data exhibit some kind of dependence, such as Markovian chains, mixing sequences, long-range memory process, among others (Ullah et al., 2022). In such cases, the statistical properties of the volatility estimators presented in the papers considering i.i.d. samples may be changed. We noticed that Wang and Tang (2016) investigated robust estimators by applying local M-estimation for conditional variance in heteroscedastic regression models and utilizing Huber's function for estimation with dependent samples, which still belongs to the mean regression estimator. This robust estimator is quite similar to the modal-based robust estimator considered in our paper; see Remark 3.2.33. However, we can theoretically show that the modal-based robust estimator is more efficient than the M-estimator. As far as we know, no attempt has been made in the existing literature to apply nonparametric regression to estimate the modal (and modal-based) volatility function for dependent samples. To fill this literature

[2]Mode is defined as the most frequent data point in a dataset, which can be achieved by maximizing a conditional distribution $f_{Y|X}(Y \mid X)$, in which Y is the dependent variable and X are covariates. In practice, we have to utilize nonparametric density estimation for $f_{Y|X}(Y \mid X)$ since the actual density function is unknown; see Chen et al. (2016). However, such nonparametric density estimation is difficult to implement with high-dimensional data. To deal with this issue, Kemp and Santos Silva (2012) and Yao and Li (2014) suggested a kernel-based objective function for estimating the modal coefficient in a parametric modal regression. In this paper, we prefer to choose nonparametric modal regression for its flexibility. We can alternatively apply parametric modal regression based upon the GARCH family of models to estimate the volatility function by imposing mode restrictions. Nevertheless, without theoretical reasons to identify the model format, parametric modal regression is prone to misspecification.

gap, we specify the dependence framework and divide the proposed estimation procedure into two steps, where in the first step, we use local linear mean regression to obtain the estimated squared residuals;[3] then in the second step, we apply local linear modal regression on the mean squared residuals to achieve the modal volatility estimator. Generally, the resulting modal estimator of volatility will carry additional bias and variance due to the first step estimation. Nevertheless, under regularity conditions, we show that the novel modal volatility estimator is fully regression-adaptive in the sense that we can estimate the volatility function asymptotically as well as if the mean regression function were known, implying that there is no loss in asymptotic efficiency due to the estimation of the unknown mean regression function. A similar "adaptive" phenomenon has been observed in the mean volatility estimator (Fan and Yao, 1998). In theoretical terms, compared to the local linear mean volatility estimator, the modal volatility estimator has completely different asymptotic properties and would have a slower convergence rate, which is the cost we need to pay in order to estimate mode (Parzen, 1962). Thus, estimating the unknown mean regression function (with a faster convergence rate) no longer has any noticeable effect on estimating modal volatility, indicating that we do not need to undersmooth the mean regression function in the first step to obtain a regression-adaptive modal volatility estimator in the second step.

We also show a new and interesting result that the asymptotic theorem for the proposed modal volatility estimator for stationary α-mixing dependent samples is the same as that for independent samples under some mild conditions, which is intrinsic in nonparamet-

[3]In the context of financial time series, more attention is placed on the volatility function $\sigma(\cdot)$ rather than on the mean function $m(\cdot)$. However, $m(\cdot)$ is not unimportant and cannot simply be set to zero; see the discussion of impacts of $m(\cdot)$ on the volatility estimator in Remark 3.2.29.

ric estimation for dependent samples; see Cai and Ould-Said (2003). The asymptotic mean squared error (MSE) expression for modal volatility estimation follows from the asymptotic theorem in the usual way. By minimizing MSE, the expression of the optimal bandwidths is derived, which offers some guidance regarding how bandwidths should be chosen in practice. To numerically estimate the proposed modal regression in the second step, we develop a computationally attractive MEM algorithm based on Li et al. (2007) and Yao (2013) and suggest a data-based bandwidth selection method for practically choosing bandwidths. Although the idea of this volatility function estimator is not very new in nonparametric kernel estimation, the application of nonparametric modal regression to the newly defined modal volatility function is novel. To avoid the "curse of dimensionality" issue in the nonparametric literature, we focus on the univariate predictor case throughout the paper. However, all of the methods introduced in this paper can be easily extended to the multivariate dimensional case at the price of more complicated expressions and discussions.[4]

Compared to parametric estimators, when the dimension of the covariates increases, the variance of nonparametric estimators will increase as well due to the slower rate of convergence. Also, as opposed to the mean estimator, the modal volatility estimator has a slower rate of convergence due to the fact that only a small portion of total observations around the mode are used. Therefore, reducing the variance of the nonparametric modal estimator becomes very essential and attractive. There exists a large number of research investigating variance reduction issues in both theory and practice. One of which has

[4]If the proposed model is extended to the multivariate case, i.e., by including several period lags in the conditional modal volatility equation, we may encounter the "curse of dimensionality" issue. In such a case, one can restrict the form of the conditional mean and volatility functions to a lower dimension without losing information to avoid the issue of the "curse of dimensionality". For example, we can choose a single index, a varying coefficient, or an additive partial linear structure for estimating, which is an interesting research direction that is beyond the scope of this paper, and so we leave it for future research.

played an important role is the variance reduction technique in nonparametric smoothing proposed by Cheng et al. (2007) through forming a linear combination of a preliminary estimator evaluated at nearby points. They demonstrated that the asymptotic MSE of the estimator is improved considerably after variance reduction, and the amount of reduction is uniform across different locations, regression functions, designs, and error distributions. Since then, a large number of researchers have devoted effort to applying the same quadratic interpolation method to reduce the variance of estimators. For example, Cheng and Peng (2007) applied the variance reduction technique to multiparameter likelihood models to improve the efficiency of the estimator and Chen et al. (2009) proposed a new efficient method for estimating the conditional variance in heteroscedasticity regression models and applied the variance reduction technique to improve the inference for the conditional variance. After we obtain the modal volatility estimator, we generalize the variance reduction technique introduced in Cheng et al. (2007) for modal regression to obtain a new *variance reduced modal volatility estimator*, which has asymptotic relative efficiency by taking a linear combination of the previous modal volatility estimator at three equally spaced points around the point of estimation. Theoretical results under mild conditions indicate that the asymptotic variance of the modal volatility estimator can be reduced by a known factor, while the asymptotic bias remains unchanged by forcing the coefficients in the linear combination to fulfill the corresponding moment conditions, which leads directly to a reduction in asymptotic MSE. We further examine the variance reduced modal volatility estimator in the aspect of bandwidth selection, where we show that the asymptotic optimal bandwidth can be achieved by a simple constant factor adjustment of that from the local linear modal volatility estimator.

Consequently, the bandwidth in practice can be obtained straightforwardly by utilizing the corresponding bandwidth values found in the previous modal volatility estimator.

It is worth pointing out that although the local linear modal volatility estimator is attractive, it cannot always be ensured to be nonnegative in finite samples.[5] To avoid this drawback and potentially improve the accuracy of the bias term, we in the end generalize the proposed method to local exponential modal estimation, which can guarantee the positivity of the volatility function by extending the results in Ziegelmann (2002), Yu and Jones (2004), and Mishra et al. (2010). This exponential modal volatility estimator is not directly equivalent to the local linear modal volatility estimator, but rather estimates the logarithm of the volatility, thereby introducing an extra bias term. The provided theoretical results show that the difference between the local linear and exponential modal volatility estimators lies in the form of the asymptotic bias. Similar to the proposed local linear modal volatility estimator, the exponential modal volatility estimator is asymptotically fully adaptive to the unknown conditional mean regression function, i.e., its asymptotic property is not sensitive to how well the conditional mean regression function is estimated. More generally, we show that under certain conditions, the exponential modal volatility estimator can have a smaller bias compared to the local linear modal volatility estimator and achieve a modal parametric convergence rate. In this regard, the present paper also makes a contribution to bias reduction for the modal estimator without any pilot parametric guide to capture some roughness features of the unknown volatility function, which differs significantly from the combined-estimation method in Mishra et al. (2010).

[5]Negative values for modal volatility in a real application indicate that the estimate of mean regression in the first step might suffer from overfitting, because negative values can occur particularly in regions in which data are sparse. It has been observed that the number of negative values decreases as sample size increases, which means that the negative volatility estimates will not be a problem asymptotically.

The structure of this paper is as follows. In Section 3.2, we investigate the modal regression estimator for the volatility function under stationary α-mixing dependent samples. We present the asymptotic distributional theory for the resulting estimator under some mild conditions, which provides guidelines for selecting reliable bandwidths in practice. We also briefly discuss the modal-based robust volatility estimator for the sake of completeness. Section 3.3 applies the variance reduction technique to improve the estimation of the modal volatility function. Asymptotic properties and optimal bandwidths are provided. Section 3.4 contains the results of finite sample numerical studies, including two simulation studies and two analyses of real datasets—Interest rate dataset and Motorcycle dataset. To avoid negative values in the volatility function estimates, we in Section 3.5 extend the proposed method to the local exponential modal volatility estimator and conduct some theoretical comparisons with the suggested local linear modal volatility estimator. We conclude the paper and present some potential future research in Section 3.6. The additional numeral results as well as all technical proofs of the theoretical results are given in the appendix.

3.2 Modal Volatility Estimator

We in this section introduce a two-step residual-based procedure for estimating the modal volatility function under a heteroskedastic regression model, where in the first step the nonparametric estimation of the conditional mean regression function is obtained; then in the second step, the local linear modal estimation is applied to the squared residuals for volatility estimation. Under the assumption of a certain degree of smoothness on the modal regression function, we present the asymptotic properties and derive the optimal bandwidths. The

adaptiveness property to the unknown conditional mean regression Fan and Yao (1998) established is also shared by the proposed local linear modal volatility estimator.

3.2.1 Local Linear Modal Estimation

Let $\{(Y_t, X_t)\}_{t=1}^n$ be a two dimensional strictly stationary process having the same marginal distribution as $(Y, X) \in R^2$, where $X_t \in R$ is $\mathcal{F}_{t-1}$ measurable and can be the lag variable of $Y_t \in R$, and $\mathcal{F}_{t-1}$ is the σ-algebra of events generated by $\{X_k\}_{k=-\infty}^t$. Naturally, this includes the scenario in which $\{(Y_t, X_t)\}_{t=1}^n$ are i.i.d.. Let $m(X_t) = E(Y_t \mid X_t)$ be the conditional mean regression function (location function) based on the past information and $\sigma^2(X_t) > 0$ ($\sigma(X_t) > 0, \forall X_t \in R$) for all X_t denote the volatility function (scale function) of the stochastic process Y_t depending on covariate, which are left unspecified and are the subjects of statistical investigation in this paper.[6] We then write the nonparametric heteroskedastic regression model as

$$Y_t = m(X_t) + \sigma(X_t)\varepsilon_t,\ t = 1, \cdots, n, \tag{3.1}$$

where $\{\varepsilon_t\}_{t=1}^n$ is a sequence of stochastic random variables with $E(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0$ and $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = 1$. There are some notable features of the preceding model. To begin, in contrast to most research, which assumes that $E|\varepsilon_t|^{4+\delta} < \infty$ for some $\delta > 0$, we do not need to impose high moment conditions on ε_t, comparable to quantile regression. When $X_t = Y_{t-1}$, (3.1) would include the nonparametric autoregressive conditional heteroskedastic (ARCH) time series model (Engle, 1982) and AR(1) process with ARCH(1) errors as special cases, indicating that both static and dynamic models are covered. If we let $Y_t = X_{t+1} - X_t$

[6] We in this paper assume that volatility depends on covariate (heteroskedasticity). If $\sigma^2(\cdot)$ is constant, the model becomes homoskedastic. We can then apply the nonparametric kernel density estimation method to obtain the corresponding modal volatility value. In particular, we maximize $\frac{1}{nh}\sum_{i=1}^n K\left(\frac{\hat{r}_t - a}{h}\right)$, where h is a bandwidth depending on sample sizes, $\hat{r}_t$ is defined in the paper, and $K(\cdot)$ is a kernel function. The numerical solution can be obtained by utilizing the mean-shit algorithm in Chen et al. (2016).

and assume that ε_t are standard normals, then the model can be viewed as the discretized version of the stochastic diffusion model $dX_t = m(X_t)dt + \sigma(X_t)dW_t$, where $\{W_t\}$ is a standard Brownian motion, including the geometric Brownian motion as in a stock price model in option pricing and the Vasicek model for interest rates. Notice that (3.1) is also of essential importance in financial econometrics, since it has the capability of accounting for nonlinearity and conditional heteroskedasticity in the modeling of financial time series.

On the basis of the above settings, we can obtain

$$r_t = \sigma^2(X_t)\varepsilon_t^2 = (Y_t - m(X_t))^2, \tag{3.2}$$

which is the primary focus of this paper. Instead of imposing moment restrictions, if the mean regression function $m(X_t)$ is known, the volatility function could be estimated directly from a modal regression of the squared residual r_t on X_t due to the fact that

$$Mode(r_t \mid X_t) = \sigma^2(X_t) = Mode((Y_t - m(X_t))^2 \mid X_t), \tag{3.3}$$

i.e., we replace $\hat{r}_t$ in the objective function (3.6) with r_t for estimating. We emphasize that studying modal volatility estimation is necessary to supplement the existing mean variance function because the mode of an arbitrary random variable is always (a) finite number(s). Especially, when the error terms are Cauchy random variables or other random variables with heavy-tailed distributions, the mean regression might be inapplicable, as the variance may not exist. Accordingly, the modal volatility provides a more natural dispersion measure than the mean variance for the non-Gaussian case.

Remark 3.2.25 *The model settings in this paper can be considered as a generalization of the Qualitative Threshold ARCH model proposed in Gouriéroux and Monfort (1992), but with a focus on modal estimation, where* $Y_t = \sum_{l=1}^{J} \alpha_j I(Y_{t-l} \in A_j) + \sum_{l=1}^{J} \beta_j I(Y_{t-l} \in A_j)\varepsilon_t$,

$\{A_j\}_{j=1}^{J}$ with fixed J denotes a partition of the set of values for Y, $I(\cdot)$ represents indicator function, and (α_j, β_j) are unknown parameter vectors and matrices, respectively. In addition, the assumption $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = 1$ distinguishes our model from existing volatility models that impose the conditional mean or median assumption. It is not strict to enforce such a conditional mode restriction. A similar median condition for the nonparametric least-absolute-deviations estimator of the volatility function has been utilized in Ziegelmann (2002), such that $Median(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = 1$.

Remark 3.2.26 *If we are intend to estimate $\sigma(X_t)\varepsilon_t$ instead of $\sigma(X_t)$, as in the ARCH case investigated in Koenker and Zhao (1996), it is not necessary to impose the condition on $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1})$. Also, if we are interested in $\sigma^2(x)$ when $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1})$ is unknown and $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) \neq 0$, we must specify the form of the scale function $\sigma(\cdot)$, otherwise, the model is unidentifiable. In this case, an estimate of the mode value of ε_t is needed in order to recover the estimate of $\sigma(\cdot)$. In practice, we can estimate ε_t using the residuals $(Y_t - \hat{m}(X_t))/\sigma'(X_t)$, where $\hat{m}(\cdot)$ is from (3.4) and $\sigma'(\cdot)$ is an estimate of $\sigma(\cdot)$; see Akritas and Van Keilegom (2001) for the detailed method.*

In reality, the value of the conditional mean regression function $m(X_t)$ is unknown, indicating that $m(\cdot)$ plays the role of a nuisance parameter that must be estimated first. A natural approach is to apply the nonparametric regression estimator, in which the local linear estimation technique is utilized to cope with $m(X_t)$, though it will be clear that the estimation can be generalized to any linear smoother (e.g., polynomial regression, smoothing splines, and wavelet). Suppose that the second derivative of $m(\cdot)$ is continuous in the domain of X, x is a given point in the domain, and Y_t is in the support of the conditional density of

Y, based on Taylor expansion, we then have $m(X) \approx m(x)+m^{(1)}(x)(X-x)$ for X in a local neighborhood of x and $m^{(1)}(x)$ is the first derivative of $m(\cdot)$ with respect to x. Throughout the paper, the notation $A \approx B$ represents $A = B(1+o(1))$. After that, we get the following least squares problem

$$(\hat{a}(x), \hat{b}(x)) = \arg \min_{a(x),b(x)} \sum_{t=1}^{n} \{Y_t - a(x) - b(x)\,(X_t - x)\}^2 K\left(\frac{X_t - x}{h}\right), \qquad (3.4)$$

where $a(x) = m(x)$, $b(x) = m^{(1)}(x)$, $K(\cdot) : R \to R$ is a bounded and symmetric kernel function, and $h = h(n) > 0$ is a sequence of positive numbers tending to zero as $n \to \infty$, which is referred to as bandwidth. The kernel $K(\cdot)$ and the bandwidth h determine the shape and width of the local neighborhood. Defining $\hat{a}(x)$ and $\hat{b}(x)$ as estimators from (3.4), the local linear estimators of $m(\cdot)$ and $m^{(1)}(\cdot)$ are simply

$$\hat{m}(x) = \hat{a}(x) = \frac{T_{n,0}S_{n,2} - T_{n,1}S_{n,1}}{S_{n,2}S_{n,0} - S_{n,1}S_{n,1}} \text{ and } \hat{m}^{(1)}(x) = \hat{b}(x) = \frac{T_{n,1}S_{n,0} - T_{n,0}S_{n,1}}{S_{n,2}S_{n,0} - S_{n,1}S_{n,1}},$$

respectively, where $S_{n,l_1} = \sum_{t=1}^{n} K_h\,(X_t - x)\,(X_t - x)^{l_1}$, $l_1 = 0, 1, 2$, $T_{n,l_2} = \sum_{t=1}^{n} K_h(X_t - x)\,(X_t - x)^{l_2}\,Y_t$, $l_2 = 0, 1$, and $K_h(\cdot) = K(\cdot/h)/h$. The asymptotic bias and variance results for $\hat{m}(x)$ are standard and can be established under certain regularity conditions (Fan and Gijbels, 1996). In particular, the asymptotic bias is

$$\text{Bias}\{\hat{m}(x)\} = \frac{1}{2}h^2 m^{(2)}(x) \int u^2 K(u)du + o_p(h^2 + (nh)^{-1/2})$$

with $m^{(2)}(\cdot)$ denoting the second derivative of $m(\cdot)$, and the asymptotic variance is

$$\text{Var}\{\hat{m}(x)\} = \frac{\text{Var}\,(Y \mid X = x)}{nhf_X(x)} \int K^2(u)du + o_p((nh)^{-1} + h^4),$$

where $f_X(\cdot)$ represents the density of X. It is worth pointing out that the first step estimation error in $\hat{m}(\cdot)$ must be controlled in the asymptotic analysis of the proposed modal

volatility estimator. Especially, we need to impose the high-level condition that the chosen mean estimator satisfies $|m(x) - \hat{m}(x)| = O_p(\vartheta_n)$ for some rate parameter $\vartheta_n \to 0$. Nonetheless, because the convergence rate of the mean estimator is faster than that of the modal estimator, we can take the results from the first step mean estimation without further processing.

Remark 3.2.27 *We can also obtain the estimate of r_t by directly estimating the conditional expectation of squared responses and setting $\hat{r}_t = \hat{g}(X_t) - \hat{m}^2(X_t)$, where $\hat{g}(X_t)$ is the nonparametric kernel-type estimate of $E(Y_t^2 \mid X_t)$. The same mean convergence rate as the one from* (3.4) *is exhibited in this estimator. However, as pointed out by Fan and Yao (1998), such a direct estimation method may result in a very large bias, and the estimator is not asymptotically design adaptive to the estimation of $m(\cdot)$. It may also produce a negative estimate of the volatility function, particularly if different smoothing parameters are utilized.*

After obtaining $\hat{m}(x)$, we consider a residual-based modal estimator of the conditional volatility, in which the volatility function is estimated by performing a modal regression of the estimated squared residuals $\hat{r}_t = (Y_t - \hat{m}(X_t))^2$ against X_t. In order to take advantage of the modal estimator and ease the structural assumptions in parametric models, instead of fitting the squared residuals against X_t via mean or parametric regression, we estimate the volatility function $\sigma^2(X_t)$ by employing Taylor expansion, such that

$$Mode(r_t \mid X_t) = \sigma^2(X_t) \approx \sigma^2(x) + (\partial\sigma^2(x)/\partial x)(X_t - x) \tag{3.5}$$

for X_t in a local neighborhood of x. After that, we can obtain the following objective function[7]

[7]The maximum of the kernel-based objective for estimating modal coefficients is originated from non-

$$Q_n(\alpha_1(x), \alpha_2(x)) = \frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi\left(\frac{\hat{r}_t - \alpha_1(x) - \alpha_2(x)(X_t - x)}{h_1}\right)K\left(\frac{X_t - x}{h_2}\right), \qquad (3.6)$$

where $\alpha_1(x) = \sigma^2(x)$, $\alpha_2(x) = \partial\sigma^2(x)/\partial x$, $\phi(\cdot) : R \to R$ is a standard kernel density with bandwidth $h_1 = h_1(n)$ (control the number of estimated modes) such that $h_1 \to 0$ as $n \to \infty$, and $K(\cdot) : R \to R$ is a rescaled kernel function associated with bandwidth $h_2 = h_2(n)$ (control the distance between observations) such that $h_2 \to 0$ as $n \to \infty$. With bounded kernel functions, the bandwidths h_1 and h_2 play an important role in estimating dependent observations because the dependency can be controlled with observations in a small window; see the discussions associated with Theorem 3.2.12. Although we can utilize different kernels for the mean regression function and the volatility function, we choose to use the same kernel $K(\cdot)$ here for simplicity. We cannot, however, employ the same bandwidths for mean and modal regressions due to differences in convergence rates. The aforementioned objective function then leads to the residual-based modal estimators $\hat{\sigma}^2(x) = \hat{\alpha}_1(x)$ and $\hat{\dot{\sigma}}^2(x) = \hat{\alpha}_2(x)$.

Remark 3.2.28 *Local linear approximation is popular in many contexts due to the advantages in boundary behavior (i.e., the boundary adjustment is not necessary) and estimating regression derivatives (Fan and Gijbels, 1996). Nonetheless, as previously discussed, such an estimate cannot guarantee the positive value of volatility in finite samples. To avoid negative values in practice, in addition to the exponential estimate developed in Section 3.5, we can apply the theoretically less satisfactory local constant modal estimator*

parametric density estimation, where $\phi(\cdot)$ is used to target the mode value of the error term and $K(\cdot)$ is applied to control the smoothness of the modal volatility function. We can also interpret modal estimation by utilizing a distance-based loss function such as $h_1^{-1}(1 - \phi(Y - X\beta)h_1^{-1})$ and $Mode(Y \mid X) = X\beta$. We will need an additional kernel function $K(\cdot)$ for nonparametric modal regression. As stated in Yao et al. (2012), the choice of kernel functions is not very crucial empirically or theoretically compared to the choice of bandwidths. We choose the standard normal kernel for all kernels in this paper to make computation easier.

$$\hat{\alpha}_1 = \arg\max_{\alpha_1} \frac{1}{nh_1h_2} \sum_{t=1}^{n} \phi\left(\frac{\hat{r}_t - \alpha_1}{h_1}\right) K\left(\frac{X_t - x}{h_2}\right),$$

where the estimation algorithm (or using the mean-shift algorithm) and asymptotic theorems shown below are still valid but with a little bit of notation change.

Remark 3.2.29 *The developed methodology provides a broadly applicable framework for constructing different volatility estimators across a wide range of diverse settings. The volatility defined in this paper is consistent with the classical definition in mean case, which is to measure variability from the average or mean but in mode sense. If we consider employing modal estimators in both stages, i.e., to estimate both the location function and the volatility function, we then need to slightly change the model settings without imposing any restrictions on moment conditions. To be more specific, let* $\{(Y_t, X_t)\}$ *be a two-dimensional strictly stationary process, having the same marginal distribution as* (Y, X)*. Then, we can write* $Y_t = m(X_t) + \sigma(X_t)\varepsilon_t$*, where* $Mode(\varepsilon_t \mid \mathcal{F}_{t-1}) = 0$*,* $Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = 1$*,* $Mode(Y_t \mid X_t) = m(X_t)$*, and*

$$\sigma^2(X_t) = Mode\{[Y_t - Mode(Y_t \mid X_t)]^2 \mid X_t\} > 0.$$

In such a case, the defined modal volatility is to measure variability from the mode over a given period of time in mode sense. The properties of $Mode(Y_t \mid X_t) = m(X_t)$ *can be established following the procedures in this paper. The undersmoothing in the first step, on the other hand, is required to achieve the adaptiveness property of the modal volatility estimator.*

Unlike local linear mean regression, there is no closed-form expression of the maximizers of (3.6), so the modal-optimal estimator should be found using numerical optimization techniques. We apply a modified MEM algorithm (Algorithm 2) including expectation

step (E-Step) and maximization step (M-Step) to solve it (Li et al., 2007; Yao, 2013). Because of the usage of a Gaussian kernel, the estimator in M-Step with log-maximization bears a close resemblance to the least squares estimator. Consistent with the conventional

Algorithm 2 MEM Volatility Algorithm

E-Step. Calculate the weight $\pi\left(t \mid \alpha_1^{(g)}, \alpha_2^{(g)}\right), t = 1, \cdots, n$, with the preliminary estimates of the modal parameters as

$$\begin{aligned}\pi\left(t \mid \alpha_1^{(g)}, \alpha_2^{(g)}\right) &= \frac{\phi\left(\frac{\hat{r}_t - \alpha_1^{(g)} - \alpha_2^{(g)}(X_t - x)}{h_1}\right) K\left(\frac{X_t - x}{h_2}\right)}{\sum_{t=1}^{n} \phi\left(\frac{\hat{r}_t - \alpha_1^{(g)} - \alpha_2^{(g)}(X_t - x)}{h_1}\right) K\left(\frac{X_t - x}{h_2}\right)} \\ &\propto \phi\left(\frac{\hat{r}_t - \alpha_1^{(g)} - \alpha_2^{(g)}(X_t - x)}{h_1}\right) K\left(\frac{X_t - x}{h_2}\right).\end{aligned}$$

M-Step. Update $\left(\alpha_1^{(g+1)}, \alpha_2^{(g+1)}\right)$ with the weight calculated in the E-Step

$$\begin{aligned}&\left(\alpha_1^{(g+1)}, \alpha_2^{(g+1)}\right) \\ =&\arg\max_{\alpha_1, \alpha_2} \sum_{t=1}^{n} \left\{\pi\left(t \mid \alpha_1^{(g)}, \alpha_2^{(g)}\right) \log \frac{1}{h_1} \phi\left(\frac{\hat{r}_t - \alpha_1 - \alpha_2(X_t - x)}{h_1}\right)\right\} \\ =&(X^* W_X X^*)^{-1} X^{*T} W_X \hat{r},\end{aligned}$$

where g is the iteration indicator, $X^* = (X_1^*, \cdots, X_n^*)^T$ with $X_t^* = (1, X_t - x)$, $\hat{r} = (\hat{r}_1, \cdots, \hat{r}_n)^T$, and W_X is an $n \times n$ diagonal matrix with $\pi(t \mid \alpha_1^{(g)}, \alpha_2^{(g)})$ as diagonal elements.

Iterate. Given the initial values, iterate E-Step and M-Step repeatedly until a stopping criteria like $\|\kappa^{(g+1)} - \kappa^{(g)}\| < \eta$ for some $\eta > 0$ with some approximate norm $\|\cdot\|$ is satisfied, where $\kappa^{(g+1)} = (\alpha_1^{(g+1)}, \alpha_2^{(g+1)})$.

Note: For the purpose of simplicity, we suppress x for $\alpha_1(x)$ and $\alpha_2(x)$ in the algorithm.

EM algorithm, the kernel-based objective function is not decreased after a MEM iteration as a result of the convergence property, i.e., $Q_n(\alpha_1^{(g+1)}, \alpha_2^{(g+1)}) \geq Q_n(\alpha_1^{(g)}, \alpha_2^{(g)})$. According to Yao and Li (2014) and Ullah et al. (2021), it cannot be guaranteed that the MEM algorithm will converge to the global maximizer. Therefore, it is much important to re-start the algorithm with different starting points (i.e., local linear mean or quantile estimates) to choose the best optimal value by comparing the value of the objective function. The algorithm may also diverge owing to numerical instability, which can be caused, for example, by a poor selection of the bandwidths (h_1 and h_2). Consequently, a maximum number of iterations that the algorithm may be allowed to run must be specified. Finally, we emphasize that the precondition for the MEM algorithm to obtain modal estimates is to let $h_1 \to 0$; by contrast, allowing $h_1 \to \infty$ will result in the MEM algorithm producing mean estimates; see Remark 3.2.33.

Remark 3.2.30 *The major difference between the mean regression by lease squares and the modal regression by MEM algorithm lies in the weight* $\pi(t \mid \alpha_1^{(g)}, \alpha_2^{(g)})$ *used in the E-Step. For the mean regression, each observation is given an equal weight* $1/n$*, whereas the weight* $\pi(t \mid \alpha_1^{(g)}, \alpha_2^{(g)})$ *depending on current estimates allows modal regression to reduce the effect of observations far away from the modal regression curve to achieve robustness, which is one of the advantages of modal regression over mean regression.*

Remark 3.2.31 (Modal-Based Robust Volatility Estimator) *We in this paper investigate the modal volatility estimator on the basis of the premise that the mode is not identical to the mean (otherwise the proposed estimation procedure is not the most efficient one). When data are symmetrically distributed, the modal regression line shall be identical*

to the mean regression line. In addition, as a central tendency measure, the mode is robust to outliers and abnormal observations (due to the requirement of no moment conditions), whereas the least squares estimate method is not the most effective one in the presence of outliers or heavy tails. In many areas of economics and finance, empirical studies have disclosed heavy-tailed distributions. The tails of high-frequency financial time series, for example, may be significantly heavier than those of Student-t distributions. Borkovec and Klüppelberg (2001) revealed that the stationary distribution of Y_t may have a heavy tail of the Pareto type and hence Y_t may not have a finite second moment despite the fact that $E(\varepsilon_t^2) < \infty$. These phenomena indicate that modal regression can be treated as an alternative way to obtain the nonparametric robust estimation of the conditional variance function, which can be further used to construct confidence intervals for the mean function. If we assume

$$E(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = Mode(\varepsilon_t^2 \mid \mathcal{F}_{t-1}) = 1,$$

the finite fourth moment of ε_t exists, and the bandwidth h_1 is a constant (tuning parameter not depend on sample size), the residual-based estimator obtained from (3.6) *can be referred to as the modal-based robust volatility estimator. The underlying mechanism is that with the use of a Gaussian kernel, for large value of h_1, $1-\exp(-u^2/h_1) \approx u^2/h_1$, and therefore the suggested modal-based estimator is equivalent to the least squares estimator in the extreme case. For a small value of h_1, large values of u will result in a small impact on the estimator. The asymptotic property of such a modal-based robust volatility estimator is given in Remark 3.2.33, which is shown to be more efficient than, or at least as efficient as, that obtained through mean estimation.*

3.2.2 Asymptotic Properties

The majority of previous studies concentrate on the asymptotic behavior of modal estimators with independent data. Our objective is somewhat different in that we consider dependent data and treat the mean regression function in the first step as a nuisance parameter. To make it easier to understand the asymptotic theorems that follow, we introduce some notations that will be used throughout the remaining part of this paper. We define modal residual $\epsilon_t = r_t - Mode(r_t \mid X_t)$, $\ddot{\sigma}^2(x) = \partial^2 \left[\sigma^2(x)\right] / \partial x^2$, $\mu_j = \int w^j K(w) \mathrm{d}w$, and $v_j = \int w^j K^2(w) \mathrm{d}w$, $j = 0, 1, 2, 3$. We call $T_n(x) = T(x) + o_p(s_n)$ (or $O_p(s_n)$) uniformly for $x \in \mathcal{X}$ if $\sup_{x \in \mathcal{X}} |T_n(x) - T(x)| = o_p(s_n)$ (or $O_p(s_n)$). To express convergence in a distribution, we use the symbol "$\xrightarrow{d}$". The integral $\int$ is taken over $(-\infty, \infty)$ unless otherwise specified. To derive the consistency and asymptotic theorems of estimators from (3.6), we impose several regularity conditions that are listed below.

C1 (Kernel Function) The kernel functions $\phi(\cdot)$ and $K(\cdot)$ are both nonnegatively symmetric continuous density functions with bounded support, each of which is integral to one. Moreover, $\int s^{2+\delta} \phi^{2+\delta}(s) < \infty$ and $\int s^{2+\delta} K^{2+\delta}(s) < \infty$ with probability one, where $\delta \in [0, 1)$ is a constant.

C2 (Regression Function) The mean regression function $m(\cdot)$ and the volatility function $\sigma^2(\cdot)$ both have at least a continuous second derivative on an open set containing the point x.

C3 (Density Function) For a fixed point x, $f_X(x)$ is greater than 0 and continuous at x, and $g_\epsilon(\epsilon \mid x) > 0$ is continuous at ϵ, where $g_\epsilon(\epsilon \mid x)$ is the conditional density function

of ϵ given x. In addition, the joint density of X_t and X_j is bounded for all $j \geq t+1$. $g_\epsilon(\epsilon \mid x)$ is assumed to have the fourth continuous derivative and $g_\epsilon(\epsilon \mid x) < g_\epsilon(0 \mid x)$ for all $\epsilon \neq 0$, where $g_\epsilon^{(c)}(\cdot \mid x)$ denotes the cth derivative of $g_\epsilon(\cdot \mid x)$.

C4 (Mixing Process) $\{(Y_t, X_t)\}$ is a stationary α-mixing process, and the mixing coefficient $\rho(n) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_t^\infty} |P(A \cap B) - P(A)P(B)|$ tending to zero for $n \to \infty$ satisfies $\sum_{n \geq 1} n^\gamma (\rho(n))^{\delta/(2+\delta)} < \infty$ for some $\gamma > \delta/(2+\delta)$, where δ is given in C1 and $\mathcal{F}$ is the σ-algebra of events generated by the random variables $\{(Y_t, X_t)\}$. Moreover, there is a sequence of positive integers d_n such that $d_n \to \infty$, $d_n h_1 h_2 \to 0$, and $h_1^4 h_2^{-(2+2\delta)/(2+\delta)} \sum_{k=d_n}^n [\rho(k)]^{\delta/(2+\delta)} = o(n h_2^{-1} h_1^{-3})$.

C5 (Moment) There exists a constant $s > 2$ such that $E(|X|^{2s}) < \infty$, and $E(|Y_t|^\delta \mid X_t) < \infty$ for some $\delta > 2$.

While these conditions may appear to be a little bit verbose at first sight, they are actually rather common in practice. The bounded support in C1 imposed on kernel functions is for the sake of conciseness of proofs, and the Gaussian kernel is allowed (Ullah et al., 2021, 2022), which corresponds to the default kernel used in this paper. The symmetric assumption is quite common. If this is not the case, points equidistant from x could be assigned different weights, which is not really attractive in reality. C2 is a commonly used condition on the smoothness of the nonparametric functions in local linear fitting. It regulates the precision of the approximation since the second derivatives of $m(\cdot)$ and $\sigma^2(\cdot)$ impact the bias. Note that the bias can be further reduced with fitting polynomials of a higher order, leading to an increase of the variability. C3 implies a certain smoothness of $g_\epsilon(\epsilon \mid \cdot)$ in the neighborhood of zero, which is necessary for identification. It imposes that the

conditional density of ϵ has a well defined global mode at zero; see Kemp and Santos Silva (2012) and Ullah et al. (2021, 2022). This unique global mode assumption is being made for a simple illustration. When the population is not homogeneous, the proposed method can also be applied to the multimode setting. C4 is the standard requirement for α-mixing process, which determines the mixing properties of the process under investigation. It is reasonably weak (milder than the standard mixing process where the coefficient decreases at a geometric rate) and is known to be satisfied by many stochastic processes (Cai and Ould-Said, 2003), for instance, the ARMA processes generated by absolutely continuous noise. When $\{(Y_t, X_t)\}$ is independent in which $\delta = 0$, the results in this paper also hold. C5 is the classic rank condition placing restrictions on the moments of covariate to ensure the existence of the asymptotic mean and variance. For modal regression, we do not need to impose moment conditions for Y_t. However, in the first step, we need this condition to achieve the conditional mean estimate. In the case of utilizing modal estimation on both steps, the condition $\delta \geq 2$ can be extended to $\delta > 0$, allowing consequently that $E(Y \mid X) = \infty$. We can then substitute Y by a truncated variable Y^- since the mode is not affected by truncation. All conditions related to bandwidths to control the effects of the dependence of the mixing processes on showing asymptotic normality are listed in the following relevant theorems.

Naturally, one might wonder whether estimating $\sigma^2(x)$ with the estimated value $\hat{r}_t$ rather than the true value r_t would result in some additional asymptotic errors. In answer to this query, we present the following asymptotic theorems and establish the adaptiveness property for the proposed modal volatility function.

Theorem 3.2.11 *Suppose that x is such that $x \pm h_2$ is in the support of $f_X(x)$. Under the conditions C1-C5, with probability approaching one, as $n \to \infty, h_1 \to 0, h_2 \to 0$, $h_2^2/h_1 \to 0$, $nh_2h_1^5 \to \infty$, and $h/h_2 \to 0$, there exist consistent maximizers $(\hat{\sigma}^2(x), h_2\hat{\dot{\sigma}}^2(x))$ of (3.6) such that*

i. $|\hat{\sigma}^2(x) - \sigma^2(x)| = O_p\left(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2\right)$,

ii. $|h_2(\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x))| = O_p\left(\left(nh_2h_1^3\right)^{-1/2} + h_1^2 + h_2^2\right)$.

It is necessary to emphasize that the same rates of convergence are observed in the setting of α-mixing dependence as well as in the case of independence for nonparametric modal regression. A closer examination of Theorem 3.2.11 reveals that the bias terms have contributions from both the estimation of mode and the approximation of function; the first term (h_1^2) in the bias results from the modal estimating process and the second term (h_2^2) is obtained by local linear estimation, which is consistent with mean regression. It can be seen that higher order local polynomial smoothing can reduce the bias incurred by approximating of $\sigma^2(\cdot)$ if $\sigma^2(\cdot)$ has the $(p+1)$th $(p \geq 2)$ order derivative continuous at x. The contribution from the error in the estimator $\hat{m}(x)$ is asymptotically omitted with the assumption that $h/h_2 \to 0$. Notice that the optimal bandwidth h that minimizes the asymptotic MSE has an optimal rate of $n^{-1/5}$, while the MSE-optimal bandwidth h_2 has the rate of $n^{-1/8}$, implying that the condition $h/h_2 \to 0$ is unambiguously met. Therefore, there is no need to undersmooth $\hat{m}(x)$ when conducting mean estimation. The next theorem gives the asymptotic distributions of the modal estimators we are interested in, which is in parallel with those arising in nonparametric modal estimators with independent data.

Theorem 3.2.12 *With $nh_2^5h_1^3 = O(1)$ and $nh_2h_1^7 = O(1)$, under the same conditions as Theorem 3.2.11, the estimators satisfying the consistency results in Theorem 3.2.11 have the following asymptotic result*

$$\sqrt{nh_2h_1^3}\left(\begin{bmatrix}\hat{\sigma}^2(x)-\sigma^2(x)\\ h_2(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))\end{bmatrix}-\Gamma^{-1}\left(\frac{h_2^2}{2}\Lambda_2\ddot{\sigma}^2(x)-\frac{h_1^2}{2}\Lambda_1\right)\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,\frac{\int\tau^2\phi^2(\tau)d\tau}{f_X(x)}\Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

If we allow $nh_2^5h_1^3\to 0$ and $nh_2h_1^7\to 0$, the asymptotic theorem becomes

$$\sqrt{nh_2h_1^3}\begin{pmatrix}\hat{\sigma}^2(x)-\sigma^2(x)\\ h_2(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(0,\frac{\int\tau^2\phi^2(\tau)d\tau}{f_X(x)}\Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

where

$$\Gamma=\begin{bmatrix}\mu_0g_\epsilon^{(2)}(0\mid x) & \mu_1g_\epsilon^{(2)}(0\mid x)\\ \mu_1g_\epsilon^{(2)}(0\mid x) & \mu_2g_\epsilon^{(2)}(0\mid x)\end{bmatrix},\ \Sigma=\begin{bmatrix}v_0g_\epsilon(0\mid x) & v_1g_\epsilon(0\mid x)\\ v_1g_\epsilon(0\mid x) & v_2g_\epsilon(0\mid x)\end{bmatrix},\ \Lambda_1=\begin{bmatrix}\mu_0g_\epsilon^{(3)}(0\mid x)\\ \mu_1g_\epsilon^{(3)}(0\mid x)\end{bmatrix},$$

and $\Lambda_2=\begin{bmatrix}\mu_2g_\epsilon^{(2)}(0\mid x)\\ \mu_3g_\epsilon^{(2)}(0\mid x)\end{bmatrix}$. *Provided that $\mu_1=0$ and $\mu_3=0$ with a symmetric kernel, the results indicate that $\hat{\sigma}^2(x)$ and $\hat{\dot{\sigma}}^2(x)$ are asymptomatically independent.*

The proof of Theorem 3.2.12 is based on the Bahadur-type representation for the nonparametric estimator of modal volatility. It states that the residual-based modal estimator $\hat{\sigma}^2(x)$, which does not require $m(\cdot)$ to be known, is asymptotically as efficient as the oracle estimator as if the knowledge of $m(\cdot)$ were known in advance. In particular, the asymptotic results do not include any additional bias or variance components due to the first step in the estimation procedure. This adaptiveness property to the unknown conditional mean regression function is shared by other residual-based volatility estimators;

see Ziegelmann (2002) and Xu and Philips (2011). Theorem 3.2.12 also indicates that the asymptotic bias term can vanish fast enough under certain conditions to have no impact on the asymptotic distribution. However, the optimal bandwidths of h_1 and h_2 have the rate $n^{-1/8}$, which does not fulfill the conditions that $nh_2^5h_1^3 \to 0$ and $nh_2h_1^7 \to 0$. As a result, consistent with most semiparametric regression literature, undersmoothing is required. Comparing Theorem 3.2.12 to Theorem 1 in Fan and Yao (1998), which investigated a mean variance estimator based on random and absolutely regular observations, the modal volatility estimator has a slower convergence rate. Also, Theorem 3.2.12 indicates that under suitable conditions, the dependence of the observations does not influence the asymptotic distribution of the modal volatility estimator. The primary reason for this is attributed to the imposed conditions on bandwidth choice. Under α-mixing process, the covariance between random variables ϵ_t and ϵ_j such that $\epsilon_t, \epsilon_j \in (\epsilon - h_1, \epsilon + h_1)$ is dominated by the variance of ϵ_t (or is nearly uncorrelated) under certain mild conditions. The similar explanation applies to variables X_t and X_j. Therefore, when the dependence is moderate, the asymptotic theorem is identical to that in the i.i.d. case.

Remark 3.2.32 (Boundary Behavior) *The behavior near the boundary, which is a well-known appealing property of local linear smoothers, can be shown to carry over to the mode case. Take $x = ch_2$ for some $0 < c < 1$ and define $\mu_j(c) = \int_{-c}^{\infty} w^j K(w)dw$ and $v_j(c) = \int_{-c}^{\infty} w^j K^2(w)dw$. Under conditions similar to those of Theorem 3.2.12, we can obtain*

$$Bias(\hat{\sigma}^2(x)) = \frac{h_2^2}{2}\ddot{\sigma}^2(x)\frac{\mu_2^2(c) - \mu_1(c)\mu_3(c)}{\mu_0(c)\mu_2(c) - \mu_1^2(c)} - \frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)}.$$

$$Var(\hat{\sigma}^2(x)) = \frac{\int \tau^2\phi^2(\tau)d\tau}{nh_2h_1^3 f_X(x)}\frac{g_\epsilon(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)^2}\frac{\mu_2^2(c)v_0(c) - 2\mu_1(c)\mu_2(c)v_1(c) + \mu_1^2(c)v_2(c)}{(\mu_0(c)\mu_2(c) - \mu_1^2(c))^2}.$$

Thus, the local linear modal estimator adapts automatically to all locations.

Remark 3.2.33 (Modal-Based Robust Volatility Estimator) *Following Remark 3.2.31, to derive the asymptotic theorem for the modal-based robust volatility estimator under stationary α-mixing dependent samples, we impose the following necessary conditions.*

D1 *The errors ϵ_t are symmetrically random errors with zero mode and zero mean, and they are independent of X_t. Also, $E(\phi_{h_1}^{(1)}(\epsilon_t) \mid X_t) = 0$, where $\phi_{h_1}(\epsilon_t) = \phi(\epsilon_t/h_1)/h_1$ and $\phi_{h_1}^{(c)}(\cdot)$ represents the cth derivative. In addition, $E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t) < 0$, $E((\phi_{h_1}^{(2)}(\epsilon_t))^2 \mid X_t)$, $(E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t))^2$, $E(\phi_{h_1}^{(1)}(\epsilon_t)^2 \mid X_t)$, $E(|\phi_{h_1}^{(1)}(\epsilon_t)^3| \mid X_t)$, and $E(\phi_{h_1}^{(3)}(\epsilon_t) \mid X_t)$ are continuous with respect to X_t.*

D2 *The random variable X has a bound support, and its density $f_X(\cdot)$ has a continuous first derivative and is bounded away from zero and infinity. In addition, the joint density of X_t and X_j is bounded for all $j \geq t+1$.*

D3 *There exists a sequence of positive integers d_n such that $d_n \to \infty$, $d_n = o\left((nh_2)^{1/2}\right)$, and $(n/h_2)^{1/2}\rho\left(d_n\right) \to 0$ as $n \to \infty$.*

D4 *The conditional moments $E(|\phi_{h_1}^{(1)}(\epsilon_t)|^{2+\gamma} \mid X_t)$ and $E(|\phi_{h_1}^{(2)}(\epsilon_t)|^{2+\gamma} \mid X_t)$ are bounded. There exits $\tau > 2+\gamma$ such that $E(|\phi_{h_1}^{(1)}(\epsilon_t)|^{\tau} \mid X_t)$ is bounded and $\rho(n) = O(n^{-\theta})$ such that $\theta \geq (2+\gamma)\tau/(2(\tau-2-\gamma))$. Also, $n^{-\gamma/4}h_1^{(2+\gamma)/\tau-1-\gamma/4} = O(1)$.*

D5 *The bandwidth h_2 is a real sequence such that as $n \to \infty$, $h_2 \to 0$ and $nh_2 \to \infty$. Also, it is necessary that $h/h_2 \to 0$.*

Completely different from the proposed modal volatility estimator, we now need to impose moment conditions for the modal-based robust volatility estimator. Condition D1 is a necessary condition for the asymptotic normality and consistency of the modal-based ro-

bust estimator. The derivative $\phi_{h_1}^{(1)}(\cdot)$ can be considered as the influence function, measuring the influence of an observation on the value of the parameter estimate. Condition D2 is a standard smoothness condition that is found in a variety of applications. Conditions D3 and D4 are utilized to ensure the asymptotic neglect of dependence between observations. Such conditions are different from the ones used in modal volatility estimation since we do not need to control the dependence between errors. Condition D5 is the classical bandwidth condition in nonparametric estimation. As opposed to modal volatility estimation, we presently need to undersmooth the first step estimators in order to asymptotically ignore their effect in the second step estimation. There is no condition imposed on bandwidth h_1 because it is treated as a constant. We then have the following asymptotic theorem, with the proof obtained by using the first-order Taylor expansion and the central limit theorem.

Theorem 3.2.13 *Suppose that the point x at which the estimator is taking place satisfies $h_2 < x < 1 - h_2$. With $nh_2^5 = O(1)$, under the conditions C1, C2, the first part of C4, and D1-D5, as $n \to \infty$, we have the following asymptotic result*

$$\sqrt{nh_2}\left[\begin{pmatrix} \hat{\sigma}^2(x) - \sigma^2(x) \\ \\ h_2(\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x)) \end{pmatrix} - \frac{h_2^2}{2}\Gamma_m^{-1}\Lambda_m \ddot{\sigma}^2(x)\right]$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{E((\phi_{h_1}^{(1)}(\epsilon))^2 \mid X = x)\Gamma_m^{-1}\Sigma_m(\Gamma_m^{-1})^T}{(E(\phi_{h_1}^{(2)}(\epsilon) \mid X = x))^2 f_X(x)}\right).$$

If we allow $nh_2^5 \to 0$, the asymptotic theorem becomes

$$\sqrt{nh_2}\begin{pmatrix} \hat{\sigma}^2(x) - \sigma^2(x) \\ \\ h_2(\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x)) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \frac{E((\phi_{h_1}^{(1)}(\epsilon))^2 \mid X = x)}{(E(\phi_{h_1}^{(2)}(\epsilon) \mid X = x))^2 f_X(x)}\Gamma_m^{-1}\Sigma_m(\Gamma_m^{-1})^T\right),$$

where $\Lambda_m = \begin{bmatrix} \mu_2 \\ \\ \mu_3 \end{bmatrix}$, $\Gamma_m = \begin{bmatrix} \mu_0 & \mu_1 \\ \\ \mu_1 & \mu_2 \end{bmatrix}$, *and* $\Sigma_m = \begin{bmatrix} v_0 & v_1 \\ \\ v_1 & v_2 \end{bmatrix}$.

We briefly outline the proof in the appendix. In comparison to the modal volatility estimator, the convergence rate now becomes $(nh_2)^{1/2}$*, which is the same as that of local linear mean estimator. Similar to the literature on semiparametric two-step estimators involving kernel estimation in the first step, the bias and variance in the first step vanish asymptotically with undersmoothing* ($h/h_2 \to 0$)*. The above theorem also shows that the modal-based robust volatility estimator is not only asymptotically equivalent to the local linear mean robust volatility estimator but is also more efficient. An asymptotic comparison with local linear mean estimation is entirely driven by a comparison of their asymptotic variance (they share the same asymptotic bias), where we can achieve an efficiency gain in the presence of certain conditions by adjusting the value of* h_1*; see Yao et al. (2012) and Ullah et al. (2022). In practice, we can minimize the asymptotic variance, utilize the plug-in method to replace unknown functions with consistent estimators, and combine with the grid search method to choose bandwidths* h_1 *and* h_2*. Due to space constraints, the comprehensive investigation of such a modal-based robust volatility estimator is omitted here.*

Remark 3.2.34 *Because of the use of only a small part of the data, quantifying uncertainty in modal statistics is challenging. The theorems stated above establish the asymptotic distribution of the proposed modal (and modal-based robust) volatility estimators. However, the asymptotic distribution involves many unknown terms, necessitating further nonparametric estimation with new bandwidths. Consequently, it would be difficult to apply the asymptotic results directly to perform (modal) inference. We will not explore the problem of inference here, but readers may consult, for example, we can adopt the stationary bootstrap in Politis and Romano (1994) to construct confidence intervals for modal (and modal-based robust) volatility.*

3.2.3 Optimal Bandwidths

It is commonly acknowledged that in a standard finite dimensional framework, the smoothing parameters must be appropriately chosen to ensure good practical performance. Different from other nonparametric methods, setting the bandwidths in modal regression not only controls the trade off between bias and variance but also affects the target of the objective (mean estimator or modal estimator). In addition, for modal regression, bandwidth choice is considerably more important since the value of bandwidth associated with error terms affects the number of estimated modes. In this part, we explore asymptotic optimal bandwidths for h_1 and h_2 and show how to obtain bandwidths suggested by data.

Notice that the mean regression function $m(X_t)$ only plays the role of a nuisance parameter during the modal estimation process, and the bandwidth h used in estimating the mean regression function is not as crucial as the bandwidths applied in the modal estimator. The asymptotic result presented above justifies the use of standard bandwidth selectors developed to estimate the mean function. We thus follow the rule of thumb to choose h for simplicity and practical convenience, i.e., $\hat{h} = h_a \hat{\sigma}_X n^{-1/5}$, where $\hat{\sigma}_X$ is the standard deviation of X, h_a is a real constant to be tuned,[8] and $n^{-1/5}$ is the rate of the MSE-optimal bandwidth. We refer readers to Fan and Yao (1998) for the expression of the asymptotic optimal bandwidth h.

With the obtained asymptotic properties, we can derive the asymptotic optimal bandwidths for h_1 and h_2 . The fundamental idea is to minimize the asymptotic MSE of an estimator. Considering the estimator of $\sigma^2(x)$ at point x, the asymptotic MSE equals

[8]The selection of the bandwidth parameter h_a has been debated at length in the literature. We choose $h_a = 1.06$ for numerical examples since it is the optimum number for density estimation with respect to the mean integrated standard error criteria (Silverman, 1986).

$$
\begin{aligned}
&MSE(\hat{\sigma}^2(x)) = \text{Bias}(\hat{\sigma}^2(x))^2 + \text{Var}(\hat{\sigma}^2(x)) \\
\approx &\left\{ e_1^T \Gamma^{-1} \left(\frac{h_2^2}{2} \ddot{\sigma}^2(x) \Lambda_2 - \frac{h_1^2}{2} \Lambda_1 \right) \right\}^2 + \frac{\int \tau^2 \phi^2(\tau) d\tau}{n h_2 h_1^3 f_X(x)} e_1^T \Gamma^{-1} \Sigma \Gamma^{-1} e_1,
\end{aligned} \tag{3.7}
$$

where $e_1 = (1, 0)^T$. By defining $(\hat{h}_1, \hat{h}_2) = \arg\min_{h_1, h_2} MSE\ (\hat{\sigma}^2(x))$, we have the following result in which $\Delta_1 = (\ddot{\sigma}^2(x) e_1^T \Gamma^{-1} \Lambda_2)^2$, $\Delta_2 = (e_1^T \Gamma^{-1} \Lambda_1)^2$, $\Delta_3 = (\ddot{\sigma}^2(x) e_1^T \Gamma^{-1} \Lambda_2)(e_1^T \Gamma^{-1} \Lambda_1)$, and $\Delta_4 = \Delta_2 / (\sqrt{\Delta_3^2 + 3\Delta_1 \Delta_2} - \Delta_3)$.

Corollary 3.2.1 *Under the same conditions as Theorem 3.3.14, the optimal bandwidths of h_1 and h_2 satisfy $\hat{h}_2 = \hat{h}_1 \Delta_4^{1/2}$, where*

$$
\hat{h}_1 = \left(\frac{(\Delta_4^{5/2} \Delta_1 - \Delta_4^{3/2} \Delta_2) f_X(x)}{\int \tau^2 \phi^2(\tau) d\tau e_1^T \Gamma^{-1} \Sigma \Gamma^{-1} e_1} \right)^{-\frac{1}{8}} n^{-\frac{1}{8}}.
$$

Remark 3.2.35 *It can be seen that the modal rate of the MSE-optimal bandwidth is smaller than that of the mean estimator. Thus, the value of $\hat{h}_1$ is larger than the value of $\hat{h}$. One can also minimize the asymptotic weighted mean integrated squared error to obtain the asymptotic global optimal bandwidth. Especially, we can minimize*

$$
\int E[\hat{\sigma}^2(x) - \sigma^2(x)]^T W [\hat{\sigma}^2(x) - \sigma^2(x)] w(x) dx,
$$

where W is a weight matrix and $w(x)$ denotes a weight function. One popular choice for W is the inverse of the asymptotic variance of $\hat{\sigma}^2(x)$.

If we replace the bandwidths in $(n h_2 h_1^3)^{1/2}$ with the MSE-optimal bandwidth values, we can achieve $n^{1/4}$ consistency, indicating that the convergence rate of modal regression is slower than that of mean regression. Notice that the optimal bandwidths in the above corollary are complicatedly dependent on the unknown densities $g_\epsilon(\cdot)$ and $f_X(\cdot)$, which are

not accessible in practice.[9] However, the expression can provide us with some guidelines on how to select the practically optimal data-driven bandwidths. To simplify the calculations, we generalize the method in Kemp and Santos Silva (2012) to choose the optimal bandwidths and let $\hat{h}_1 = 1.6\text{MAD}n^{-0.13}$ (-0.13 comes from the rate -1/8 and undersmoothing requirement), where

$$MAD = med_j\{|(\hat{r}_j - \hat{\sigma}_m^2(X_j)) - med_t(\hat{r}_t - \hat{\sigma}_m^2(X_t))|\} \tag{3.8}$$

is the median absolute deviation and $\hat{\sigma}_m^2(\cdot)$ represents the mean volatility estimator. We set $\hat{h}_2 = 1.06\sigma_X n^{-0.13}$, in which σ_X is the standard deviation of the sample X_t. Although these informal selection procedures may not provide the optimal estimates in practice, they are formally consistent with the shrinking rates of bandwidths and the requirement for undersmoothing. How to precisely choose the optimal bandwidths within the content of modal volatility estimation would be quite involved and requires additional research in the future to fully understand.

3.3 Variance Reduced Modal Volatility Estimator

Although the proposed modal volatility estimator has a slower convergence rate, there is room for improvement in terms of variance. Motivated by the increasing attention in the literature on variance reduction, we extend the variance reduction technique in Cheng et al.

[9] One method related to bandwidth selection for modal regression is the plug-in method, which is proposed in Yao and Li (2014), Yao and Xiang (2016), and Ullah et al. (2021, 2022). They estimated the optimal bandwidths following the expressions of the asymptotic MSE-optimal bandwidths by using the estimated densities of $g_\epsilon(\cdot)$ and $f_X(\cdot)$. They then replaced the unknown terms in expressions with the corresponding estimates. Nevertheless, for nonparametric modal regression, the computation burden associated with the plug-in method will increase dramatically. Also, the traditional cross validation based on MSE criteria cannot be used here, as modal regression is intended to maximize a kernel-based objective function. Although kernel-based cross validation may be applied, the asymptotic property has not yet been completed investigated.

(2007) to improve the estimation of the modal volatility estimator $\hat{\sigma}^2(x)$ by constructing a linear combination of modal estimators at three points around x.[10] The main idea of variance reduction is to incorporate more data points around the target one to reduce variance while remaining the asymptotic bias unchanged through certain moment conditions derived from asymptotic bias expansions, which can ultimately result in a significant improvement in the asymptotic MSE. It is worth noting that a similar technique has been adopted by Choi and Hall (1998), in which they fitted a straight line segment to a curve in a symmetric way to reduce bias while keeping variance unchanged.

3.3.1 Variance Reduced Modal Estimation

For any x, we define three equally spaced points $x-(r+1-j)\beta h_2$, $j = 0, 1, 2$, to form a linear combination of the values $\tilde{\sigma}^2(\cdot)$, where the shift parameter $r \in (-1, 0) \cup (0, 1)$ represents the relative location and $\beta h_2 > 0$ indicates the spacing of the grid. Under the assumption that $\beta > 0$, $\tilde{\sigma}^2(x)$ is identical to $\hat{\sigma}^2(x)$ if and only if $r \in \{-1, 0, 1\}$. When $\beta = 0$, $\tilde{\sigma}^2(\cdot)$ is degenerated to $\hat{\sigma}^2(\cdot)$, which is not of any interest. Then, the variance reduced modal volatility estimator for $\hat{\sigma}^2(x)$ is formally given by the interpolated curve at x

$$\tilde{\sigma}^2(x) = \frac{r(r-1)}{2}\hat{\sigma}^2(x-(r+1)\beta h_2)+(1-r^2)\hat{\sigma}^2(x-r\beta h_2)+\frac{r(r+1)}{2}\hat{\sigma}^2(x-(r-1)\beta h_2), \quad (3.9)$$

where the moment condition $2^{-1}r(r-1)(-1-r)^j+(1-r^2)(-r)^j+2^{-1}r(r+1)(1-r)^j = \beta_{0,j}$ in which $\beta_{0,j} = 1$ if $j = 0$ and 0 otherwise is satisfied to ensure the asymptotic bias of $\tilde{\sigma}^2(x)$ is the same as that of $\hat{\sigma}^2(x)$. Taking x in (3.9) to be $X_1, \cdots, X_n$, respectively, we can obtain the variance reduced modal volatility estimators of $\sigma^2(x)$ at all of the design points. For

[10] The choice of three nearby points is based on the minimal requirement imposed by the moment conditions, while the solutions will become more complicated as the number of nearby points increases beyond three.

explicitness and simplicity of presentation, we only consider the variance reduced estimator for the modal function $\sigma^2(x)$. However, the method proposed in this subsection can also be straightforwardly applied to any other high order modal derivatives.

As with the argument in Cheng et al. (2007), the variance reduction is accomplished because the correlation coefficients of the above three estimators are smaller than one. Furthermore, assuming that $Supp(\sigma^2(x))$ is bounded, i.e., $Supp(\sigma^2(x)) = [0,1]$, in order to ensure all points are inside $Supp(\sigma^2(x))$ all the time, we choose $\beta(x) = \min\{\beta, x/[(r+1)h_2], (1-x)/[(1-r)h_2]\}$ since $x-(1-r)\beta h_2 < x < x+(1+r)\beta h_2$. Then, $\tilde{\sigma}^2(x)$ will have the same asymptotic bias as $\hat{\sigma}^2(x)$. Before we present the asymptotic theorem for $\tilde{\sigma}^2(x)$, we list the asymptotic distribution for $\hat{\sigma}^2(x)$ that originates from the Theorem 3.2.12.

Theorem 3.3.14 *Suppose that x is any given point in the interior of the support of $f_X(\cdot)$. With $nh_2^5h_1^3 = O(1)$ and $nh_2h_1^7 = O(1)$, under the same conditions as Theorem 3.2.11, the estimator $\hat{\sigma}^2(x)$ satisfying the consistency result in Theorem 3.2.11 is asymptotically normal. That is,*

$$P\left(\frac{\hat{\sigma}^2(x) - \sigma^2(x) - B(x)}{\sqrt{V^2(x)/(nh_2h_1^3)}} \le t\right) = \Phi(t) + o_p(1),$$

where $\Phi(\cdot)$ is the standard normal distribution function,

$$B(x) = \frac{h_2^2}{2}\mu_2\ddot{\sigma}^2(x) - \frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)}, \text{ and } V^2(x) = \frac{g_\epsilon(0 \mid x)\int \tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0 \mid x)^2 f_X(x)}v_0.$$

If we allow $nh_2^5h_1^3 \to 0$ and $nh_2h_1^7 \to 0$, the asymptotic normality becomes

$$P\left(\frac{\hat{\sigma}^2(x) - \sigma^2(x)}{\sqrt{V^2(x)/(nh_2h_1^3)}} \le t\right) = \Phi(t) + o_p(1).$$

We obtain in what follows the asymptotic theorem for the variance reduced modal volatility estimator $\tilde{\sigma}^2(x)$, which shows that $\tilde{\sigma}^2(x)$ is asymptotically efficient relative to $\hat{\sigma}^2(x)$ by reducing asymptotic variance by a known factor. Such a result is to be expected given that the estimator $\tilde{\sigma}^2(x)$ is simply a linear combination of local linear modal volatility estimators evaluated at nearby points without any pilot estimation.

Theorem 3.3.15 *With $nh_2^5h_1^3 = O(1)$ and $nh_2h_1^7 = O(1)$, under the same conditions as Theorem 3.2.11, we have the following asymptotic result*

$$\sqrt{nh_2h_1^3}\left(\tilde{\sigma}^2(x) - \sigma^2(x) - \left(\frac{h_2^2}{2}\mu_2\ddot{\sigma}^2(x) - \frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)}\right)\right)$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid x)\int \tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0 \mid x)^2 f_X(x)}\left(v_0 - r^2(1-r^2)C(\beta)\right)\right).$$

If we allow $nh_2^5h_1^3 \to 0$ and $nh_2h_1^7 \to 0$, the asymptotic theorem becomes

$$\sqrt{nh_2h_1^3}\left(\tilde{\sigma}^2(x) - \sigma^2(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{g_\epsilon(0 \mid x)\int \tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0 \mid x)^2 f_X(x)}\left(v_0 - r^2(1-r^2)C(\beta)\right)\right),$$

where $C(\beta) = 1.5C(0,\beta) - 2C(0.5,\beta) + 0.5C(1,\beta)$ and $C(t,\beta) = \int K(w - t\beta)K(w + t\beta)dw$.

Remark 3.3.36 *Because of the use of local linear estimation, when x is a boundary point, that is, when x is close to the endpoints of the support of X ($supp(X)$), $\tilde{\sigma}^2(x)$ still has an asymptotic normal distribution, with only the constant factors in the asymptotic bias and variance changing.*

The function $C(\cdot)$ in Theorem 3.3.15 shares the same form as the corresponding function in Cheng et al. (2007) and $C(\beta) \geq 0$ for any $\beta \geq 0$ with the symmetric kernel $K(\cdot)$.[11] Comparing Theorem 3.3.15 to Theorem 3.3.14, the asymptotic bias is not changed,

[11] If $K(\cdot)$ has a unique maximum and is concave, $C(\beta)$ is increasing in $\beta \geq 0$.

while the asymptotic variance is reduced by the amount $\{nh_2h_1^3 f_X(x)g_\epsilon^{(2)}(0 \mid x)^2\}^{-1} g_\epsilon(0 \mid x) r^2(1-r^2)C(\beta)$ that depends on β. Because r is an arbitrary constant that is not equal to 0, 1, or -1, the asymptotic variance of $\tilde{\sigma}^2(x)$ is always smaller than that of $\hat{\sigma}^2(x)$, with the maximum achieved at $r = \pm\sqrt{1/2}$ regardless of the other parameters. We can get the most variance reduction with estimators

$$\begin{aligned}\tilde{\tilde{\sigma}}_1^2(x) = &\frac{1/2 - \sqrt{1/2}}{2}\hat{\sigma}^2(x - (\sqrt{1/2}+1)\beta h_2) + \frac{1}{2}\hat{\sigma}^2(x - \sqrt{1/2}\beta h_2) \\ &+ \frac{1/2 + \sqrt{1/2}}{2}\hat{\sigma}^2(x - (\sqrt{1/2}-1)\beta h_2),\end{aligned}$$

$$\begin{aligned}\tilde{\tilde{\sigma}}_2^2(x) = &\frac{1/2 + \sqrt{1/2}}{2}\hat{\sigma}^2(x - (1-\sqrt{1/2})\beta h_2) + \frac{1}{2}\hat{\sigma}^2(x + \sqrt{1/2}\beta h_2) \\ &+ \frac{1/2 - \sqrt{1/2}}{2}\hat{\sigma}^2(x - (1-\sqrt{1/2})\beta h_2),\end{aligned}$$

where both of them have asymptotic variance $\{nh_2h_1^3 f_X(x)g_\epsilon^{(2)}(0 \mid x)^2\}^{-1} g_\epsilon(0 \mid x)(v_0 - C(\beta)/4)$. It can be observed that either of the variance reduction estimators $\tilde{\tilde{\sigma}}_1^2(x)$ and $\tilde{\tilde{\sigma}}_2^2(x)$ utilizes more information from data points on one side of x than the other side. To balance the finite sample bias caused by $\tilde{\tilde{\sigma}}_1^2(x)$ and $\tilde{\tilde{\sigma}}_2^2(x)$, we define the final modal volatility estimator by equally averaging the above two estimators

$$\tilde{\tilde{\sigma}}^2(x) = \frac{1}{2}\tilde{\tilde{\sigma}}_1^2(x) + \frac{1}{2}\tilde{\tilde{\sigma}}_2^2(x), \tag{3.10}$$

which can further improve the asymptotic efficiency of the modal volatility estimator. The proposed variance reduced modal volatility estimator is, as can be seen, a simple linear combination of local linear modal volatility estimators evaluated at nearby points. The following theorem regarding $\tilde{\tilde{\sigma}}^2(x)$ follows immediately from Theorem 3.3.15.

Theorem 3.3.16 *With $nh_2^5h_1^3 = O(1)$ and $nh_2h_1^7 = O(1)$, under the same conditions as Theorem 3.2.11, we have the following asymptotic result*

$$\sqrt{nh_2h_1^3}\left(\tilde{\tilde{\sigma}}^2(x)-\sigma^2(x)-\left(\frac{h_2^2}{2}\mu_2\ddot{\sigma}^2(x)-\frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0\mid x)}{g_\epsilon^{(2)}(0\mid x)}\right)\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,\frac{g_\epsilon(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0\mid x)^2 f_X(x)}\left(\frac{v_0}{2}-\frac{C(\beta)}{8}-\frac{D(\beta)}{2}\right)\right).$$

If we allow $nh_2^5h_1^3\to 0$ *and* $nh_2h_1^7\to 0$*, the asymptotic theorem becomes*

$$\sqrt{nh_2h_1^3}\left(\tilde{\tilde{\sigma}}^2(x)-\sigma^2(x)\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{g_\epsilon(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0\mid x)^2 f_X(x)}\left(\frac{v_0}{2}-\frac{C(\beta)}{8}-\frac{D(\beta)}{2}\right)\right),$$

where $D(\delta)=\frac{1}{16}\{4(1+\sqrt{2})C(\sqrt{2}-1,\beta/2)\ +(3+2\sqrt{2})C(2-\sqrt{2},\beta/2)\ +2C(\sqrt{2},\beta/2)+$ $4(1-\sqrt{2})C(\sqrt{2}+1,\beta/2)\ +(3-2\sqrt{2})C(\sqrt{2}+2,\beta/2)\}$.

It can be shown that the variance of $\tilde{\tilde{\sigma}}^2(x)$ is always smaller than the variance of $\tilde{\sigma}^2(x)$, indicating that $\tilde{\tilde{\sigma}}^2(x)$ enjoys an appealing advantage in terms of the improvement of efficiency (sharing the same asymptotic bias). In order to guarantee that all points are contained inside $Supp(\sigma^2(x))$, for a given positive constant β, we choose

$$\beta(x)=\min\left\{\beta,\frac{x}{(\sqrt{1/2}+1)h_2},\frac{1-x}{(\sqrt{1/2}+1)h_2}\right\}. \tag{3.11}$$

Note that while theoretical results suggest that implementing larger β values can achieve more variance reduction, doing so may introduce significant finite sample bias effects. We thus follow the instructions in Cheng et al. (2007) to take $\beta=1$ for general purposes.

Remark 3.3.37 *Given the present model settings, applying variance reduction to* $m(x)$ *does not provide any gain in the asymptotic results of the modal volatility estimator. If we apply the variance reduction technique to the modal-based robust volatility estimator, the variance term becomes* $f_X^{-1}(x)E((\phi_{h_1}^{(1)}(\epsilon))^2\mid X=x)(E(\phi_{h_1}^{(2)}(\epsilon)\mid X=x))^{-2}\left(v_0-r^2(1-r^2)C(\beta)\right)$*. The final modal-based robust volatility estimator, based on the preceding arguments, has the variance* $f_X^{-1}(x)E((\phi_{h_1}^{(1)}\ (\epsilon))^2\mid X=x)(E(\phi_{h_1}^{(2)}(\epsilon)\mid X=x))^{-2}\,(v_0/2-C(\beta)/8-D(\beta)/2)$.

3.3.2 Optimal Bandwidths

With the asymptotic properties established in the previous section, we can derive the asymptotic optimal bandwidths for the variance reduced modal volatility estimator. The essential idea is to minimize the asymptotic MSE of the estimator. Taking into consideration the estimator $\tilde{\sigma}^2(x)$, the asymptotic MSE equals

$$\begin{aligned} &MSE(\tilde{\sigma}^2(x)) = \text{Bias}(\tilde{\sigma}^2(x))^2 + \text{Var}(\tilde{\sigma}^2(x)) \\ &= \left(\frac{h_2^2}{2} \mu_2 \ddot{\sigma}^2(x) - \frac{h_1^2}{2} \frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)} \right)^2 + \left(\frac{g_\epsilon(0 \mid x) \int \tau^2 \phi^2(\tau) d\tau}{n h_2 h_1^3 g_\epsilon^{(2)}(0 \mid x)^2 f_X(x)} \left(v_0 - r^2(1-r^2) C(\beta) \right) \right). \end{aligned} \tag{3.12}$$

In comparison to $MSE(\hat{\sigma}^2(x))$, we now have an extra term $-(nh_2h_1^3 g_\epsilon^{(2)}(0 \mid x)^2 f_X(x))^{-1} g_\epsilon (0 \mid x) \int \tau^2 \phi^2(\tau) d\tau r^2(1-r^2) C(\beta)$. Note that $0 < r^2(1-r^2) \leq 1/4$ for any $r \in (-1,1) \backslash \{0\}$, which reaches the maximum at $r = \pm 2^{-1/2}$. Moreover, for any symmetric kernel $K(\cdot)$, $0 \leq C(\beta) \leq 3 \int K^2(u) du/2$ for all $\beta > 0$. As a result, the variance reduced modal volatility estimator performs significantly better than the local linear modal volatility estimator in terms of asymptotic MSE. By defining $(\tilde{h}_1, \tilde{h}_2) = \arg\min_{h_1, h_2} MSE(\tilde{\sigma}^2(x))$, we have the following result, where $\Delta_5 = -g_\epsilon^{(3)}(0 \mid x)(3\mu_2 \ddot{\sigma}^2(x) g_\epsilon^{(2)}(0 \mid x))^{-1}$.

Corollary 3.3.2 *Under the regularity conditions C1-C5, the optimal bandwidths of h_1 and h_2 satisfy $\tilde{h}_2 = \tilde{h}_1 \Delta_5^{1/2}$, where*

$$\tilde{h}_1 = \left(\frac{g_\epsilon^{(2)}(0 \mid x)^2 f_X(x) \Delta_5^3 \mu_2 \ddot{\sigma}^2(x) \left(\Delta_5^2 \mu_2 \ddot{\sigma}^2(x) - g_\epsilon^{(3)}(0 \mid x) g_\epsilon^{-(2)}(0 \mid x) \right)}{g_\epsilon(0 \mid X) \int \tau^2 \phi^2(\tau) d\tau \left(v_0 - \frac{1}{4} C(\beta) \right)} \right)^{-\frac{1}{8}} n^{-\frac{1}{8}}.$$

Remark 3.3.38 *The optimal bandwidths result in $MSE(\tilde{\sigma}^2(x)) = (v_0 - \frac{C(\beta)}{4})^{\frac{7}{8}} v_0^{-\frac{7}{8}} \; MSE(\hat{\sigma}^2(x))$, which indicates the asymptotic relative efficiency*

$$\frac{MSE(\hat{\sigma}^2(x))}{MSE(\hat{\sigma}^2(x))} = (v_0 - \frac{C(\beta)}{4})^{-\frac{7}{8}} v_0^{\frac{7}{8}} \geq 1.$$

Considering the estimator $\tilde{\tilde{\sigma}}^2(x)$, the asymptotic MSE equals

$$\begin{aligned} &MSE(\tilde{\tilde{\sigma}}^2(x)) = \text{Bias}(\tilde{\tilde{\sigma}}^2(x))^2 + \text{Var}(\tilde{\tilde{\sigma}}^2(x)) \\ &= \left(\frac{h_2^2}{2} \mu_2 \ddot{\sigma}^2(x) - \frac{h_1^2}{2} \frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)} \right)^2 + \left(\frac{g_\epsilon(0 \mid x) \int \tau^2 \phi^2(\tau) d\tau}{n h_2 h_1^3 g_\epsilon^{(2)}(0 \mid x)^2 f_X(x)} \left(\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2} \right) \right). \end{aligned} \tag{3.13}$$

By defining $(\tilde{\tilde{h}}_1, \tilde{\tilde{h}}_2) = \arg\min_{h_1, h_2} MSE(\tilde{\tilde{\sigma}}^2(x))$, we have the following result.

Corollary 3.3.3 *Under the regularity conditions C1-C5, the optimal bandwidths of h_1 and h_2 satisfy $\tilde{\tilde{h}}_2 = \tilde{\tilde{h}}_1 \Delta_5^{1/2}$, where*

$$\tilde{\tilde{h}}_1 = \left(\frac{g_\epsilon^{(2)}(0 \mid x)^2 f_X(x) \Delta_5^3 \mu_2 \ddot{\sigma}^2(x) \left(\Delta_5^2 \mu_2 \ddot{\sigma}^2(x) - g_\epsilon^{(3)}(0 \mid x) g_\epsilon^{-(2)}(0 \mid x) \right)}{g_\epsilon(0 \mid x) \int \tau^2 \phi^2(\tau) d\tau \left(\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2} \right)} \right)^{-\frac{1}{8}} n^{-\frac{1}{8}}.$$

Remark 3.3.39 *The optimal bandwidths produce $MSE(\tilde{\tilde{\sigma}}^2(x)) = (\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2})^{\frac{7}{8}} v_0^{-\frac{7}{8}}$ $MSE(\hat{\sigma}^2(x))$, which means the asymptotic relative efficiency*

$$\frac{MSE(\tilde{\tilde{\sigma}}^2(x))}{MSE(\hat{\sigma}^2(x))} = (\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2})^{-\frac{7}{8}} v_0^{\frac{7}{8}} \geq 1.$$

Remark 3.3.40 *We can deduce from the preceding corollaries that the optimal bandwidths for the original modal volatility estimator and the variance reduced modal volatility estimator differ by a constant factor that depends only on the known v_0 and β. To choose the data-driven bandwidths in practice, following Cheng et al. (2007), we set $\beta = 1$ as default value and let*

$$\tilde{\tilde{h}}_1 = (\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2})^{1/8} \hat{h}_1 \text{ and } \tilde{\tilde{h}}_2 = (\frac{v_0}{2} - \frac{C(\beta)}{8} - \frac{D(\beta)}{2})^{1/8} \hat{h}_2$$

on the basis of asymptotic optimal bandwidth expressions.

3.4 Numerical Examples

To further explain the difference between mean and modal volatility estimators as well as to demonstrate the advantage of applying the variance reduction technique, we carry out several simulation studies and real data analyses in this section, where we compare various estimation methods, i.e., local linear mean estimation, local linear modal estimation, and variance reduced modal estimation. To keep all estimates positive, we take the value zero whenever a negative estimate is obtained, so that $\hat{\sigma}^2(x) = \max(\hat{\sigma}^2(x), 0)$. The Gaussian kernel is used for all examples and the bandwidth selection for modal estimation in practice is accomplished using the steps introduced in the preceding sections. For the local linear mean estimation, we utilize the cross-validation technique to select the bandwidths. One additional simulation related to variance reduction and the prediction advantage of modal regression is listed in the appendix, in which we show that the variance reduction technique indeed works well for all univariate functions.

3.4.1 Monte Carlo Experiments

We conduct two simulation studies, one with independent observations and the other with nonlinear time series. Following Fan and Yao (1998), we use the mean absolute deviation error ($MADE$) to evaluate the performance of the estimator

$$MADE(\sigma_E^2(x_t)) = \frac{1}{n_{grid}} \sum_{t=1}^{n_{grid}} |\sigma_E^2(x_t) - \sigma^2(x_t)|,$$

where $\sigma_E^2(x_t)$ denotes the corresponding estimate at time t, $\sigma^2(x_t)$ represents the true volatility function, and the lowest $MADE$ value means the best fit. Since the data near the boundary may be very sparse, we take $\{x_t, t = 1, \cdots, n_{grid}\}$ as grid points in the range

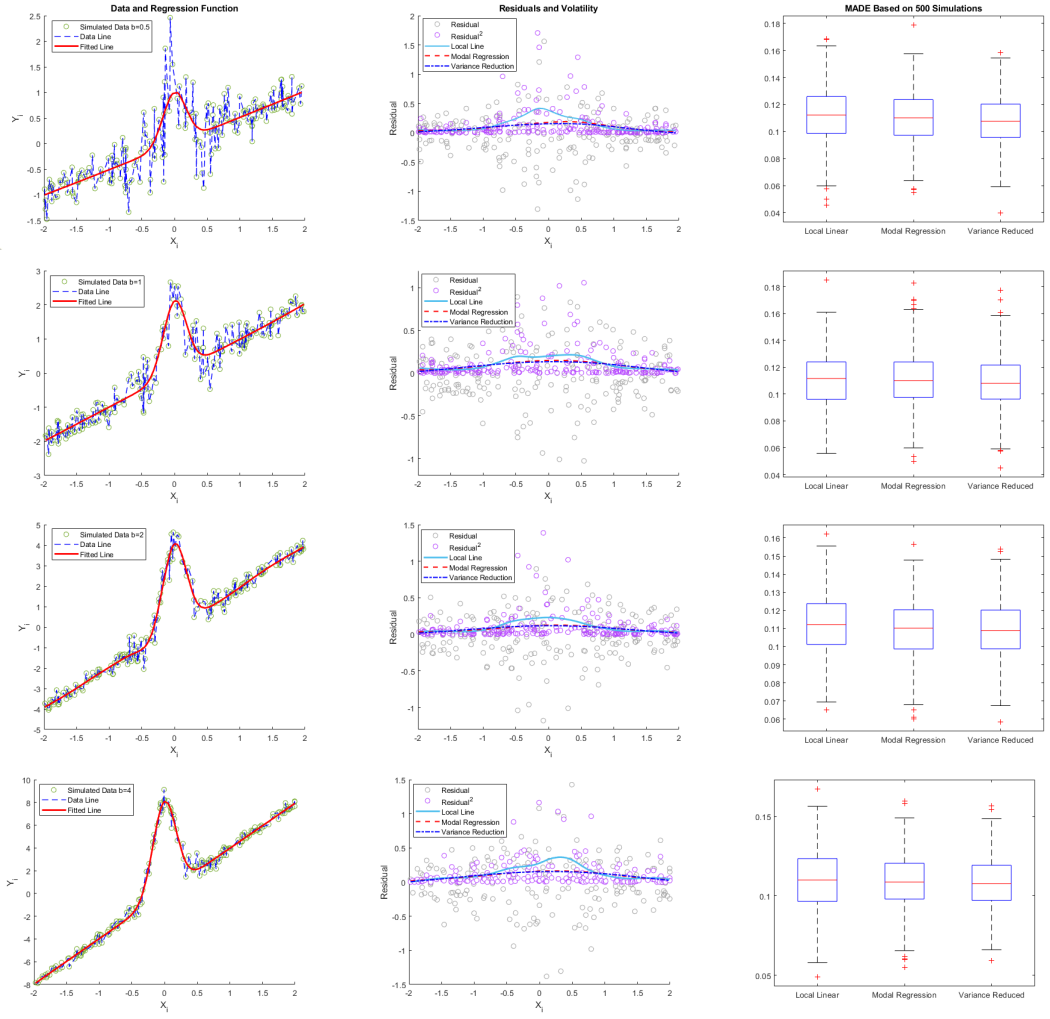

Figure 3.1: Simulation Results of Example 1

Note: the first row is for the case $b = 0.5$; the second row is for the case $b = 1$; the third row is for the case $b = 2$; and the last row is for the case $b = 4$. For each row, the left plot shows the simulated data, true mean regression line, and mean fitted line; the middle plot depicts the residual and squared residual points, as well as the fitted results for three different estimated volatility functions; and the last plot represents the boxplot results of $MADE$ for the considered three different volatility estimators.

of x and $n_{grid} = 101$. We consider the following models with simulated 500 random samples of size $n = 200$ for each of the settings.

(1) Example 1 The data are generated from the following heteroskedastic regression model

$$Y_t = b(X_t + 2\exp(-16X_t^2)) + \sigma(X_t)\varepsilon_t.$$

We let $\sigma(X_t) = 0.4\exp(-2X_t^2)+0.2$, where X_t is generated from uniform distribution $U[-2, 2]$, and ε_t is independent of X_t and follows $N(0, 1)$. Four different values of b are considered, i.e., $b \in \{0.5, 1, 2, 4\}$, which corresponds to the model setting in Fan and Yao (1998).

Figure 3.1 displays the estimation results, where the right plot represents the $MADE$ boxplots of the estimators, indicating that the modal volatility estimator does not vary with different values of b and shows the same property as the mean volatility estimator. The fitted lines of different volatility functions in Figure 3.1 reveal that the modal volatility function can be utilized as a complement to the mean volatility function to indicate risk. Nevertheless, there is not much difference between the local linear modal volatility estimator and the variance reduced modal volatility estimator with regard to fitted lines. Additionally, Figure 3.1 shows that the modal volatility estimator has a smaller $MADE$ compared to the mean volatility estimator, and the variance reduced modal volatility estimator performs better than the local linear modal volatility estimator in terms of $MADE$. Such findings are not surprising, and they are in accordance with the theoretical results presented in Section 3.3. The simulation results also suggest that $\beta = 1$ is indeed an appropriate default value for the variance reduced modal estimator in univariate regression.

(2) Example 2 We consider the nonlinear time series model following Fan and Yao (1998) and Yao and Tong (1994), such that

$$Y_t = 0.235X_t(16 - X_t) + e_t,$$

where $e_t \sim t(3)$ is independent of X_t, and Y_t represents the lag value of X_t. Because the variance in the one-step-ahead case is constant, we consider the two-step-ahead and three-step-ahead cases, i.e., $Y_t = X_{t+2}$ and $Y_t = X_{t+3}$, respectively. Notice that the volatility functions are not constant for these two-step and three-step cases, and both are dependent on covariate X_t.

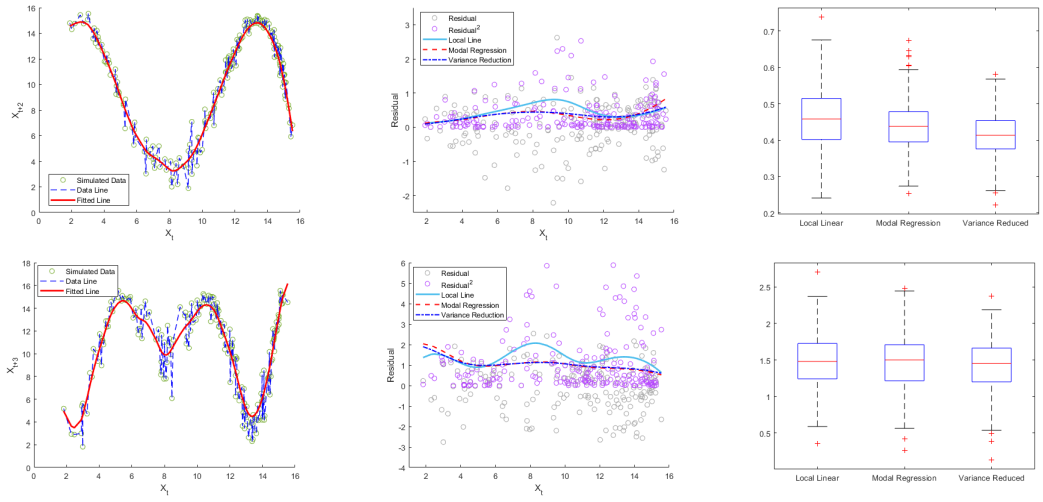

Figure 3.2: Simulation Results of Example 2

Note: the first row is for the case $Y_t = X_{t+2}$, whereas the second row is for the case $Y_t = X_{t+3}$. For each row, the left plot represents the simulated data, true mean regression line, and mean fitted line; the middle plot depicts the fitted results for three different estimated volatility functions, along with the residual and squared residual points; and the last plot shows the boxplot results of $MADE$ for the considered three different volatility estimators.

The estimation results are reported in Figure 3.2, in which the left plots depict the shape of the mean regression function and the corresponding local linear mean fitted

line, and the middle plots show that the modal volatility function is smoother, indicating less risk than the mean volatility function. The middle plots also demonstrate that modal estimation is intended to capture the "most likely" values and can be utilized to reveal mode preference. Figure 3.2 conveys similar conclusions as the previous example, where the local linear modal volatility estimator and the variance reduced modal volatility estimator show remarkable superiority compared to the mean volatility estimator in terms of $MADE$. In line with the theoretical results, the variance reduced modal volatility estimator can achieve a comparable bias while having a smaller variance compared to the local linear modal volatility estimator.

3.4.2 Real Data Analyses

We in this part use two real examples to demonstrate the practical application of the proposed modal volatility and variance reduced modal volatility, which can be utilized to expose a variety of distinct characteristics of data.

(1) Interest Rate To establish a plausible connection between risk and the modal return on financial assets, we consider the yields of the three month Treasury Bill from the secondary market rates, which are annualized using a 360-day year of bank interest and quoted on a discount basis. For comparing the results to those in Fan and Yao (1998), we choose weekly observations from January 5, 1962 to March 31, 1995 as well (Figure 3.3). The total number of observations is 1735. At first, we fit the data using an AR(5) model with the order selected by the Akaike information criterion

$$z_t = \underset{(0.011)}{1.082} z_{t-1} - \underset{(0.018)}{0.045} z_{t-2} + \underset{(0.017)}{0.015} z_{t-3} + \underset{(0.016)}{0.030} z_{t-4} - \underset{(0.012)}{0.083} z_{t-5} + Y_t, \tag{3.14}$$

where z_t represents the interest rate series and Y_t denotes the residual. The values listed below each coefficient are standard errors. The results of (3.14) are consistent with those of Fan and Yao (1998). Then, we can obtain the mean regression function $\hat{m}(x) = E\{Y_t \mid z_{t-1} = x\}$ and the conditional volatility of Y_t given $z_{t-1} = x$. The overall fitted model is

$$z_t = \hat{m}(z_{t-1}) + 1.082z_{t-1} - 0.045z_{t-2} + 0.015z_{t-3} + 0.030z_{t-4} - 0.083z_{t-5} + \hat{\sigma}(z_{t-1})\varepsilon_t. \quad (3.15)$$

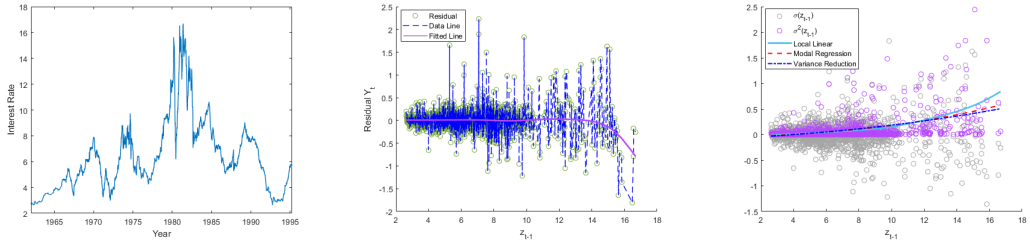

Figure 3.3: Results for Three-Month Treasure Bill Data

Note: the left one is the plot of real data; the middle one indicates the residual point and fitted line of the mean regression function; and the right one represents the fitted results for the considered three different volatility functions.

The main estimation results are shown in Figure 3.3, from which we can see that modal regression is able to capture the majority of data points and disclose the characteristics of data that mean regression cannot reveal (capture the "most likely" effect). When the interest rate is smaller than 12, the modal regression produces nearly identical results to the mean regression. In contrast, when the value of the interest rate is beyond 12, the results from modal regression suggest a lower level of risk in the return of financial assets. The variance reduced modal volatility estimator is not appreciably different from the modal volatility estimator, since the primary function of the variance reduction method is to de-

crease variance while maintaining asymptotic bias constant. This example shows that for some individuals who have a mode preference rather than the classical expectation/mean preference, the modal volatility should be regarded as a risk indicator for risk management.

(2) Motorcycle Data To further illustrate the proposed modal volatility, we consider the well-known motorcycle data from Silverman (1985), in which the dependent variable is the acceleration force on the head of the rider (in gram) and the independent variable is the time after a simulated impact (in milliseconds) with motorcycles. The data are available in the software R library MASS. The sample size is 133. Following Chen et al. (2009), we model both the mean and volatility of acceleration as nonparametric functions of time

$$Acceleration = m(Time) + \sigma(Time)\varepsilon, \tag{3.16}$$

where the volatility function is estimated with the squared residuals.

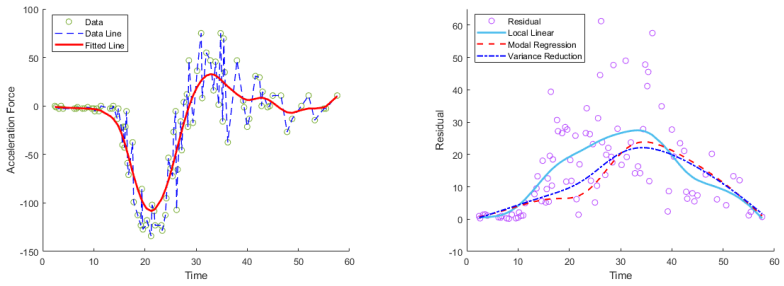

Figure 3.4: Results for Motorcycle Data

Note: the left one indicates the data points and the fitted line of mean regression function; and the right one represents the fitted results for the considered three different volatility functions.

The estimation results are reported in Figure 3.4, where the left one depicts the original data and the local linear mean regression estimate and the right one reveals that

both modal and mean regressions can finely explore the underlying heteroskedastic structure of the dataset and reveal a downward U-shape structure. However, the data features indicated by modal and mean estimations are different. The modal volatility estimators are significantly smoother than the mean volatility estimator. Compared to mean regression, the variability is smaller for modal regression when the time is not beyond 40. Nevertheless, there is a larger value of modal variability when the time is between 40 and 52. In addition, there exists an obvious downward-upward trend in modal volatility around the time of 20, while a similar downward trend appears in mean volatility around the time of 45. In accordance with the previous findings, the variance reduced modal volatility estimator has a fitted line that is quite close to that of the modal volatility estimator, but with a smaller amount of variation.

3.5 Exponential Modal Volatility Estimator

Local linear modal regression performs well in estimating the modal volatility function. However, it is not guaranteed that all volatility estimators based on (3.6) with finite samples are positive. Generally, large or small bandwidths can lead to negative volatility estimates, which are commonly observed at design points with fewer observations, and thus we need to take a further step to correct for negativity; see the numerical results in Section 3.4. Despite the fact that we can apply the local constant approach, the statistical property does not appear promising. To enlarge the applicability of the modal volatility and ensure the guarantee of positivity, we generalize the results in Ziegelmann (2002) to introduce a local exponential modal volatility estimator and establish its asymptotic property under

mild regularity conditions, where we show that the proposed exponential modal volatility estimator can achieve a smaller asymptotic bias compared to the local linear modal volatility estimator in some cases.

Particularly, when an unknown function $r(X)$ is positive almost surely, for any given X_t that is close to x, we can approximate it as

$$r(X_t) = \exp(\log(r(X_t))) \approx \exp(\beta_1(x) + \beta_2(x)(X_t - x)), \tag{3.17}$$

where $\beta_1(x) = \log(r(x))$ and $\beta_2(x) = \partial \log(r(x))/\partial x = r^{-1}(x)\partial r(x)/\partial x$. On the basis of this, after obtaining the estimate of r_t from the mean regression in (3.3), we maximize the following local kernel-based objective function

$$Q_n(\beta_1(x), \beta_2(x)) = \frac{1}{nh_3h_4} \sum_{t=1}^{n} \phi \left(\frac{\hat{r}_t - \Psi(\beta_1(x) + \beta_2(x)(X_t - x))}{h_3} \right) K \left(\frac{X_t - x}{h_4} \right) \tag{3.18}$$

with respect to $\beta_1(x)$ and $\beta_2(x)$, where the function $\Psi(\cdot)$ takes the form $\Psi(s) = \exp(s)$ and the bandwidths $h_3 = h_3(n)$ and $h_4 = h_4(n)$ are approaching zero as the sample size $n \to \infty$. With the estimators $\hat{\beta}_1(x)$ and $\hat{\beta}_2(x)$, we can then define the exponential modal volatility estimator as $\hat{\sigma}_e^2(x) = \exp(\hat{\beta}_1(x))$, which is always positive, and the derivative as $\hat{\dot{\sigma}}_e^2(x) = \exp(\hat{\beta}_1(x))\hat{\beta}_2(x)$. When $\Psi(s) = s$, the above objective function is identical to (3.6). In fact, according to Mishra et al. (2010), we can substitute the exponential function with any well-defined monotone function that has at least two continuous derivatives on its support.

It should be highlighted that the application of local exponential estimation may be restricted by its computational complexity, as maximizing (3.18) does not have an explicit solution. Because $\Psi(\cdot)$ is a nonlinear function, the previous MEM algorithm cannot be used without adaptation. Following Ullah et al. (2022) and Algorithm 2, we propose a nonlinear

MEM algorithm with the help of Taylor expansion to simplify computations (Algorithm 3). All of the comments for Algorithm 2 are carried over here as well.

Algorithm 3 MEM Exponential Volatility Algorithm

E-Step. Define $\theta = (\beta_1(x), \beta_2(x))$. Calculate the weight $\pi\left(t \mid \theta^{(g)}\right), t = 1, \cdots, n$ as

$$\pi\left(t \mid \theta^{(g)}\right) = \frac{\phi\left(\frac{\hat{r}_t - L(X_t - x, \theta^{(g)})}{h_3}\right) K\left(\frac{X_t - x}{h_4}\right)}{\sum_{t=1}^{n} \phi\left(\frac{\hat{r}_t - L(X_t - x, \theta^{(g)})}{h_3}\right) K\left(\frac{X_t - x}{h_4}\right)} \propto \phi\left(\frac{\hat{r}_t - L(X_t - x, \theta^{(g)})}{h_3}\right) K\left(\frac{X_t - x}{h_4}\right).$$

Taylor Expansion. Approximate $L(X_t - x, \theta)$ by a first order Taylor expansion around $\theta^{(g)}$

$$L(X_t - x, \theta) \approx L(X_t - x, \theta^{(g)}) + \frac{\partial L(X_t - x, \theta)}{\partial \theta^T}\Big|_{\theta = \theta^{(g)}} \left(\theta - \theta^{(g)}\right).$$

M-Step. Update $\theta^{(g+1)}$ by

$$\theta^{(g+1)} = \arg\max_{\theta} \sum_{t=1}^{n} \left\{ \pi\left(t \mid \theta^{(g)}\right) \log \frac{1}{h_3} \phi\left(\frac{\hat{r}_t - L(X_t - x, \theta)}{h_3}\right) \right\} =$$

$$\left[\sum_{t=1}^{n} \pi\left(t \mid \theta^{(g)}\right) \frac{\partial L(X_t - x, \theta^{(g)})}{\partial \theta} \frac{\partial L(X_t - x, \theta^{(g)})}{\partial \theta^T}\right]^{-1} \left[\sum_{t=1}^{n} \pi\left(t \mid \theta^{(g)}\right) \frac{\partial L(X_t - x, \theta^{(g)})}{\partial \theta} \hat{r}_t^{(g)}\right]$$

where g is the iteration indicator and $\hat{r}_t^{(k)} - L(X_t - x, \theta^{(g)}) + \frac{\partial L(X_t - x, \theta^{(g)})}{\partial \theta^T} \theta^{(g)}$.

Iterate. Given the initial values, iterate E-Step and M-Step repeatedly until a stopping criteria is satisfied, i.e., $\|\theta^{(g+1)} - \theta^{(g)}\| < \tau$ for a small tolerance value, say $\tau = 10^{-5}$.

To present the asymptotic results, we define

$$L(X_t - x, \theta) = \Psi(\beta_1(x) + \beta_2(x)(X_t - x)) \tag{3.19}$$

and $L^{(i)}(X_t - x, \theta) = (\partial/\partial(X_t - x))^i L(X_t - x, \theta)$. It is simple to verify that $L(0, \theta) = \Psi(\beta_1(x))$, $L^{(1)}(0, \theta) = \Psi(\beta_1(x))\beta_2(x)$, and $L^{(2)}(X_t - x, \theta) = \beta_2^2(x)\Psi(\beta_1(x) + \beta_2(x)(X_t - x))$. We can then rewrite the objective function (3.18) as

$$\frac{1}{nh_3h_4}\sum_{t=1}^{n}\phi\left(\frac{\hat{r}_t-\Psi(\beta_1(x)+\beta_2(x)(X_t-x))}{h_3}\right)K\left(\frac{X_t-x}{h_4}\right)=\frac{1}{nh_3h_4}\sum_{t=1}^{n}$$
$$\phi\left(\frac{\hat{r}_t-\exp(\beta_1(x))-\beta_2(x)\exp(\beta_1(x))(X_t-x)-2^{-1}L^{(2)}(\lambda_t(X_t-x),\theta)(X_t-x)^2}{h_3}\right)$$
$$K\left(\frac{X_t-x}{h_4}\right), \tag{3.20}$$

where $\lambda_t \in [0,1]$. The asymptotic theorem shown below is followed.

Theorem 3.5.17 *With $nh_4^5h_3^3 = O(1)$ and $nh_4h_3^7 = O(1)$, under the regularity conditions C1-C5, as $n \to \infty, h_3 \to 0, h_4 \to 0$, $h/h_4 \to 0$, $h_4^2/h_3 \to 0$, and $nh_4h_3^5 \to \infty$, we have*

$$\sqrt{nh_4h_3^3}\left(\hat{\sigma}_e^2(x)-\sigma^2(x)-\left(\frac{h_4^2}{2}\mu_2\left(\ddot{\sigma}^2(x)-L^{(2)}(0,\theta)\right)-\frac{h_3^2}{2}\frac{g_\epsilon^{(3)}(0\mid x)}{g_\epsilon^{(2)}(0\mid x)}\right)\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,\frac{g_\epsilon(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0\mid x)^2f_X(x)}v_0\right).$$

If we allow $nh_4^5h_3^3 \to 0$ and $nh_4h_3^7 \to 0$, the asymptotic theorem becomes

$$\sqrt{nh_4h_3^3}\left(\hat{\sigma}_e^2(x)-\sigma^2(x)\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{g_\epsilon(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{g_\epsilon^{(2)}(0\mid x)^2f_X(x)}v_0\right).$$

Theorem 3.5.17 indicates that the exponential modal volatility estimator is asymptotically fully adaptive to the unknown mean regression function, i.e., the asymptotic results are the same as those in the case where $m(\cdot)$ is known. The estimator $\hat{\sigma}_e^2(x)$ can asymptotically ignore the bias term by undersmoothing ($\lim_{n\to\infty} nh_4^5h_3^3 = 0$ and $\lim_{n\to\infty} nh_4h_3^7 = 0$) at the expense of the increase in variance. However, the exponential modal volatility estimator is not fully equivalent to the local linear modal volatility estimator as it estimates the logarithm of the volatility rather than the volatility itself. If $\Psi(s) = s$, the asymptotic results will reduce to Theorem 3.3.14. Comparing the results to those of the local linear modal volatility estimator $\hat{\sigma}^2(x)$, it is clear that the two estimators have the same convergence rate.

Also, the two estimators share exactly the same asymptotic variance but different biases. The difference is governed by the term related to h_4, which is $2^{-1}h_4^2\mu_2(\ddot{\sigma}^2(x)-L^{(2)}(0,\theta))$ for the exponential estimator and $2^{-1}h_4^2\mu_2\ddot{\sigma}^2(x)$ for the linear estimator. Because $L^{(2)}(0,\theta)$ is a nonnegative quantity, we conclude that the exponential modal volatility estimator $\hat{\sigma}_e^2(x)$ could have a smaller bias compared to $\hat{\sigma}^2(x)$ when $\ddot{\sigma}^2(x)$ is nonnegative and greater than $L^{(2)}(0,\theta)$. This is an interesting observation since it implies a different approach to reducing variance. For instance, in the case of $\ddot{\sigma}^2(x) = L^{(2)}(0,\theta)$, the bias term associated with h_4 will vanish. Then, we can choose an arbitrarily large value for h_4 to reduce variance as well as achieve a modal parametric convergence rate.

Remark 3.5.41 *We can also apply the variance reduction technique described in Section 3.3 on the exponential modal volatility estimator to achieve variance reduction. As a result of employing the identical processes as before, the asymptotic bias remains unchanged but the asymptotic variance becomes* $(f_X(x)g_\epsilon^{(2)}(0 \mid x)^2)^{-1}g_\epsilon(0 \mid x)\int \tau^2\phi^2(\tau)d\tau\left(v_0 - r^2(1-r^2)C(\beta)\right)$. *In addition, following the result in Remark 3.2.33, we can straightforwardly obtain the asymptotic result for the local exponential modal-based robust volatility estimator by treating bandwidth* h_3 *as a constant, where the asymptotic bias is* $h_4^2 2^{-1}\mu_2\left(\ddot{\sigma}^2(x) - L^{(2)}(0,\theta)\right)$ *but the asymptotic variance is unchanged.*

3.6 Concluding Remarks

Modal regression has grown in popularity because of its ability to fit a large variety of datasets with skewness or heavy tails. However, in the literature, there is no research focusing on the modal volatility estimator, which is directly related to option pricing and

risk measure quantification. It is frequently found that market returns display negative skewness and excess kurtosis. In this paper, we apply nonparametric modal regression on volatility function on the basis of mode value and develop an extension of the variance reduction technique based on Cheng et al. (2007) for the modal volatility estimator. The asymptotic results and the expressions of the optimal bandwidths are presented under some mild conditions. The simulation results and empirical analyses indicate that the introduced estimation method is applicable in practice to complement the existing mean or median volatility estimation. In addition, it can be difficult in practice to find statistics that are both resistant and have the robustness of efficiency. We briefly argue that the modal-based volatility estimator is not only robust but also as asymptotically efficient as the least squares estimator. To avoid negative estimates of the volatility function, we discuss the extension of the proposed method to the local exponential modal estimation without providing numerical justification. Due to the fact that both conditional heteroskedasticity and asymmetric error distributions have been observed on a regular basis in empirical finance and economics, the proposed modal volatility with the relaxation of the symmetry assumption on the error density can be very helpful for practical applications.

Several extensions of the present work are immediate. We concentrate on the stationary data in this paper. It has been recognized for a long time that in financial time series, nonstationarity is an important factor that needs to be considered. In the future, we can let X_t be an integrated or near-integrated process and utilize modal regression to develop some nonstationary volatility models with the assumption that $X_t = (1 - c/n)X_{t-1} + v_t$, where $c \geq 0$ and v_t is generated by $v_t = \varphi(L)\eta_t = \sum_{k=0}^{\infty} \varphi_k \eta_{t-k}$ in which $\varphi_0 = 1, \varphi(1) \neq 0$

with $\sum_{k=0}^{\infty} k|\varphi_k| < \infty$, and η_t are i.i.d. random variables with mean zero and $E|\eta_t|^p < \infty$ for some $p > 2$. Also, long memory structure is a more general dependence structure than mixing. In practice, stock return or exchange rate return series commonly exhibit the long memory property in volatility, as the lag k auto-covariances decays to zero like $k^{-\theta}$ for some $0 < \theta < 1$. It would be an interesting topic that deserves to be researched further in the future.

In addition, it is well-known that nonparametric estimation has a slower convergence rate compared to parametric estimation. For practical purposes, it might be important and interesting to test whether the modal volatility function follows a specified parametric form, which can be formulated as $H_0 : \sigma^2(X) = f(X, \kappa)$, where $f(X, \kappa)$ is a given family of modal functions indexed by an unknown parameter vector κ. We can easily adopt a wild bootstrap approach based on comparing the residual sum of kernel-based objective functions from the restricted and unrestricted estimates. Besides this, we can also develop a test to detect structural changes in modal volatility. The presence of variance change easily confuses traditional time series analysis procedure, leading to incorrect conclusions. Thus, detecting and locating these change points is a critical practice. We leave all of these interesting directions for future research.

Chapter 4

Modal Regression Discontinuity Designs

4.1 Introduction

Regression discontinuity designs, which were originally introduced by Thistlethwaite and Campbell (1960) to investigate the impact of student scholarships on future academic outcomes, have emerged as one of the state-of-the-art quasi-experimental approaches for identifying, estimating, and inferring local treatment effects on the target population in economics, statistics, social science, biomedicine, and other related fields. The most distinctive feature of RD designs is that there exists a continuous variable of interest, known as the running variable, for each unit in the sample that determines the treatment assignment either deterministically or probabilistically. In accordance with the ways of determination by a running variable, the RD designs explored in the literature of causal analysis are divided into

two types—the sharp RD (SRD) design and the fuzzy RD (FRD) design. It is convenient to employ the SRD design when the outcome or dependent variable exhibits discontinuity at a cutoff, in which an individual is allocated to the treatment when the value of the running variable surpasses the given cutoff. Under weak smoothness conditions, the treatment near the cutoff appears almost random (Lee, 2008), enabling researchers to identify and estimate the analogous treatment effect. Significantly different from the SRD design, the treatment of the FRD design is partially influenced by the running variable, and the probability of treatment assignment jumps at the cutoff. Following that, researchers can produce an instrumental variable estimate of the treatment effect. More recently, with the increasing availability of richer datasets, there has been a large amount of theoretical and empirical research investigating RD designs based on the mean or quantile regression with the crucial assumption that units around the cutoff do not systematically differ in their unobservable characteristics (Hahn et al., 2001; Van Der Klaauw, 2008; Frandsen et al., 2012). Due to space constraints, we refer interested readers to the review papers written by Imbens and Lemieux (2008) and Lee and Lemieux (2010), as well as the references therein, for further information on the theoretical details and practical applications of RD designs.

In traditional mean regression, researchers are interested in the conditional expectations of the outcomes given covariates in order to recover the average causal effect of the treatment at the cutoff, which is estimated by $\lim_{X \downarrow \bar{X}} E(Y_1 \mid X) - \lim_{X \uparrow \bar{X}} E(Y_0 \mid X)$ (SRD design) under the assumption of the smoothness of the conditional expectation functions, where $X \in R$ is the running variable having a continuous distribution, $\bar{X}$ is the cutoff, $Y_1 \in R$ denotes the outcome of the treatment group, $Y_0 \in R$ represents the outcome of the

group without treatment, and $X \downarrow \bar{X}$ and $X \uparrow \bar{X}$ mean taking the limits from the right and left sides of $\bar{X}$, respectively. For FRD design in mean regression, the aforementioned limit expression is divided by a similar difference in the conditional expectations of treatment assignment D (defined in Sections 4.2), which is determined by the running variable; see Hahn et al. (2001) for the solid theoretical work exploring the RD designs with the average causal effect in the treatment model framework. To extend the conditional average treatment effect, some researchers propose methods to estimate the conditional τth quantile treatment effect such that $\lim_{X \downarrow \bar{X}} Q_\tau (Y_1 \mid X) - \lim_{X \uparrow \bar{X}} Q_\tau (Y_0 \mid X)$, under the assumptions of rank invariance and the smoothness of the conditional distributions of the potential outcomes, where $Q_\tau(Y \mid X)$ is the τth conditional quantile of Y given X; see Frandsen et al. (2012) for the formal theoretical work in the econometrics literature on identification and estimation in the quantile RD designs. Since RD designs are related to the local causal effects at a certain cutoff of the running variable, it is common to adopt a standard nonparametric regression with local linear approximation to estimate the treatment effects at the boundary points, which places more weight on observations closer to the cutoff. In addition, it is well-known that we in general have three popular location or central tendency measures—mean, median (quantile), and mode. The RD designs should identify not only the mean or quantile treatment effect but also the *mode treatment effect* of policies. However, to the best of our knowledge, almost all research related to RD designs in the literature has a concentration on estimating the conditional average or quantile treatment effect on the basis of the mean or quantile regression, which only provides a partial picture of the effects of the treatment. Also, little is known about the behavior of the regression based on the mode value near the

boundary. This paper attempts to develop a complementary treatment effect estimator in the RD designs for *modal regression at the boundary point* (the cutoff) without making any strong assumptions about the shapes of the regression functions.

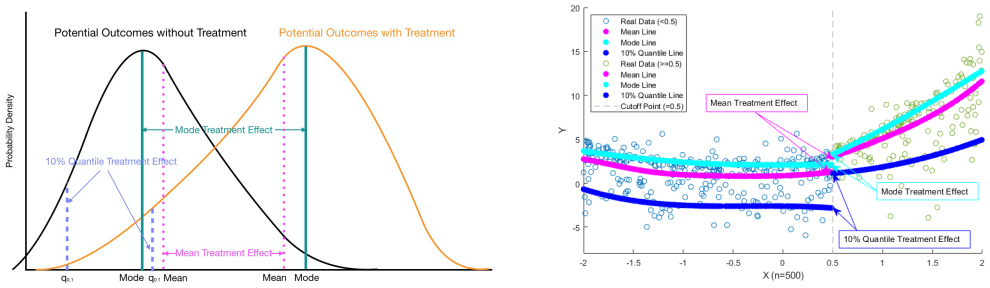

Figure 4.1: Mean, Mode, and Quantile Treatment Effects

Note: the right plot represents the results of DGP 1 in Section 4.3 with $n = 500$; the left plot shows that the proposed mode treatment effect incorporates the standard mean treatment effect as a special case, i.e., with a symmetric distribution, the mode treatment effect is identical to the mean treatment effect. It is feasible to obtain mode treatment effect from quantile treatment effect due to the special connection between mode and quantile via the distribution function, which is the motivation for us to propose a quantile-based estimation method for the CMTE in the modal fuzzy RD design.

Particularly, to fill the literature gap, we devote attention to the mode treatment effect and establish its formal identification, estimation, and statistical properties, which can potentially disclose a novel and intriguing treatment structure that may have been overlooked by the existing treatment effects. An illustration of different treatment effects is depicted in Figure 4.1, where we can observe that the existing mean and quantile treatment effects, as well as the new mode treatment effect, are targeting different population

parameters generally. From a practical perspective, each treatment effect has its own merits and can be used in conjunction with the others. In comparison to the mean or quantile treatment effect, the mode treatment effect has two appealing features—*it represents the "most likely" effect and can be resistant to outliers, heavy-tailed distributions, and certain types of measurement error*. In reality, when the data have a skewed distribution (such as salaries, prices, and expenditures), it is more informative to investigate the mode treatment effect due to the representative of the benefit of the policy to the majority of people. For example, the mode treatment effect can typically better measure whether the implementation of an income-raising policy has benefited most people from policy-making views, whereas the mean treatment effect will be influenced much more by the single high income increase due to the policy, and the quantile treatment effect cannot directly reflect the "most likely" effect and could result in producing low density point predictions. Also, a treatment effect with a zero mean may obscure a significant offsetting effect in the mode view. All of these indicate the necessity of investigating the mode treatment effect. It is important to mention that the proposed CMTE in this paper has a close relationship with the previous theoretical work on mode estimation, which has attracted much more attention in statistics and econometrics accompanied by substantial theoretical contributions for estimation and inference recently. Among the notable results in this line of modal regression work include Lee (1989, 1993), Kemp and Santos Silva (2012), Yao and Li (2014), Chen et al. (2016), Yao and Xiang (2016), Krief (2017), Chen (2018), Ota et al. (2019), Feng et al. (2020), Kemp et al. (2020), Ullah et al. (2021, 2022), among others.[1] However, the objectives of

[1]The mode, as one of the three measures of central tendency (the other two being the mean and the median (quantile)), offers an important description of data in many analytical scenarios but has received little attention in the regression literature, in part due to the computational complexity involved in calculations.

those studies are completely different from the ones of this paper, which focuses on modal boundary estimation. By providing a systematic statistical analysis of the modal estimator at the boundary point, we contribute significantly to the large and still rapidly growing literature on mode estimation and treatment effects in the RD designs.

Paralleling the traditional mean or quantile treatment effect, the purpose of this paper is to propose a novel (local) CMTE in the RD designs

$$\tau_{RD} = Mode\left(Y_1 \mid X = \bar{X}\right) - Mode\left(Y_0 \mid X = \bar{X}\right), \tag{4.1}$$

which provides a more valuable measure for skewed or heavy-tailed data and can be considered as an appealing complement to the existing treatment effects. Nevertheless, making causal interpretations of the CMTE in the form (4.1) is perhaps more challenging than the mean treatment effect, since the mode is not an additive parameter in the vast majority of instances. As a consequence, $Mode(Y_1 - Y_0 \mid X = \bar{X}) \neq Mode(Y_1 \mid X = \bar{X}) - Mode(Y_0 \mid X = \bar{X})$ in general. The expression $Mode(Y_1 - Y_0 \mid X = \bar{X})$ has a causal interpretation, which is a measure of the mode benefit to the individual from being a member of the treatment group, while the expression $Mode(Y_1 \mid X = \bar{X}) - Mode(Y_0 \mid X = \bar{X})$ represents the difference in modes between two distributions, which measures the effect of a policy on the mode of the distribution of the outcome of interest rather than the effects of treatment on a typical individual. To satisfy the above equation, we can impose a deterministic relation-

Modal regression, which is based on the principle of mode, seeks to determine the "most likely" conditional value (mode) of a dependent variable Y given covariates X, denoted by $Mode(Y \mid X)$. Let $f_{Y|X}(Y \mid X)$ be the conditional density function of Y given X and $f_{Y,X}(Y, X)$ be the joint density function, we can write the estimator of the conditional mode of Y given X as

$$Mode(Y \mid X) = \arg\max_{Y} f_{Y|X}(Y \mid X) \propto \arg\max_{Y} f_{Y,X}(Y, X),$$

where " $\propto$" means "is proportional to". However, owing to the "curse of dimensionality", such a modal regression based on the distribution function is difficult to implement when the dimension of the covariates is large. To address this issue, Kemp and Santos Silva (2012) and Yao and Li (2014) allowed a more general kernel function and established the consistency of the linear modal regression estimator under very mild conditions even when the error density is skewed; see Yao and Li (2014) and Ullah et al. (2021, 2022) for the summary of the advantages of modal regression compared to the existing regressions.

ship between Y_1 and Y_0. For instance, suppose that $Y_1 = \mu + \sigma Y_0$ with $0 < \mu, \sigma < \infty$. Then, $Mode(Y_1 - Y_0) = \mu + (\sigma - 1)Y_0 = Mode(Y_1) - Mode(Y_0)$, implying a constant mode treatment effect. However, this determinism constraint is overly strict, and the constant effect in practice may be uninteresting.

To allow the mode effect to be equal to the individual treatment effect, we require that the *mode rank* of an individual remains the same regardless of being treated or not throughout the whole paper, which is referred to as *mode rank invariance*. Without this condition, the differences of the potential outcome distributions at mode may be zero, but the true treatment effects are not zero due to the individuals moving up and down in the distribution. To better elaborate this mode rank condition, we assume that the potential outcomes $(Y_{1,i}^m, Y_{0,i}^m)$ of individual i correspond to the quantiles $(\xi_{1,i}^m, \xi_{0,i}^m)$ of the conditional distributions of Y_1 and Y_0 given X that achieve the mode points, respectively. Following that, the mode effect of the policy on individual i is represented by

$$\tau_m = Y_{1,i}^m - Y_{0,i}^m = F_1^{-1}(\xi_{1,i}^m) - F_0^{-1}(\xi_{0,i}^m), \tag{4.2}$$

where $F(\cdot)$ denotes the cumulative distribution of the outcome. Nonetheless, this individual effect will never be attained because we cannot estimate $F_1^{-1}(\xi_{1,i}^m)$ and $F_0^{-1}(\xi_{0,i}^m)$ simultaneously. We therefore rewrite the effect on individual i as

$$\tau_m = Y_{1,i}^m - Y_{0,i}^m = \underbrace{F_1^{-1}(\xi_{1,i}^m) - F_0^{-1}(\xi_{1,i}^m)}_{\text{mode treatment effect}} + \underbrace{F_0^{-1}(\xi_{1,i}^m) - F_0^{-1}(\xi_{0,i}^m)}_{\text{flexibility effect}}, \tag{4.3}$$

where the first difference is a mode treatment effect and the second difference is a flexibility effect, capturing the change in outcomes caused by the movement of individuals to different modes within the same distribution. We would have $\tau_{RD} = \tau_m$ if everyone remains the same

mode rank in the corresponding distributions. That is, the flexibility effect is zero if $\xi^m_{1,i} = \xi^m_{0,i}$ for all individuals. This *mode rank invariance* condition is weaker than the usual rank invariance restriction imposed in the quantile RD designs, which requires an individual's rank in the potential outcome distribution to be the same across treatment states. If the usual rank invariance condition is met, then all features of the rank distributions, including mode, would be the same.

As a matter of exposition, we focus primarily on the SRD design to illustrate the main idea of this paper, and briefly address the extension of the results to the FRD design in the end. Under certain mild conditions, the estimation of the mode treatment effect is equivalent to the problem of estimating the magnitude of a discontinuity in a conditional mode. For the purpose of simplicity, we restrict our attention to the running variable rather than other covariates, as is common in most RD studies, and make a comment on the effect of including additional covariates in the modal SRD design in Section 4.4. Furthermore, it is discovered that the modal regression line is identified to the mean regression line whenever the distribution of the data is symmetric. Compared to the mean, it is commonly recognized that the mode is less susceptible to outliers and heavy-tailed distributions. If we investigate treatment effects with symmetric data based on the proposed modal RD designs, we have $\tau_{RD} = Mode(Y_1 - Y_0 \mid X = \bar{X}) = E(Y_1 - Y_0 \mid X = \bar{X}) = Mode(Y_1 \mid X = \bar{X}) - Mode(Y_0 \mid X = \bar{X}) = E(Y_1 \mid X = \bar{X}) - E(Y_0 \mid X = \bar{X})$, which can be interpreted as the modal-based robust causal effect without the use of any rank invariance assumptions. Nevertheless, the estimation procedure and asymptotic properties of such a modal-based robust treatment effect are significantly different from those of the mode treatment effect.

We thus concentrate on the asymmetric case in this paper and only present some simulation results for illustrating modal-based robust treatment effect in the appendix.[2]

In light of the fact that the estimated CMTE in the SRD design only applies to individuals who have $X = \bar{X}$, it is essential to assess the overall stability of the CMTE estimate. Along with that, we demonstrate that when modal regression functions are subject to the continuous differentiability condition, we can nonparametrically identify the *derivative of the CMTE*

$$\tau_{RD}^{(1)}(\bar{X}) = \frac{\partial \tau_{RD}(\bar{X})}{\partial \bar{X}} = \frac{\partial \left[Mode\left(Y_1 \mid X = \bar{X}\right) - Mode\left(Y_0 \mid X = \bar{X}\right) \right]}{\partial \bar{X}}, \tag{4.4}$$

which can be utilized to measure the impact of small discrete changes in the running variable on the treatment effect and test the external validity or generality of the estimated CMTE. Notice that $\tau_{RD}^{(1)}(\bar{X}) = 0$ is a crucial condition for the mode treatment effect to remain unchanged regardless of the running variable. We then simply need to check the magnitude of the derivative of CMTE to verify the external validity. In addition, to reflect the impact of changing the cutoff, we propose a *modal marginal cutoff treatment effect* (MMCTE) and show that with the local policy invariance (defined in Lemma 4.2.3), the MMCTE is identified as $\tau_{RD}^{(1)}(\bar{X})$. This allows us to apply a Taylor expansion to provide an approximate estimate of the effect of a discrete change in the cutoff. Because the focus of this paper is on the CMTE, we only present theoretical analyses of the derivative and the MMCTE without any numerical examples.

In recent RD literature, nonparametric local estimation has gained considerable attention and has emerged as the preferred method for estimating RD treatment results.

[2]When the data contain outliers or have a heavy-tailed distribution, utilizing modal estimation instead of mean estimation will yield robust and efficient estimators. In this case, we must regard the bandwidths associated with error terms in the matching kernel functions as constants in order to attain a mean convergence rate. Such a modal-based robust treatment effect is investigated separately in other research.

In order to maintain the generality and sufficiently flexible modeling of function forms, the proposed CMTE estimator in the SRD design is also estimated by local linear modal regression because of its superior performance near the boundary, provided that the bandwidths decay towards zero asymptotically at an approximate rate. Since the cutoff can be assimilated to the boundary point, the fact that the modal estimator has no boundary effects is a particularly appealing property when dealing with RD designs. To be more specific, the developed modal estimation methodology involves a weighted local linear regression to approximate the modal function above and below the cutoff with weights calculated by the application of a kernel function at the distance between each observation and the cutoff. We adopt a so-called modified MEM algorithm by virtue of the normal kernel function to efficiently obtain the numerical solutions for estimators. We develop modal identification and present asymptotic theorems for the CMTE estimator in the SRD design under modest assumptions, where we show that the resulting estimator is consistent at the nonparametric modal rate. Consequently, the bias complexity of the suggested estimator is no worse than that of the interior point estimator. Due to the use of a tiny fraction of total observations around the mode, the CMTE estimator has a slower convergence rate $n^{-1/4}$ (with the mean squared error (MSE) optimal bandwidths) compared to the estimators of the mean and quantile treatment effects.

After achieving the asymptotic theorem, it is natural to construct a confidence interval for the suggested CMTE estimate. In general, one can apply the asymptotic normality result to consistently estimate the asymptotic bias and variance. However, due to the ignorance of the extra variability produced by the bias term and the unknown quantities

in the bias and variance terms, it is neither simple nor valid to construct a Wald-type confidence interval for the CMTE based on the asymptotic limiting theorem directly. To build a reliable confidence interval in the modal SRD design, we develop an effective bootstrap procedure for the practical application relying on undersmoothing (i.e., choosing a bandwidth that vanishes at a rate that is faster than the MSE-optimal bandwidth). Currently, there is no widely accepted method for the selection of optimal bandwidths in modal regression. We in this paper extend the results in Kemp and Santos Silva (2012) to choose the undersmoothed bandwidths on the right and left sides of the cutoff, separately, taking into consideration that the target modal functions are different on both sides of the cutoff, to implement the developed estimation procedure. We then conduct Monte Carlo simulations and an empirical analysis to further investigate the practical application of the developed estimation and inference procedures for CMTE, which demonstrate the good finite sample performance of the suggested procedures. To our best knowledge, there do not exist any studies that provide formal identification, estimation, and inference of CMTE in a general framework for SRD design, where nonparametric modal regression with local boundary estimation is applied.[3] Several potential extensions of the proposed CMTE estimator in the modal SRD design, such as including additional covariates, multiple running variables, and multiple cutoffs, are also discussed in the paper.

[3]We are aware that a recent work by Chang (2020) also discussed the mode treatment effect by providing two estimation methods, the traditional kernel density method and the machine learning method, which is closest to yet substantially different from the current paper. As we pointed out in Footnote 1, modal regression considered from the conditional distribution is practically infeasible when the dimension of covariates is moderate or high due to the "curse of dimensionality". In addition, Chang (2020) did not investigate CMTE for the modal RD designs, smoothness conditions for identification, local linear approximation for estimation, practical implementation of the estimators, or the bootstrap method for the confidence interval. In contrast, the setting in this paper provides more practical guidance on CMTE in the RD designs for theoretical investigation and empirical applications.

Finally, because the mode does not possess the additive property, generalizing the results of the SRD design presented in this paper to the FRD design with a local Wald estimator (instrumental variables type estimator) is not conceptually straightforward, unless we impose a strict symmetric distribution assumption under which the mode is identical to the mean. Providing that the modal regression line coincides with a quantile regression line, we briefly discuss a simple method considering from quantile regression to estimate CMTE for compliers (individuals who receive the treatment if and only if they are eligible) in the modal FRD design without theoretical justification. It should be noted that while this quantile-based estimating approach for the CMTE can be applied to the modal SRD design as well, the theoretical properties are fundamentally different from those described in this paper. Due to space constraints, we will defer to other research for the comprehensive exploration of such an estimating approach and asymptotic theorems. With consideration of both sharp and fuzzy cases, the present paper extends the previous literature on treatment effects to the estimation of mode treatment effect in the modal RD designs.

The remainder of this paper is structured as follows. Section 4.2 formally establishes the econometric framework of the modal SRD design, proposes a nonparametric modal regression with local boundary approximation to estimate CMTE, settles the asymptotic properties of the resulting estimators rigorously, and develops a bootstrap procedure to construct the confidence interval. Section 4.3 presents the numerical results of the CMTE estimator, which contain Monte Carlo simulations as well as an empirical analysis to illustrate how well the proposed estimator performs with finite samples. We discuss various extensions in Section 4.4, including the estimation of CMTE in the FRD design using quan-

tile regression. The paper is concluded in Section 4.5. All technical proofs and additional numerical and theoretical results are included in the appendix.

4.2 Modal Sharp Regression Discontinuity

We focus on the modal SRD design at the boundary point (the cutoff) in this section, using the potential outcome framework to characterize the two underlying counterfactual states (with and without receiving treatment) (Rubin, 1974), where we first establish conditions under which the newly proposed CMTE in the SRD design can be identified, and then introduce a local linear modal estimation method to obtain the estimator. The well-known appealing feature of local linear smoothers, i.e., the behavior near the boundary, is shown to carry over to the mode case. Following this, we investigate the inference of CMTE and construct a confidence interval in practice through a bootstrap procedure depending on undersmoothing.

4.2.1 Econometric Identification

Consider samples $\{(Y_i, X_i, D_i)\}_{i=1}^n$ in a standard RD design, where $Y_i \in supp(Y_i) \subset R$ is the observed outcome for individual i in which $supp(Y_i)$ denotes the support of Y_i, $X_i \in R$, called running variable, is the pretreatment covariate that determines the assignment of treatment with support $[X_l, X_u] \subset R$, and the treatment assignment D_i is completely determined by the running variable, which is equal to 1 if individual i receives treatment and 0 otherwise, i.e.,

$$D_i = \mathbf{1}\left(X_i \geq \bar{X}\right) = \begin{cases} 1 & \text{if } X_i \geq \bar{X}, \\ 0 & \text{if } X_i < \bar{X}, \end{cases} \tag{4.5}$$

where $\mathbf{1}(\cdot)$ is the indicator function that takes the value 1 when the condition within the bracket is true, and the cutoff $\bar{X} \in [X_l, X_u]$ is supposed to be known for econometricians. With the previous definitions of $Y_{1,i}$ and $Y_{0,i}$ in the introduction, we have $Y_i = Y_{1,i}D_i + Y_{0,i}(1 - D_i)$. The objective of causal analysis in this paper is to get knowledge about the features of the conditional mode value of potential outcomes. Therefore, to avoid model misspecification, we assume that the outcome variable Y_i is determined by a function Y of the individual characteristics and treatment such that $Y_i = Y(X_i, D_i, \epsilon_i)$, and use a nonparametric modal regression to explicitly describe the relationship between Y_i and X_i in a reduced form

$$\begin{cases} Y_i = m(X_i) + D_i\tau_{RD} + \epsilon_i, \\ \\ Mode(Y_i \mid X_i) = m(X_i) + D_i\tau_{RD}, \end{cases} \tag{4.6}$$

where $Mode(\epsilon_i \mid X_i, D_i) = 0$ almost surely, $m(\cdot)$ is an unknown baseline effect function that is characterized by regularity conditions near the cutoff, and τ_{RD} is the parameter of interest (CMTE). Throughout the paper, we assume that $\{(Y_i, X_i)\}_{i=1}^n$ are *i.i.d.* observations.

Remark 4.2.42 *Given that D is a deterministic function of X and that there is no variation in the treatment by conditional on X, the unconfounded treatment assignment $Y_{1,i}, Y_{0,i} \perp D_i \mid X_i$ holds trivially, where $\perp$ denotes conditional independence. It states that the treatment assignment D is independent of the potential outcomes Y_1 and Y_0 conditional on X. The distributions $f_{Y_1}(\cdot)$ and $f_{Y_0}(\cdot)$ can then be identified because $f_{Y|D=1,X}(Y \mid X) = f_{Y_1|D=1,X}(Y \mid X) = f_{Y_1|X}(Y \mid X)$.*

Since there is no value of X for which we can observe the potential outcomes $Y_{1,i}$ and $Y_{0,i}$ simultaneously, we need a tractable representation of τ_{RD} in terms of modal

function in (4.6), which can be directly estimated from data. We therefore impose the following assumption for identification without any parametric functional form constraints on $Mode\,(Y \mid X)$.

Assumption 1 *(Unimode and Continuity) The conditional distributions* $f_{Y_1|X}\,(Y \mid X)$ *and* $f_{Y_0|X}\,(Y \mid X)$ *are strictly positive, unimodal, and continuous in* X *for all* Y *such that* $\sup_{Y:|Y-Y_1^m|>\eta} f_{Y_1|X}(Y \mid X) < f_{Y_1|X}\,(Y_1^m \mid X)$ *and* $\sup_{Y:|Y-Y_0^m|>\eta} f_{Y_0|X}(Y \mid X) < f_{Y_0|X}(Y_0^m \mid X)$ *for* $\eta > 0$.

Assumption 1 is a novel and fundamental condition, which is imposed to ensure that both the distributions of Y_1 and Y_0 are unimodal in the presence of X. It indicates that $Mode(Y_1 \mid X)$ and $Mode(Y_0 \mid X)$ are continuous in X for all Y and represents that $E\left[\text{Mode}\left(Y_1 \mid X = \bar{X}\right)\right] = \lim_{X\downarrow\bar{X}} E[\text{Mode}(Y \mid X)]$ and $E\left[\text{Mode}\left(Y_0 \mid X = \bar{X}\right)\right] = \lim_{X\uparrow\bar{X}} E[\text{Mode}(Y \mid X)]$. The continuity of conditional distributions guarantees, both intuitively and informally, that the difference in the mode of outcomes on each side of the cutoff is ascribed to the change in assignment of treatment. The violation of this continuity assumption would suggest that a change in the mode value of Y is driven by a change in X rather than a change in D. In reality, we shall utilize continuity for conditional distributions only at $X = \bar{X}$, but it is uncommon to assume continuity for one value of the covariate but not for others.[4] The underlying mechanism of modal RD designs is illustrated in Figure 4.2. We then have the following lemma.

Lemma 4.2.1 *Under the aforementioned model settings and Assumption 1, by defining* $m_{Y_1}(\bar{X}) = \lim_{X\downarrow\bar{X}} m_{Y_1}(X)$ *and* $m_{Y_0}(\bar{X}) = \lim_{X\uparrow\bar{X}} m_{Y_0}(X)$*, the conditional mode effect of the treatment on the outcome at the cutoff can be identified as*

[4]Assumption 1 implies that the probability density $f_X(X)$ is continuous and strictly positive at $X = \bar{X}$.

$$\tau_{RD} = Mode\left(Y_1 \mid X = \bar{X}\right) - Mode\left(Y_0 \mid X = \bar{X}\right) = \lim_{X \downarrow \bar{X}} Mode\left(Y_1 \mid X\right) - \lim_{X \uparrow \bar{X}} Mode\left(Y_0 \mid X\right)$$

$$= \lim_{X \downarrow \bar{X}} Mode(Y \mid X) - \lim_{X \uparrow \bar{X}} Mode(Y \mid X) = m_{Y_1}(\bar{X}) - m_{Y_0}(\bar{X}),$$

where the second and the forth equations follow under Assumption 1, and the third equation is a consequence of $Y = Y_1 D + Y_0(1 - D)$ *and* $D = \mathbf{1}\left(X \geq \bar{X}\right)$.

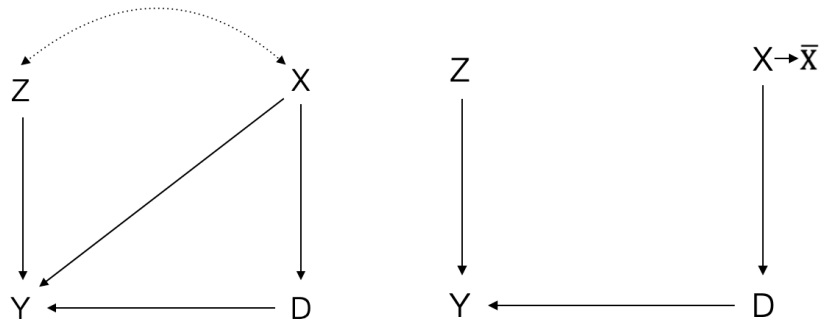

Figure 4.2: Modal Regression Discontinuity

Note: it depicts the mechanism of modal RD designs, with Z representing an extra covariate that may influence outcomes. Since the treatment state is determined entirely by the assignment rule and is independent of Z given X, there is no arrow from Z to D. The left plot shows that the running variable confounds the relation $D \to Y$ as X influences both D and Y. The right plot indicates that under the continuity assumption, there will be no arrow from X to Y, and the treatment effect can be identified since the relation $D \to Y$ is neither confounded by X nor by Z.

Remark 4.2.43 *Throughout the paper, whenever we take a limit on a function, we implicitly presume that this limit exists with almost certainty and is finite. By releasing the unique global mode assumption, the suggested approach can also be applied to the multi-mode treatment effects setting with a local mode rank invariance condition. In particular,*

we can estimate different local modal regression lines to reveal the treatment effect structure for clustered or inhomogeneous data when the population consists of multiple homogeneous latent sub-populations (heterogeneous treatment effects). To a large extent, the proposed CMTE can only be estimated at the cutoff and lacks generalizability, owing to the structure that there is no overlap in X between the treatment and the control groups and the counterfactual cannot be accessed.

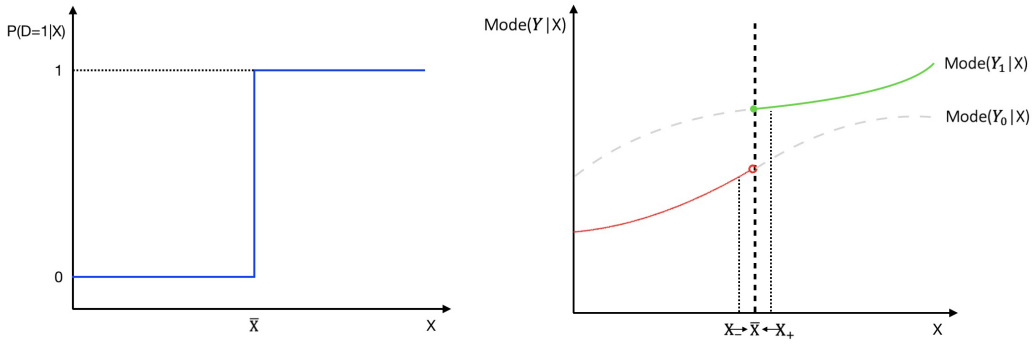

Figure 4.3: Modal Sharp Regression Discontinuity

Note: Figure 4.3 indicates that $Mode(Y_0 \mid X = \bar{X} - \varepsilon)$ can be an arbitrarily good approximation to $Mode(Y_0 \mid X = \bar{X})$ when ε is small enough. It also shows that there are no observations in which the X is precisely equal to the cutoff $\bar{X}$. As a consequence, similar to the existing treatment effects, the local approximation in the modal SRD design depending on observations farther away from the cutoff is inevitable.

The fundamental concept of identifying CMTE in the SRD design is illustrated in Figure 4.3 (modal FRD design is elucidated in Figure 4.9). The left plot represents the conditional probability of receiving the treatment, which has a jump from 0 to 1 at

the cutoff $\bar{X}$. The right plot shows that there is a sharp upward jump at the cutoff $\bar{X}$ in the modal relationship between Y and X, where the data identify the modal regressions with solid lines, the counterfactual modes are unobservable (dashed lines), and the CMTE is the vertical distance between the two modal regression curves at the cutoff $\bar{X}$. Figure 4.3 shows that the mode treatment effect for the entire population cannot be identified nonparametrically without the smoothness condition because of a violation of the positive probability assumption. It also implies that we cannot learn about the CMTE away from the cutoff without making further assumptions, which motivates us to investigate the derivative and the MMCTE in what follows.

Remark 4.2.44 (**Identification without Continuity**) *While the continuity assumption is crucial for identifying mode treatment effect, in many empirical situations, it may fail and* $m(X_i)$ *might have a break at* $X = \bar{X}$*. In such a scenario, we can restore the CMTE by redefining* (4.6) *as* $Y_i = \dot{m}(X_i) + D_i\dot{\tau}_{RD} + \epsilon_i$*, where* $\dot{m}(X_i) = m(X_i) - (m_{Y_1}(X_+) - m_{Y_0}(X_-))D_i$*, and*

$$\dot{\tau}_{RD} = \tau_{RD} + \underbrace{m_{Y_1}(X_+) - m_{Y_0}(X_-)}_{Indirect\ Effect}$$

is the adjusted mode treatment effect that contains the standard CMTE at the cutoff as well as the indirect effect of the treatment due to the break of $m(X_i)$ *(the break magnitude is* $m_{Y_1}(X_+) - m_{Y_0}(X_-)$*), where the signs* $+$ *and* $-$ *denote quantities in the regression associated with* $X_i \geq \bar{X}$ *and* $X_i < \bar{X}$*, respectively. To show the continuity, we obtain* $Mode(\dot{m}(X_i) \mid X_+) = m_{Y_1}(X_+) - (m_{Y_1}(X_+) - m_{Y_0}(X_-)) = m_{Y_0}(X_-)$ *and* $Mode(\dot{m}(X_i) \mid X_-) = m_{Y_0}(X_-)$*. As* $Mode(\dot{m}(X_i) \mid X_+) - Mode(\dot{m}(X_i) \mid X_-) = 0$*, the continuity of* $Mode(\dot{m}(X) \mid X)$ *at* $X = \bar{X}$ *is fulfilled.*

Remark 4.2.45 (**Identification with Monotonicity**) *It is common to observe monotonous RD designs in reality. For instance, when measuring the effect of capital investment on firm's production, we expect that production will not decrease in the size of the capital investment due to increasing returns to scale. If we incorporate the monotonicity into the modal estimation procedure, we can release the full continuity imposed in Assumption 1 and utilize weaker one-sided continuity conditions to identify CMTE; see Lemma 4.2.3 in the appendix for the identification.*

Practically, we may wonder how the mode effect of the policy would change if the cutoff is marginally altered. We can then use the derivative of the CMTE to test the external validity of the estimated CMTE. In terms of modal regressions, if we impose a slightly stronger differentiable assumption, we are able to nonparametrically identify the derivative of the mode treatment effect with respect to the running variable at the cutoff.

Assumption 2 *(Continuous Differentiability) The modal regression functions* $Mode(Y_1 \mid X)$ *and* $Mode\,(Y_0 \mid X)$ *are continuously differentiable in* X *for all* Y*.*

This stronger smoothness assumption is easily satisfied by local linear modal estimators. In accordance with Assumption 1, we only need continuous differentiability at $X = \bar{X}$ to achieve identification, but this is not typically assumed for a single point of the covariate. By virtue of continuity, the mode treatment effect derivative ($\tau_{RD}^{(1)}$), like the mean treatment effect derivative discussed in Dong and Lewbel (2015), equals the difference of the right and left limits of the derivatives of $Mode(Y_1 \mid X)$ and $Mode(Y_0 \mid X)$ evaluated at the cutoff $\bar{X}$.

Lemma 4.2.2 *Under the preceding model settings and Assumption 2, the mode treatment effect derivative is defined as*

$$\tau_{RD}^{(1)}(\bar{X}) = m_{Y_1}^{(1)}(\bar{X}) - m_{Y_0}^{(1)}(\bar{X}),$$

where $m_{Y_1}^{(1)}(\bar{X}) = \lim_{X \downarrow \bar{X}} Mode^{(1)}(Y_1 \mid X)$ *and* $m_{Y_0}^{(1)}(\bar{X}) = \lim_{X \uparrow \bar{X}} Mode^{(1)}(Y_0 \mid X)$ *are the right and left limits of the derivatives of modal regression functions with respect to* X*.*

The proof of Lemma 4.2.2 is included in the appendix. The magnitude of $\tau_{RD}^{(1)}(\bar{X})$ can be used to measure the impact of a marginal change in the running variable X on the treatment effect and test the external validity or generality of the estimated CMTE. A large value of $\tau_{RD}^{(1)}(\bar{X})$ indicates that a minor change in the running variable will cause a substantial change in the mode treatment effect. In addition, the sign of $\tau_{RD}^{(1)}(\bar{X})$ is also relevant since it indicates whether the CMTE for individuals with the value X, which is somewhat larger or smaller than $\bar{X}$, is likely to be stronger or lower.

As previously stated, the mode treatment effect identified by the modal SRD design applies only to a small subpopulation having $X = \bar{X}$. In fact, we may be interested in evaluating the effects of policy intervention in a variety of contexts and situations by looking at whether individuals with other values of X around $\bar{X}$ would have predicted treatment effects of comparable sign and magnitude. One simple example of a policy intervention involves altering the eligibility cutoff. For example, a change in income tax brackets may have an impact on people's behavior accordingly. To explore such a cutoff change effect, we propose the MMCTE and demonstrate that with the local policy invariance (Abbring and Heckman, 2007; Dong and Lewbel, 2015),[5] the MMCTE is identified as $\tau_{RD}^{(1)}(\bar{X})$.

[5]The local policy invariance in the content of mode treatment effect means that the change in the CMTE is negligible compared to a change ε in the cutoff $\bar{X}$ when $\varepsilon \to 0$.

Lemma 4.2.3 *Let Λ denote a possible cutoff and $M(X,\Lambda) = Mode(Y_1 \mid X,\Lambda) - Mode(Y_0 \mid X,\Lambda)$ be a hypothetical treatment effect function ($X \neq \Lambda$ or $\Lambda \neq \bar{X}$). Under the previous model settings, suppose that Assumption 2 and the local policy invariance $\frac{\partial M(X,\Lambda)}{\partial \Lambda}\big|_{X=\bar{X},\Lambda=\bar{X}}$ =0 hold. Assuming $M(\cdot)$ is differentiable, we have*

$$MMCTE = \frac{\partial M(\Lambda,\Lambda)}{\partial \Lambda}\Big|_{\Lambda=\bar{X}} = \tau_{RD}^{(1)}(\bar{X}) + \frac{\partial M(X,\Lambda)}{\partial \Lambda}\Big|_{X=\bar{X},\Lambda=\bar{X}} = \tau_{RD}^{(1)}(\bar{X}),$$

where the MMCTE is nonparametrically identified as $\tau_{RD}^{(1)}(\bar{X})$.

Given the identified MMCTE, we can apply the Taylor expansion to obtain an approximate estimate of the new CMTE when the cutoff is changed to a new one, $\bar{X}^*$, such that $|\bar{X}^* - \bar{X}| = o(1)$ and

$$\tau_{RD}(\bar{X}^*) \approx \tau_{RD}(\bar{X}) + \tau_{RD}^{(1)}(\bar{X})(\bar{X}^* - \bar{X}), \tag{4.7}$$

where "$\approx$" denotes an approximation in probability that excludes higher-order terms. We can observe from (4.7) that if the magnitude of $\tau_{RD}^{(1)}(\bar{X})$ is small, we have $\tau_{RD}(\bar{X}^*) \approx \tau_{RD}(\bar{X})$, indicating that the the CMTE or the associated policy is stable; otherwise, the CMTE is unstable. We would like to point out that (4.7) may also be utilized to extrapolate the CMTE far away from the cutoff when the local policy invariance is not satisfied, but $\frac{\partial M(X,\Lambda)}{\partial \Lambda}\big|_{X=\bar{X},\Lambda=\bar{X}}$ is negligibly small.

Remark 4.2.46 (**Various CMTEs**) *Similar to the quantile SRD design in Qu and Yoon (2019), the technique developed in this paper can be utilized to investigate various other mode treatment effects in addition to estimating CMTE. For example, we can compare the CMTE between subgroups (i.e., males and females) specified by a covariate Z such that*

$\tau^z_{RD} = \{\lim_{X\downarrow\bar{X}} Mode(Y \mid X, Z = Z_1) - \lim_{X\uparrow\bar{X}} Mode(Y \mid X, Z = Z_1)\} - \{\lim_{X\downarrow\bar{X}} Mode(Y \mid X, Z = Z_2) - \lim_{X\uparrow\bar{X}} Mode(Y \mid X, Z = Z_2)\}$. *If we substitute Z with a time variable, the above equation can be used to examine how the CMTE changes between two periods, t_1 and t_2, such that* $\tau^t_{RD} = \{\lim_{X\downarrow\bar{X}} Mode(Y \mid X, t = t_1) - \lim_{X\uparrow\bar{X}} Mode(Y \mid X, t = t_1)\} - \{\lim_{X\downarrow\bar{X}} Mode(Y \mid X, t = t_2) - \lim_{X\uparrow\bar{X}} Mode(Y \mid X, t = t_2)\}$. *This can be useful when there exists a confounding policy at the cutoff or when we need to ensure that the findings are as resilient as possible.*

4.2.2 Local Modal Boundary Estimation

In light of the identification results presented above, the estimation for CMTE is concerned with estimating the jump size of a discontinuity in the conditional modes, i.e., $\lim_{X\downarrow\bar{X}} Mode(Y_1 \mid X) - \lim_{X\uparrow\bar{X}} Mode(Y_0 \mid X)$. In order to obtain an unbiased estimate of the preceding difference, we need unbiased estimates of each limit. Rather than applying nonparametric kernel density estimation, we develop a local linear modal regression for estimating the mode treatment effect τ_{RD} due to the absence of the edge effects. While polynomial regression with a higher order (i.e., polynomial model with the entire sample) can theoretically capture more features of the unknown modal regression functions, it may exhibit erratic behavior when estimating boundary points, a well-accepted reality known as Runge's phenomenon (Calonico et al., 2015). Furthermore, Gelman and Imbens (2019) have shown that inference based on high-order polynomials is often inaccurate. The local linear estimation technique, on the other hand, provides robustness by guaranteeing that observations distant from the cutoff have no impact on the estimate. To justify the proposed CMTE estimator, we divide (4.6) into two equations

$$\begin{cases} Y_{1,i} = m_{Y_1}(X_{+,i}) + \epsilon_{+,i} & \text{if } X_i \geq \bar{X}, \\ \\ Y_{0,i} = m_{Y_0}(X_{-,i}) + \epsilon_{-,i} & \text{if } X_i < \bar{X}, \end{cases} \tag{4.8}$$

where $Mode(\epsilon_{+,i} \mid X_{+,i}) = 0$ and $Mode(\epsilon_{-,i} \mid X_{-,i}) = 0$. Then, the estimation of τ_{RD} requires consistently estimating two functions, $m_{Y_1}(\bar{X})$ and $m_{Y_0}(\bar{X})$, with data near the cutoff.

Under the assumption that the modal regression functions $m_{Y_1}(\cdot)$ and $m_{Y_0}(\cdot)$ have at least second derivatives in the region near the cutoff, we develop separate local linear modal regressions on each side of $\bar{X}$ to estimate (4.8), where

$$\begin{cases} m_{Y_1}(X_{+,i}) \approx m_{Y_1}(x) + m_{Y_1}^{(1)}(x)(X_{+,i} - x), \\ \\ m_{Y_0}(X_{-,i}) \approx m_{Y_0}(x) + m_{Y_0}^{(1)}(x)(X_{-,i} - x), \end{cases} \tag{4.9}$$

$|X_{+,i} - x| = o(1)$, and $|X_{-,i} - x| = o(1)$. Hence, the proposed CMTE estimator remains local to the cutoff. We thereupon maximize the following two local kernel-based objective functions[6]

$$\frac{1}{n_+ h_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi\left(\frac{Y_{1,i} - a_+ - b_+(X_{+,i} - x)}{h_{1,+}}\right) K\left(\frac{X_{+,i} - x}{h_{2,+}}\right), \tag{4.10}$$

$$\frac{1}{n_- h_{1,-} h_{2,-}} \sum_{i=1}^{n_-} \phi\left(\frac{Y_{0,i} - a_- - b_-(X_{-,i} - x)}{h_{1,-}}\right) K\left(\frac{X_{-,i} - x}{h_{2,-}}\right), \tag{4.11}$$

with respect to a_+, b_+, a_-, and b_-, where $\phi(\cdot)$ and $K(\cdot)$ are two nonnegatively symmetric kernels, $a_+ = m_{Y_1}(x)$, $b_+ = m_{Y_1}^{(1)}(x)$, $a_- = m_{Y_0}(x)$, $b_- = m_{Y_0}^{(1)}(x)$, and $h_+ = h_+(n_+) \to 0$ as $n_+ \to \infty$ and $h_- = h_-(n_-) \to 0$ as $n_- \to \infty$ are two positive bandwidth sequences that decay at an appropriate rate depending on the degree of smoothness assumed for the unknown

[6]They can also be written as

$$\frac{1}{n h_{1,+} h_{2,+}} \sum_{i=1}^{n} \mathbf{1}\left(X_i \geq \bar{X}\right) \phi\left(\frac{Y_i - a_+ - b_+(X_i - x)}{h_{1,+}}\right) K\left(\frac{X_i - x}{h_{2,+}}\right),$$

$$\frac{1}{n h_{1,-} h_{2,-}} \sum_{i=1}^{n} \mathbf{1}\left(X_i < \bar{X}\right) \phi\left(\frac{Y_i - a_- - b_-(X_i - x)}{h_{1,-}}\right) K\left(\frac{X_i - x}{h_{2,-}}\right).$$

functions. Specially, the bandwidth h_1 associated with $\phi(\cdot)$ is used to capture mode value, whereas the bandwidth h_2 related to $K(\cdot)$ regulates kernel width to localize the regression fit near the cutoff. Throughout the paper, we allow for different bandwidths on both sides of the cutoff and drop n_+ or n_- for bandwidths whenever possible to simplify notation. As stated in Yao and Li (2014), the choice of kernel functions is not very crucial for the statistical performance of the modal estimators compared to the bandwidth selection. For ease of computation, we choose a standard normal kernel for $\phi(\cdot)$ to form estimators in this paper.[7] The detailed conditions on kernel functions can be found in the following subsection.

Since our point of interest in (4.9) is $x = \bar{X}$, we can express the estimator for τ_{RD} as the difference in intercepts of the above modal estimators

$$\hat{\tau}_{RD}(\bar{X}) = \hat{m}_{Y_1}(\bar{X}) - \hat{m}_{Y_0}(\bar{X}), \tag{4.12}$$

where $\hat{m}_{Y_1}(\bar{X})$ and $\hat{m}_{Y_0}(\bar{X})$ are the estimators of a_+ and a_- that evaluated at the cutoff $x = \bar{X}$ from (4.10) and (4.11), respectively. Similarly, the estimator for the mode treatment effect derivative, evaluated at the cutoff $x = \bar{X}$, is

$$\hat{\tau}^{(1)}_{RD}(\bar{X}) = \hat{m}^{(1)}_{Y_1}(\bar{X}) - \hat{m}^{(1)}_{Y_0}(\bar{X}), \tag{4.13}$$

in which the estimators of b_+ and b_-, denoted by $\hat{m}^{(1)}_{Y_1}(\bar{X})$ and $\hat{m}^{(1)}_{Y_0}(\bar{X})$, are from (4.10) and (4.11), accordingly.

Remark 4.2.47 (**Mechanism of Modal Estimation**) *In the preceding local kernel-based objective functions, $\phi(\cdot)$ is served to find the mode, whereas $K(\cdot)$ is merely a rescale function that can put more weight on observations that are closer to the cutoff relative to*

[7] We choose a normal kernel for $\phi(\cdot)$ to form a closed-form expression for the proposed estimators in Algorithm . Dimitriadis et al. (2020) also argued that the log-concave kernels with infinite support, i.e., normal kernel, can identify the generalized modal midpoint with the unimodal assumption.

the bandwidth. To gain an insight into the proposed modal estimation, we assume that $\phi(t) = 2^{-1}1(|t| \leq h)$ *and* $K(t) = 2^{-1}1(|t| \leq h)$ *are two uniform kernels. Then,* (4.10) *tries to find the value of* a_+ *such that the band* $a_+ \pm h$ *contains the largest number of response* Y_1 *given* X_+ *within* $x \pm h$ *(Yao and Li, 2014). A similar argument is applied to* (4.11)*. We can also interpret the above objective functions on the basis of the mode loss function* $1 - 2\phi([Y - X\beta]/h)$*, where* $Mode(Y \mid X) = X\beta$ *(Silverman, 1986). As we are considering nonparametric estimation, we need to utilize another kernel* $K(\cdot)$ *to control the smoothness of the modal function.*

Remark 4.2.48 *It is widely known that the local linear estimation method in mean or quantile regression has benefits in boundary behavior and in estimating regression derivatives (Fan and Gijbels, 1996). Such a property of automatic adjustment near boundary regions is also shown in the modal regression; see the asymptotic results in the following part. In practice, we can also apply the local constant modal estimation for* (4.8) *borrowing the idea of kernel density estimation in Chen et al. (2016), where we maximize the following two local kernel-based objective functions*

$$\frac{1}{n_+ h_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi\left(\frac{Y_{1,i} - a_+}{h_{1,+}}\right) K\left(\frac{X_{+,i} - x}{h_{2,+}}\right),$$

$$\frac{1}{n_- h_{1,-} h_{2,-}} \sum_{i=1}^{n_-} \phi\left(\frac{Y_{0,i} - a_-}{h_{1,-}}\right) K\left(\frac{X_{-,i} - x}{h_{2,-}}\right),$$

in which modal regression functions are approximated by a constant at a fixed point. Then, the estimation algorithm and asymptotic theorems shown below in this paper may remain valid but with some modifications. However, such local constant estimators are generally

unappealing for RD designs due to the poor performance at boundary points. It will also occur the "curse of dimensionality" when there exist additional covariates and cannot give the estimator of the mode treatment effect derivative unless we take the finite difference approximation. We highlight that in addition to the local approximation method, other basis systems, including B-splines, Fourier bases, and polynomial bases, can also be adopted for estimation. Regardless of which expansion we use, the CMTE and the derivative will be uniquely identified.

Remark 4.2.49 (**Semiparametric Modal Estimation**) *We in this paper utilize two separate modal regression functions to capture local modes on both sides of the cutoff to estimate the desired CMTE. As an alternative, we can also interpret the estimation problem of the effect parameter in the modal SRD as that of the slope coefficient in a partial linear model*

$$\frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{Y_i-a-b_1(X_{+,i}-x)-b_2(X_{-,i}-x)-D_i\pi_{RD}}{h_1}\right)K\left(\frac{X_i-x}{h_2}\right),$$

where $a=m(x)$, $b_1=m^{(1)}_{Y_1}(x)$, $b_2=m^{(1)}_{Y_0}(x)$, *and* $h_1>0$ *and* $h_2>0$ *are two bandwidths depending on sample size* n. *This semiparametric model enables the use of all available data to estimate* τ_{RD} *and achieve a parametric modal convergence rate. To be more specific, we need to update estimates in different steps to achieve the optimal convergence rates for the nonparametric and parametric components, respectively. Nevertheless, the computational burden will be increased and the choice of bandwidths will be much more complicated due to the presence of nonparametric and parametric parts.*

It is clear that, as opposed to the traditional local linear mean regression, there is no closed-form expression for the maximizers of (4.10) and (4.11). We therefore apply the

modified MEM Algorithm to tackle such a challenging problem (Li et al., 2007; Yao, 2013), which has the ascending property to guarantee the almost sure convergence of the MEM algorithm to a stationary point; see Remark 4.2.50. For simplicity, we only present the algorithm for maximizing (4.10). The maximizers of (4.11) can be obtained accordingly. In analogy to an EM algorithm, the proposed algorithm consists primarily of two steps: E-Step (calculating weight) and M-Step (updating estimates), in which we maximize the l-

Algorithm 4 MEM Algorithm for Modal SRD Design

E-Step. Calculate the weight $\pi\left(i \mid a_+^{(g)}, b_+^{(g)}\right), i = 1, \cdots, n_+$, with the preliminary estimates of the modal parameters as

$$\pi\left(i \mid a_+^{(g)}, b_+^{(g)}\right) = \frac{\phi\left(\frac{Y_{1,i} - a_+^{(g)} - b_+^{(g)}(X_{+,i}-x)}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)}{\sum_{i=1}^{n_+} \phi\left(\frac{Y_{1,i} - a_+^{(g)} - b_+^{(g)}(X_{+,i}-x)}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)}.$$

M-Step. Update $\left(a_+^{(g+1)}, b_+^{(g+1)}\right)$ with the weight calculated in the E-Step

$$\begin{aligned}\left(a_+^{(g+1)}, b_+^{(g+1)}\right) &= \arg\max_{a_+, b_+} \sum_{i=1}^{n^+} \left\{\pi\left(i \mid a_+^{(g)}, b_+^{(g)}\right) \log \frac{1}{h_{1,+}} \phi\left(\frac{Y_{1,i} - a_+ - b_+(X_{+,i}-x)}{h_{1,+}}\right)\right\} \\ &= (X_+^{*T} W_+ X_+^*)^{-1} X_+^{*T} W_+ Y_+,\end{aligned}$$

where g denotes the iteration indicator, $X_+^* = (X_{+,1}^*, \cdots, X_{+,n_+}^*)^T$ with $X_{+,i}^* = (1, X_{+,i}-x)$, $Y_+ = (Y_{1,1}, \cdots, Y_{1,n_+})^T$, and W_+ is an $n_+ \times n_+$ diagonal matrix with diagonal elements $\{\pi(i \mid a_+^{(g)}, b_+^{(g)})\}_{i=1}^{n_+}$.

Iterate. Given the initial values, iterate E-Step and M-Step repeatedly until a stopping criteria is satisfied. For instance, the Euclidean norm $\|(a_+^{(g+1)}, b_+^{(g+1)}) - (a_+^{(g)}, b_+^{(g)})\| < 10^{-5}$ or a pre-specified maximum number of iterations is achieved.

og objective function instead of the original objective function for ease of calculation. With the use of a normal kernel function, we can form a closed expression in the M-step. It is important to notice that, as there may exist multiple maxima for the objective functions with small bandwidths, we should try different starting points, such as local linear mean or quantile estimates, in practice to guarantee that the MEM algorithm converges to the global maximizer (Ullah et al., 2021, 2022).

Remark 4.2.50 *We can prove that each iteration of the MEM algorithm monotonically nondecreases the objective functions. Define* $f(Y_{1,i}, X_{+,i}; a_+, b_+) = \phi\left(\frac{Y_{1,i}-a_+-b_+(X_{+,i}-x)}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)$*, by Jensen's inequality we can rewrite maximizing* (4.10) *in the M-Step as*

$$\begin{aligned}\log\sum_{i=1}^{n_+} f(Y_{1,i}, X_{+,i}; a_+^{(g+1)}, b_+^{(g+1)}) &= \log\sum_{i=1}^{n_+} \frac{f(Y_{1,i}, X_{+,i}; a_+^{(g+1)}, b_+^{(g+1)})}{\pi(i \mid a_+^{(g)}, b_+^{(g)})}\pi(i \mid a_+^{(g)}, b_+^{(g)}) \\ &\geq \sum_{i=1}^{n_+} \pi(i \mid a_+^{(g)}, b_+^{(g)}) \log\frac{f(Y_{1,i}, X_{+,i}; a_+^{(g+1)}, b_+^{(g+1)})}{\pi(i \mid a_+^{(g)}, b_+^{(g)})} \\ &\geq \log\sum_{i=1}^{n_+} f(Y_{1,i}, X_{+,i}; a_+^{(g)}, b_+^{(g)}),\end{aligned}$$

which ensures that the sequence $(a_+^{(g)}, b_+^{(g)})$ *is convergent to some* $(\hat{a}_+, \hat{b}_+)$*.*

4.2.3 Asymptotic Properties near the Boundary

We are now in a position to provide the limiting distributions for the proposed modal estimators near the boundary of the region of interest, in the sense that τ_{RD} is estimated nonparametrically using modal regression. Before presenting asymptotic theorems, it is convenient to introduce some notations that will be used throughout the remainder of this paper. Since the boundary point is $\bar{X}$ in the modal RD designs, we define the target point $x = \bar{X} + \bar{c}h_{2,+}$ or $x = \bar{X} - \bar{c}h_{2,-}$ with $h_{2,+} \to 0$ and $h_{2,-} \to 0$ in this subsection,

where $\bar{c}$ is a positive constant of the support of kernel $K(\cdot)$, and allow $g_{\epsilon_+}(\epsilon_+ \mid x)$ and $g_{\epsilon_-}(\epsilon_- \mid x)$ to be the conditional density functions of ϵ_+ and ϵ_- given x, correspondingly. We use $m_{Y_j}^{(c)}(x) = \partial^{(c)} m_{Y_j}(x)/\partial x^c$ to indicate the cth derivative of $m_{Y_j}(x)$ for $j = 0, 1$, and let $g_{\epsilon_+}^{(c)}(\epsilon_+ \mid x) = \partial^{(c)} g_{\epsilon_+}(\epsilon_+ \mid x)/\partial \epsilon_+^c$ and $g_{\epsilon_-}^{(c)}(\epsilon_- \mid x) = \partial^{(c)} g_{\epsilon_-}(\epsilon_- \mid x)/\partial \epsilon_-^c$ represent the cth derivatives of $g_{\epsilon_+}(\epsilon_+ \mid x)$ and $g_{\epsilon_-}(\epsilon_- \mid x)$, respectively, for $c = 1, 2, 3$. We define $T_n(x) = T(x) + O_p(s_n)$ uniformly for $x \in \mathcal{X}$ if $\sup_{x \in \mathcal{X}} |T_n(x) - T(x)| = O_p(s_n)$ and use "$\xrightarrow{d}$" to denote convergence in distribution. To facilitate the investigation, we state the following assumptions from which the limiting distributions of the proposed estimators are derived.

C1 (Data Structure) $\{Y_i, X_i\}_{i=1}^n$ is an *i.i.d.* random sequence drawn from a joint probability distribution $F_{X,Y}(X, Y)$ on $R \times R$.

C2 (Kernel Function) The kernel functions $\phi(\cdot)$ and $K(\cdot)$: $R \to R$ are nonnegatively symmetric continuous density functions with bounded support and integrated to one, where the bounded support of $K(\cdot)$ is denoted as $[-M, M]$.

C3 (Smoothness) The modal regression functions $m_{Y_1}(X_+)$ and $m_{Y_0}(X_-)$, as well as their first and second derivatives, have right and left limits at the boundary point $\bar{X}$, respectively.

C4 (Conditional Density) (i) For a fixed point x, $f_{X_+}(x) > 0$ and $f_{X_-}(x) > 0$, where $f_{X_+}(x)$ and $f_{X_-}(x)$ are the marginal density functions that right and left continuous differentiable at $x = \bar{X}$, separately; (ii) Given a certain point x, $g_{\epsilon_+}(\epsilon_+ \mid x) > 0$ and $g_{\epsilon_-}(\epsilon_- \mid x) > 0$. Furthermore, $g_{\epsilon_+}(\epsilon_+ \mid x)$ and $g_{\epsilon_-}(\epsilon_- \mid x)$ have the fourth right and left continuous derivatives at $x = \bar{X}$, severally, and $g_{\epsilon_+}(\epsilon_+ \mid x) < g_{\epsilon_+}(0 \mid x)$ and

$g_{\epsilon_-}(\epsilon_- \mid x) < g_{\epsilon_-}(0 \mid x)$ for all $\epsilon_+ \neq 0$ and $\epsilon_- \neq 0$; (iii) all density functions are bounded away from infinity.

Most of the above assumptions are compatible with the conditions imposed in the majority of the existing modal regression literature. C1 is pretty typical in describing the sample generating process for cross sectional data in most applications of RD designs. The *i.i.d.* condition is shared by many prior analyses on regression discontinuity and kink designs. We can extend to the dependent data (strictly stationary process) under mixing conditions but with more tedious arguments and proofs. C2 is a mild condition on the kernel functions. The compact support condition for $\phi(\cdot)$ and $K(\cdot)$ is not essential and can be relaxed as long as certain integrability restrictions are imposed on the tail of the kernel functions. Especially, a normal kernel is permitted, which is the default kernel for $\phi(\cdot)$ used in this paper. C3 is a frequently employed condition on the smoothness of the nonparametric functions in local linear fitting to control the leading bias of the RD estimator. It allows the existence of the difference between the right and left derivatives of the conditional mode evaluated at the cutoff $\bar{X}$. Notice that the existing second derivative ensures that the bias of the local linear estimator is of order $O_p(h_{1,+}^2 + h_{2,+}^2)$ or $O_p(h_{1,-}^2 + h_{2,-}^2)$ even close to the boundary. As a result, the local linear modal estimation exhibits automatically excellent behavior near the boundary, eliminating the need for boundary correction. C4 imposes a certain smoothness on distributions. It excludes discontinuous changes in the density of the running variable X and overcomes the technical issues associated with a near-zero denominator. The presence of a positive density in the neighborhood of $\bar{X}$ guarantees that there exist sufficient data for estimating the treatment effect. Notably, the global mode

for $\{\epsilon_+\}_{i=1}^{n_+}$ or $\{\epsilon_-\}_{i=1}^{n_-}$ is enforced to simplify the illustration and computation (Kemp and Santos Silva, 2012; Ullah et al., 2021, 2022). The method can be easily modified to accommodate the multi-mode case. In contrast to the classical mean regression, the modal regression does not require the existence of error term moments (i.e., Cauchy distribution has no moments). Even when the conditional variance of the error terms is infinite, the proposed modal estimators still enjoy asymptotic normality. In practice, if we intend to compare the CMTE to the mean treatment effect, we must impose a condition that there is a constant $s > 2$ such that $E(|Y|^{2s}) < \infty$ and $E(|X|^{2s}) < \infty$. The bandwidths are the key parameters to consider when implementing the CMTE estimator. All bandwidth-related conditions are specified in each of the following theorems, and the choice of bandwidths is addressed in depth below.

Remark 4.2.51 *If X is instead assumed to have a discrete distribution with a limited number of points of support, it may not be convenient to anticipate local linear estimation to be particularly effective in estimating treatment effects. More specifically, the treatment effect parameter is usually not point identified because there is no data within the specified bandwidth range. One then may be compelled to arbitrarily choose a broad and ad hoc bandwidth, resulting in an inappropriately centered confidence interval with a nonignored asymptotic bias. The effective way to deal with such an issue is to utilize the discrete kernel function for RD designs.*

Although the asymptotic theory for both $\hat{m}_{Y_1}(\cdot)$ and $\hat{m}_{Y_0}(\cdot)$ near the boundary can be driven given the aforementioned assumptions, we only present the results for modal estimators $\hat{m}_{Y_1}(\cdot)$ and $\hat{m}^{(1)}_{Y_1}(\cdot)$ associated with $x = \bar{X} + \bar{c}h_{2,+}$ to conserve space, which

are shown as follows. The asymptotic properties for modal estimators $\hat{m}_{Y_0}(\cdot)$ and $\hat{m}^{(1)}_{Y_0}(\cdot)$ related to $x = \bar{X} - \bar{c}h_{2,-}$ are listed in the appendix.

Theorem 4.2.18 *Under the regularity conditions C1-C4, with probability approaching one, as* $n_+ \to \infty$, $h_{1,+} \to 0$, $h_{2,+} \to 0$, $h^2_{2,+}/h_{1,+} \to 0$, *and* $n_+h_{2,+}h^5_{1,+} \to \infty$, *there exist consistent maximizers* $(\hat{m}_{Y_1}(x), h_{2,+}\hat{m}^{(1)}_{Y_1}(x))$ *of* (4.10) *such that*

i. $|\hat{m}_{Y_1}(x) - m_{Y_1}(x)| = O_p\left(\left(n_+h_{2,+}h^3_{1,+}\right)^{-1/2} + h^2_{1,+} + h^2_{2,+}\right)$,

ii. $|h_{2,+}(\hat{m}^{(1)}_{Y_1}(x) - m^{(1)}_{Y_1}(x))| = O_p\left(\left(n_+h_{2,+}h^3_{1,+}\right)^{-1/2} + h^2_{1,+} + h^2_{2,+}\right)$.

Theorem 4.2.18 expresses the magnitudes of the estimation bias and variance of modal estimators near the boundary, which exhibits the same convergence rate as the interior points. The results indicate that local linear modal regression has an attractive boundary behavior in the sense that it maintains $O(h^2_{1,+}+h^2_{2,+})$ bias across the design space. In addition, it can be seen that the optimal choices of $h_{1,+}$ and $h_{2,+}$ are of order $O(n_+^{-1/8})$ by minimizing the asymptotic MSE of estimators over $h_{1,+}$ and $h_{2,+}$, and the corresponding convergence rate is $O_p(n_+^{-1/4})$, which is slower than that of mean or quantile regression due to the characteristic of mode (Parzen, 1962). Such MSE-optimal bandwidths are too large for inference as they lead to non-negligible biases of order $O_p(h^2_{1,+})$ and $O_p(h^2_{2,+})$, respectively. In practice, we can modify the MSE-optimal bandwidths through multiplying them by $n_+^{-\gamma}$, where $\gamma > 0$ is a small integer.

Theorem 4.2.19 *With* $n_+h^5_{2,+}h^3_{1,+} = O(1)$ *and* $n_+h_{2,+}h^7_{1,+} = O(1)$, *under the same conditions as Theorem 4.2.18, the parameters satisfying the consistency results in Theorem 4.2.18 have the following asymptotic result*

$$\sqrt{n_+h_{2,+}h_{1,+}^3}\left(\begin{bmatrix}\hat{m}_{Y_1}(x)-m_{Y_1}(x)\\ h_{2,+}(\hat{m}_{Y_1}^{(1)}(x)-m_{Y_1}^{(1)}(x))\end{bmatrix}-\Gamma^{-1}\Big(\frac{h_{2,+}^2}{2}m_{Y_1}^{(2)}(x)\Lambda_2-\frac{h_{1,+}^2}{2}\frac{g_{\epsilon_+}^{(3)}(0\mid x)}{g_{\epsilon_+}^{(2)}(0\mid x)}\Lambda_1\Big)\right)$$
$$\xrightarrow{d}\mathcal{N}\left(0,\frac{g_{\epsilon_+}(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{\left(g_{\epsilon_+}^{(2)}(0\mid x)\right)^2 f_{X_+}(x)}\Gamma^{-1}\Sigma\Gamma^{-1}\right).$$

If we allow $n_+h_{2,+}^5h_{1,+}^3\to 0$ *and* $n_+h_{2,+}h_{1,+}^7\to 0$*, the asymptotic theorem becomes*

$$\sqrt{n_+h_{2,+}h_{1,+}^3}\begin{pmatrix}\hat{m}_{Y_1}(x)-m_{Y_1}(x)\\ h_{2,+}(\hat{m}_{Y_1}^{(1)}(x)-m_{Y_1}^{(1)}(x))\end{pmatrix}\xrightarrow{d}\mathcal{N}\left(0,\frac{g_{\epsilon_+}(0\mid x)\int\tau^2\phi^2(\tau)d\tau}{\left(g_{\epsilon_+}^{(2)}(0\mid x)\right)^2 f_{X_+}(x)}\Gamma^{-1}\Sigma\Gamma^{-1}\right),$$

where

$$\Lambda_1=\begin{bmatrix}\int_{-\bar{c}}^{M}K(w)dw\\ \int_{-\bar{c}}^{M}wK(w)dw\end{bmatrix},\ \Lambda_2=\begin{bmatrix}\int_{-\bar{c}}^{M}w^2K(w)dw\\ \int_{-\bar{c}}^{M}w^3K(w)dw\end{bmatrix},\ \Gamma=\begin{bmatrix}\int_{-\bar{c}}^{M}K(w)dw & \int_{-\bar{c}}^{M}wK(w)dw\\ \int_{-\bar{c}}^{M}wK(w)dw & \int_{-\bar{c}}^{M}w^2K(w)dw\end{bmatrix},$$

and $\Sigma=\begin{bmatrix}\int_{-\bar{c}}^{M}K^2(w)dw & \int_{-\bar{c}}^{M}wK^2(w)dw\\ \int_{-\bar{c}}^{M}wK^2(w)dw & \int_{-\bar{c}}^{M}w^2K^2(w)dw\end{bmatrix}.$

This asymptotic result for points close to the boundary is of interest in its own right because it serves as the foundation for the subsequent derivation of the limiting distribution of the CMTE process. The modal convergence rate $(n_+h_{2,+}h_{1,+}^3)^{1/2}$ can be regarded as a novel one in the RD design literature. The orders of the bias and variance are consistent with the usual asymptotic results for the interior points in nonparametric modal regression, but the exact expressions are different since we concentrate on the local linear approximation near the boundary (half the weights are not defined). The bias term linked with $h_{1,+}$ is due to the estimation of the mode value, whereas the bias term attributed to $h_{2,+}$ is caused by local linear approximation, which has the same format as that from local linear mean estimation. Also, the variance term is not dependent on any moments, indicating that no

moment conditions are required for modal estimation. The results show that the proposed modal estimators enjoy the property of automatic boundary correction, which is a nice feature of the local linear estimator. In addition, because of the use of data on the right side of the cutoff, we do not have asymptotic independence between $\hat{m}_{Y_1}(x)$ and $\hat{m}^{(1)}_{Y_1}(x)$ even with a symmetric kernel function.[8] The theorem further reveals that the asymptotic bias term is substantially influenced by bandwidths and can be effectively eliminated under specific circumstances so that the estimator is centered at the actual value. The estimator $\hat{m}_{Y_0}(\cdot)$ is subjected to the same comments.

Due to the sensitivity of CMTE to bandwidth choice, the estimation in the modal SRD design requires specifying a bandwidth around the cutoff, much as it does for the mean or quantile treatment effect. The difference is that we also need to select another bandwidth for capturing mode. We then apply the results in the above theorem to obtain the asymptotic MSE of $\hat{m}_{Y_1}(x)$. Define $\mu_{+,l} = \int_{-\bar{c}}^{M} w^l K(w)dw$ and $v_{+,l} = \int_{-\bar{c}}^{M} w^l K^2(w)dw$ for $l = 0, 1, 2, 3$. The MSE of $\hat{m}_{Y_1}(\cdot)$ near the boundary can be approximated by

$$E(\hat{m}_{Y_1}(x) - m_{Y_1}(x))^2 \approx B_1 h_{1,+}^4 + B_2 h_{2,+}^4 + 2B_3 h_{1,+}^2 h_{2,+}^2 + \frac{1}{n_+ h_{2,+} h_{1,+}^3}\mathcal{V}, \tag{4.14}$$

where
$$\begin{cases} B_1 = 4^{-1}(g^{(2)}_{\epsilon_+}(0 \mid x))^{-2}(g^{(3)}_{\epsilon_+}(0 \mid x))^2, \\ B_2 = 4^{-1}[m^{(2)}_{Y_1}(x)]^2(\mu^2_{+,2} - \mu_{+,1}\mu_{+,3})(\mu_{+,0}\mu_{+,2} - \mu^2_{+,1})^{-1}, \\ B_3 = -4^{-1}m^{(2)}_{Y_1}(x)(g^{(2)}_{\epsilon_+}(0 \mid x))^{-1}g^{(3)}_{\epsilon_+}(0 \mid x)(\mu^2_{+,2} - \mu_{+,1}\mu_{+,3})(\mu_{+,0}\mu_{+,2} - \mu^2_{+,1})^{-1}, \\ \mathcal{V} = \int \tau^2\phi^2(\tau)d\tau f^{-1}_{X_+}(x)g_{\epsilon_+}(0 \mid x)(g^{(2)}_{\epsilon_+}(0 \mid x))^{-2} \\ \qquad (\mu^2_{+,2}v_{+,0} - 2\mu_{+,1}\mu_{+,2}v_{+,1} + \mu^2_{+,1}v_{+,2})(\mu_{+,0}\mu_{+,2} - \mu^2_{+,1})^{-2}. \end{cases}$$

[8] With the defined values of $\mu_{+,l}$ and $v_{+,l}$, the covariance between $\hat{m}_{Y_1}(x)$ and $\hat{m}^{(1)}_{Y_1}(x)$ is $(n_+h_{2,+}h^3_{1,+})^{-1}$ $(\mu_0\mu_2 - \mu_1^2)^{-2}(\mu_1^2 v_1 - \mu_2 v_0 v_1 + \mu_2 v_1 \mu_0 - \mu_1 v_2 \mu_0)$. For the interior points, this value is 0 with a symmetric kernel.

By minimizing the above equation as a function of the bandwidths, we can achieve the optimal rate for bandwidths

$$\hat{h}_{2,+} = \left(\frac{3\mathcal{V}}{4n_+ B_4^5 (B_1 B_4^2 + B_3)} \right)^{1/8} \text{ and } \hat{h}_{1,+} = B_4 \hat{h}_{2,+}, \tag{4.15}$$

where $B_4 = ([B_3^2 + 3B_1B_2]^{1/2} + B_3)/B_1$. Thus, $\hat{h}_{1,+} = O(n_+^{-1/8})$ and $\hat{h}_{2,+} = O(n_+^{-1/8})$, implying that $n_+(\hat{h}_{1,+}^8) \to \alpha_c \in (0, \infty)$ and $n_+(\hat{h}_{2,+}^8) \to \alpha_c \in (0, \infty)$. Combining this with Theorem 4.2.19, it is clear that the MSE-optimal bandwidths of $h_{1,+}$ and $h_{2,+}$ with rate $n_+^{-1/8}$ do not satisfy the requirements of $\lim_{n_+ \to \infty} n_+ h_{2,+}^5 h_{1,+}^3 = 0$ and $\lim_{n_+ \to \infty} n_+ h_{2,+} h_{1,+}^7 = 0$, and will lead to a first-order bias in the distributional approximation and bring undercoverage of confidence interval in inference. Additionally, applying this MSE-optimal bandwidth selection in reality is likely to introduce further variability into the chosen bandwidths due to the estimation of so many unknown terms, potentially resulting in the use of extremely large bandwidths. We may utilize other bandwidth choice methods, such as the revised MSE-optimal bandwidths by taking bias correction into consideration. After correcting the bias, the optimal bandwidths can balance $Bias^2(\hat{m}(\cdot) - Bias(\hat{m}(\cdot)))$ with the adjusted variance term (see Remark 4.2.52). Nevertheless, given that $Bias(\hat{m}(\cdot)) = O(h_{1,+}^2 + h_{2,+}^2)$, we have to expand the bias expression up to a higher order and develop the limit process to calculate $Bias(\hat{m}(\cdot) - Bias(\hat{m}(\cdot)))$, which will impose a significant burden on computation as well.

Practically, it may be more reasonable to employ different bandwidths for the modal regressions on both sides of the cutoff to prevent the offset of the biases in those two regression functions from one another. Developing such data-driven bandwidths with the plug-in method is much more difficult due to the large number of unknown terms

in the above expressions, which necessitates the use of pilot bandwidths and additional smoothness assumptions. Several studies in the RD design literature suggest reducing the MSE-optimal bandwidths by an arbitrary number. As Hall (1993) discovered that undersmoothing outperforms naive bias-corrected confidence intervals, we in this paper limit ourselves to applying the undersmoothing technique, which is a common requirement in the nonparametric or RD design literature.

We generalize the results in Kemp and Santos Silva (2012) to choose undersmoothed bandwidths guided by MSE-optimal rates, where in empirical analysis we let $\hat{h}_{1,+} = 1.6\text{MAD}{n_+}^{-0.13}$ (-0.13 comes from the rate -1/8 and undersmoothing requirement), $MAD = med_{1,i}\{|(Y_{1,i} - \tilde{m}_{Y_1}(x)) - med_{1,i}(Y_{1,i} - \tilde{m}_{Y_1}(x))|\}$ in which $\tilde{m}_m(\cdot)$ represents the mean estimator, med means taking the median value, and $\hat{h}_{2,+} = 1.06\sigma_{X_+} n_+^{-0.13}$ in which σ_{X_+} is the standard deviation for samples $\{X_{+,i}\}_{i=1}^{n_+}$. According to the previous discussion, we know that the bandwidth $h_{1,+}$, in comparison to the bandwidth $h_{2,+}$, can have an impact on the number of estimated local modes. To minimize the effect of $h_{1,+}$ in simulation examples, we select 50 alternative values of $h_{1,+}$ ranging from $50MAD$ to $0.5MADn^{-0.13}$ and provide the best findings (in terms of MSE). The same procedures are applied to calculate the bandwidths $h_{1,-}$ and $h_{2,-}$. We note that only observations with the running variable falling inside the chosen neighborhood $[x - \hat{h}_2, x + \hat{h}_2]$ are used for estimating CMTE. Although Monte Carlo simulations show that the undersmoothed MSE-based bandwidth selection rule allows the proposed estimation procedure to perform well in finite samples, we emphasize that the selection process can only provide a benchmark estimate, and may not produce the optimal modal estimator since it considers the performance of the estimator

over the entire support. A formal investigation into practically optimal bandwidths in the modal RD designs (e.g., the bandwidths that minimize the coverage error of the confidence interval) is deferred to future work.

4.2.4 Modal Inference on the Boundary

With the nice boundary performance of modal estimators, we can now establish the main results of this paper at the boundary point $x = \bar{X}$ on the basis of the studentized statistic. To conserve space, we concentrate on the inference of CMTE and do not pay much attention to the mode treatment effect derivative. However, using the previous asymptotic results, it is straightforward to show that $(\hat{\tau}^{(1)}_{RD} - \tau^{(1)}_{RD} - Bias(\hat{\tau}^{(1)}_{RD})) \xrightarrow{d} \mathcal{N}(0, Var(\hat{\tau}^{(1)}_{RD}))$, where $Bias(\hat{\tau}^{(1)}_{RD}) = Bias(\hat{m}^{(1)}_{Y_1}(\bar{x})) - Bias(\hat{m}^{(1)}_{Y_0}(\bar{x}))$ and $Var(\hat{\tau}^{(1)}_{RD}) = Var(\hat{m}^{(1)}_{Y_1}(\bar{x})) + Var(\hat{m}^{(1)}_{Y_0}(\bar{x}))$ because the data used in the estimation are independent. The exact expressions for $Bias(\hat{\tau}^{(1)}_{RD})$ and $Var(\hat{\tau}^{(1)}_{RD})$ are given in the appendix.

Theorem 4.2.20 *Define* $\mu_{-,l} = \int_{-M}^{\bar{c}} w^l K(w)dw$ *and* $v_{-,l} = \int_{-M}^{\bar{c}} w^l K^2(w)dw$ *for* $l = 0, 1, 2, 3$. *Under the regularity conditions C1-C4, with* $n_+ h^5_{2,+} h^3_{1,+} = O(1)$, $n_+ h_{2,+} h^7_{1,+} = O(1)$, $n_- h^5_{2,-} h^3_{1,-} = O(1)$, *and* $n_- h_{2,-} h^7_{1,-} = O(1)$, *as both* $n_+ \to \infty$ *and* $n_- \to \infty$, *we have*

$$\frac{\hat{\tau}_{RD} - \tau_{RD} - Bias(\hat{\tau}_{RD})}{\sqrt{Var(\hat{\tau}_{RD})}} \xrightarrow{d} \mathcal{N}(0,1).$$

If we allow $n_+ h^5_{2,+} h^3_{1,+} \to 0$, $n_+ h_{2,+} h^7_{1,+} \to 0$, $n_- h^5_{2,-} h^3_{1,-} \to 0$, *and* $n_- h_{2,-} h^7_{1,-} \to 0$, *as both* $n_+ \to \infty$ *and* $n_- \to \infty$, *we have*

$$(Var(\hat{\tau}_{RD}))^{-1/2}(\hat{\tau}_{RD} - \tau_{RD}) \xrightarrow{d} \mathcal{N}(0,1),$$

where $Bias(\hat{\tau}_{RD}) = Bias(\hat{m}_{Y_1}(\bar{X})) - Bias(\hat{m}_{Y_0}(\bar{X})) = \left\{ \frac{h^2_{2,+}}{2} m^{(2)}_{Y_1}(\bar{X}) p_+(\bar{c}) - \frac{h^2_{1,+}}{2} \frac{g^{(3)}_{\epsilon_+}(0 \mid \bar{X})}{g^{(2)}_{\epsilon_+}(0 \mid \bar{X})} \right.$

$$- \frac{h_{2,-}^2}{2} m_{Y_0}^{(2)}(\bar{X}) p_-(\bar{c}) + \frac{h_{1,-}^2}{2} \frac{g_{\epsilon_-}^{(3)}(0 \mid \bar{X})}{g_{\epsilon_-}^{(2)}(0 \mid \bar{X})} \Bigg\} (1 + o_p(1)),$$

$$Var(\hat{\tau}_{RD}) = Var(\hat{m}_{Y_1}(\bar{X})) + Var(\hat{m}_{Y_0}(\bar{X})) = \Bigg\{ \frac{\int \tau^2 \phi^2(\tau) d\tau f_{X_+}^{-1}(\bar{X})}{n_+ h_{2,+} h_{1,+}^3} \frac{g_{\epsilon_+}(0 \mid \bar{X})}{(g_{\epsilon_+}^{(2)}(0 \mid \bar{X}))^2} \xi_+(\bar{c})$$

$$+ \frac{\int \tau^2 \phi^2(\tau) d\tau f_{X_-}^{-1}(\bar{X})}{n_- h_{2,-} h_{1,-}^3} \frac{g_{\epsilon_-}(0 \mid \bar{X})}{(g_{\epsilon_-}^{(2)}(0 \mid \bar{X}))^2} \xi_-(\bar{c}) \Bigg\} (1 + o_p(1)),\ p_+(\bar{c}) = \frac{\mu_{+,2}^2 - \mu_{+,1}\mu_{+,3}}{\mu_{+,0}\mu_{+,2} - \mu_{+,1}^2},$$

$$p_-(\bar{c}) = \frac{\mu_{-,2}^2 - \mu_{-,1}\mu_{-,3}}{\mu_{-,0}\mu_{-,2} - \mu_{-,1}^2},\ \xi_+(\bar{c}) = \frac{\mu_{+,2}^2 v_{+,0} - 2\mu_{+,1}\mu_{+,2} v_{+,1} + \mu_{+,1}^2 v_{+,2}}{(\mu_{+,0}\mu_{+,2} - \mu_{+,1}^2)^2},$$

$$and\ \xi_-(\bar{c}) = \frac{\mu_{-,2}^2 v_{-,0} - 2\mu_{-,1}\mu_{-,2} v_{-,1} + \mu_{-,1}^2 v_{-,2}}{(\mu_{-,0}\mu_{-,2} - \mu_{-,1}^2)^2}.$$

Provided that a symmetric kernel is utilized, the values of $\mu_{+,l}$ and $v_{+,l}$ stay unchanged if kernel moments on the opposite side of the cutoff are used, implying that $p_+(\bar{c}) = p_-(\bar{c})$ and $\xi_+(\bar{c}) = \xi_-(\bar{c})$. When the bandwidths above and below the cutoff are assumed to be the same, we can deduce the following corollary.

Corollary 4.2.4 *Under the regularity conditions C1-C4, as both $n_+ \to \infty$ and $n_- \to \infty$, with the restrictions that $h_{1,+} = h_{1,-} = h_1$, $h_{2,+} = h_{2,-} = h_2$, $n_+ h_2^5 h_1^3 = O(1)$, $n_+ h_2 h_1^7 = O(1)$, $n_- h_2^5 h_1^3 = O(1)$, and $n_- h_2 h_1^7 = O(1)$, we have*

$$Bias(\hat{\tau}_{RD}) = \Bigg\{ \frac{h_2^2}{2} (m_{Y_1}^{(2)}(\bar{X}) p_+(\bar{c}) - m_{Y_0}^{(2)}(\bar{X}) p_-(\bar{c})) - \frac{h_1^2}{2} \left(\frac{g_{\epsilon_+}^{(3)}(0 \mid \bar{X})}{g_{\epsilon_+}^{(2)}(0 \mid \bar{X})} - \frac{g_{\epsilon_-}^{(3)}(0 \mid \bar{X})}{g_{\epsilon_-}^{(2)}(0 \mid \bar{X})} \right) \Bigg\}$$

$$\{1 + o_p(1)\}, and$$

$$Var(\hat{\tau}_{RD}) = \frac{\int \tau^2 \phi^2(\tau) d\tau}{h_2 h_1^3} \left(\frac{(g_{\epsilon_+}^{(2)}(0 \mid \bar{X}))^{-2} g_{\epsilon_+}(0 \mid \bar{X})}{n_+ f_{X_+}(\bar{X})} \xi_+(\bar{c}) + \frac{(g_{\epsilon_-}^{(2)}(0 \mid \bar{X}))^{-2} g_{\epsilon_-}(0 \mid \bar{X})}{n_- f_{X_-}(\bar{X})} \xi_-(\bar{c}) \right)$$

$$\{1 + o_p(1)\}.$$

Theorem 4.2.20 can be directly proved by following the results of Theorem 4.2.19. For simple presentation, we assume that the same bandwidths are used for estimation

and inference. The asymptotic distribution of $\hat{\tau}_{RD}$ is not centered at zero because of the presence of the asymptotic bias without undersmoothing. Since the data used in estimating $m_{Y_1}(\bar{X})$ and $m_{Y_0}(\bar{X})$ are not related, the processes associated with $m_{Y_1}(\bar{X})$ and $m_{Y_0}(\bar{X})$ are asymptotically uncorrelated. Furthermore, because the zero-mean normal processes are fully characterized by their covariance function, the asymptotic variance of $\hat{\tau}_{RD}$ consists of the sum of asymptotic variances of $\hat{m}_{Y_1}(\bar{X})$ and $\hat{m}_{Y_0}(\bar{X})$. If we allow the bandwidths to go to zero fast enough with the sample size (undersmoothing), the bias can be asymptotically negligible. Under the premise that the bandwidths on both sides of the cutoff are the same, if the left and right limits of the second derivatives of the unknown functions and the second and third derivatives of the unknown densities are identical, the bias would converge to zero faster, allowing for the estimation of τ_{RD} at a faster rate of convergence. Using the increased convergence rate in such a scenario, however, is problematic, as it would be difficult to establish sufficiently fast convergence in practice so that the above mentioned components are really equal.

We can use Theorem 4.2.20 for the modal inference on the boundary. For example, we can construct a conventional $100(1-\alpha)\%$ confidence interval for τ_{RD} by following the large-sample approximation of the standardized t-statistic, which is justified by

$$CI^u_{RD} = \left[\hat{\tau}_{RD} - \hat{Bias}(\hat{\tau}_{RD}) \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\tau}_{RD})}\right] \text{ or } CI^w_{RD} = \left[\hat{\tau}_{RD} \pm \Phi^{-1}_{1-\alpha/2}\sqrt{\hat{Var}(\hat{\tau}_{RD})}\right], \tag{4.16}$$

where $\hat{Bias}(\cdot)$ and $\hat{Var}(\cdot)$ are the corresponding estimates, CI^w_{RD} and CI^u_{RD} denotes the confidence intervals with and without undersmoothing, and Φ^{-1}_{α} is the appropriate α-quantile of the standard normal distribution. The estimators for the second derivatives of modal

regressions $m_{Y_1}(\cdot)$ and $m_{Y_0}(\cdot)$ can be computed using local boundary estimation with a second-order polynomial. The consistent estimates of the unknown terms in asymptotic bias and variance related to density at the cutoff can be obtained via kernel density estimation (Ullah et al., 2021). Especially, we can use the local linear mean regression to get the estimate of $\epsilon_{+,i}$, denoted by $\hat{\epsilon}^m_{+,i}$, and apply the nonparametric kernel density estimation method to obtain the mode of $\hat{\epsilon}^m_{+,i}$, say $\hat{\epsilon}_{+,m}$. We can then approximate $g^{(c)}_{\epsilon_+}(\cdot)$ by

$$g^{(c)}_{\epsilon_+}(0 \mid \bar{X}) \approx \frac{1}{n_+ \lambda_+^{(c+1)}} \sum_{i=1}^{n_+} K^{(c)} \left(\frac{\hat{\epsilon}^m_{+,i} - \hat{\epsilon}_{+,m}}{\lambda_+} \right), \tag{4.17}$$

where λ_+ is a new bandwidth and $K^{(c)}(\cdot)$ represents the cth derivative of kernel function. Similarly, we can obtain the estimate for $g^{(c)}_{\epsilon_-}(\cdot)$. In addition, although using the same bandwidths for estimation and inference can greatly reduce the complexity of implementation, a tuning parameter that is useful for estimation purpose may not necessarily be optimal for conducting inference (Pagan and Ullah, 1999). This suggests that the confidence region CI^u_{RD} built using the results obtained from the MSE-optimal bandwidths in the estimation may not provide good coverage accuracy for the probability limit of $\hat{\tau}_{RD}$. In that case, we need to minimize the asymptotic MSE of $\hat{\tau}_{RD}$ in Theorem 4.2.20 to obtain the optimal bandwidth expressions and utilize the plug-in method to consistently estimate them.

Remark 4.2.52 (Robust Bias-Corrected Estimator) *Calonico et al. (2014) developed a robust bias-corrected inference method by taking the additional variability introduced by the bias term into account for mean treatment effect. Such a method is resistant to large bandwidths that obey $h_1 \propto n^{-1/8}$ and $h_2 \propto n^{-1/8}$ in the mode content. As we can rewrite the bias-correct modal estimator as $\hat{\tau}^{bc}_{RD} = \hat{m}_{Y_1} - \hat{Bias}(\hat{m}_{Y_1}) - (\hat{m}_{Y_0} - \hat{Bias}(\hat{m}_{Y_0}))$, which is the difference between two normal random variables, we have the adjusted t-statistic*

$$(\,Var(\hat{\tau}^{bc}_{RD}))^{-1/2}(\hat{\tau}^{bc}_{RD}-\tau_{RD})\xrightarrow{d}\mathcal{N}\,(0,1)\,,$$

where $Var(\hat{\tau}^{bc}_{RD}) = Var(\hat{\tau}_{RD} - \hat{Bias}(\hat{\tau}_{RD})) = Var(\hat{\tau}_{RD}) + Var(\hat{Bias}(\hat{\tau}_{RD})) - 2Cov(\hat{\tau}_{RD}, \hat{Bias}(\hat{\tau}_{RD})) = Var(\hat{m}_{Y_1}(\bar{X})) + Var(\hat{m}_{Y_0}(\bar{X})) + Var(\hat{Bias}(\hat{m}_{Y_1}(\bar{X}))) + Var(\hat{Bias}(\hat{m}_{Y_0}(\bar{X}))) - 2Cov(\hat{m}_{Y_1}(\bar{X}), \hat{Bias}(\hat{m}_{Y_1}(\bar{X}))) - 2Cov(\hat{m}_{Y_0}(\bar{X}), \hat{Bias}(\hat{m}_{Y_0}(\bar{X})))$, *and* $Cov(\cdot)$ *represents covariance. Because of the additional variability introduced by bias correction, we can construct a* $100(1-\alpha)\%$ *bias-corrected confidence interval* (CI^{bc}_{RD}) *for* τ_{RD} *to improve coverage probability, in the sense of* $P[\tau_{RD} \in CI^{bc}_{RD}] = 1-\alpha+o(1)$. *The difficulty comes from the derivation and approximation of the aforementioned terms. We discuss such a bias-corrected estimator in detail in a separate paper.*

Nevertheless, the above inference procedure based on the plug-in method has at least two drawbacks. The first is that the bias is taken into account when constructing CI^{u}_{RD}, but the additional variability caused by the bias is ignored; see Remark 4.2.52. Thus, the confidence interval CI^{u}_{RD} may undercover the true estimate and the hypothesis test could over reject a valid null hypothesis. The other is related to the calculation issue. The mentioned plug-in calculations may not be correct enough due to the complicated and unknown quantities in the bias and variance terms. Despite the fact that the bias can be asymptotically ignored by undersmoothing and the unknown terms can be estimated by kernel estimators, the estimation process will necessitate the introduction of new tuning parameters to formulate additional nonparametric estimation. An alternative approach that can be used to approximate the variance of the modal estimator is the bootstrap (Horowitz, 1998; Zhang et al., 2020). We thus focus on undersmoothing to ensure that the finite-sample bias of $\hat{\tau}_{RD}$ converges in probability to zero quickly enough than the

variance, and propose a bootstrap procedure in Algorithm 5 that includes four steps to construct a reliable confidence interval in practice. With large data, we should expect the bootstrapped parameter acquired through undersmoothing to be close to the true value of CMTE, i.e., $(Var(\hat{\tau}_{b,RD}))^{-1/2}(\hat{\tau}_{b,RD} - \tau_{RD})$ can be used to approximate the sampling distribution of $(Var(\hat{\tau}_{RD}))^{-1/2}(\hat{\tau}_{RD} - \tau_{RD})$.[9] Notice that different from the traditional mean bootstrap algorithm, we must use the pseudo random sample $\{\hat{\epsilon}_i^*\}_{i=1}^n$ drawn from its empirical distribution and calculate the centered-in-mode instead of the centered-in-mean residual to ensure $Mode(\hat{\epsilon}_i^*) = Mode(\hat{\epsilon}_i - Mode(\hat{\epsilon}_i))= 0$ in **S-2**.

Remark 4.2.53 *With the use of undersmoothing, we do not incorporate the bias correction strategy into the above bootstrap algorithm. However, it can be easily generated to a new bootstrap algorithm to account for the bias and corresponding variability, where we include a step **S-4'** before **S-4** to repeat **S-2**-**S-3** for B times to obtain the bias, i.e., $\Delta_{bias}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\tau}_{b,RD} - (\hat{m}_{Y_1}(\bar{X}) - \hat{m}_{Y_0}(\bar{X}))$. We can then repeat **S-2**-**S-4'** to calculate the variability; see He and Bartalotti (2020), which proposed a similar wild bootstrap to construct the bias-corrected confidence interval for the mean treatment effect in the FRD design and showed that the bootstrap procedure is asymptotically equivalent to that of Calonico et al. (2014). To mitigate the effect of the second derivative, we prefer to use the second-order local polynomial to generate the bootstrapped data in the first step **S-1** instead of the local linear approximation in terms of the implementation of this new bias-corrected bootstrap algorithm. Under the same conditions as those in Theorem 4.2.20, it can be demonstrated that*

[9] We have checked the validity of this bootstrap algorithm via simulations in Section 4.3. We can also establish the bootstrap consistency for the modal SRD design, i.e., $P[\hat{\tau}_{RD} - z_\alpha^* \leq \tau_{RD} \leq \hat{\tau}_{RD} + z_\alpha^*] \xrightarrow{P} 1 - \alpha$, indicating that the bootstrapped confidence interval has an asymptotic coverage probability of $1 - \alpha$. The detailed proof of this result requires elaborate, lengthy, and sophisticated calculations, which is beyond the scope of this paper. Further investigation into the bootstrap method in the context of modal RD designs is definitely worthwhile.

$$(\,Var(\hat{\tau}_{RD} - \Delta^*_{bias}))^{-1/2}(\hat{\tau}_{RD} - \Delta^*_{bias} - \tau_{RD}) \xrightarrow{d} \mathcal{N}\,(0,1)\,,$$

which implies that the bias-corrected bootstrapped estimator has the same asymptotic distribution as the one described in Remark 4.2.52. The specifics will be illustrated in subsequent research.

Algorithm 5 Bootstrap Algorithm for Modal SRD Design

S-1 Estimate $\hat{m}_{Y_1}(X_+)$ and $\hat{m}_{Y_0}(X_-)$ from Algorithm , let $\hat{m}(X) = \hat{m}_{Y_1}(X_+)$ if $X_i \geq \bar{X}$ and $\hat{m}(X) = \hat{m}_{Y_0}(X_-)$ if $X_i < \bar{X}$, and obtain the residual $\hat{\epsilon}_i = Y_i - \hat{m}(X_i)$ for all i.

S-2 Compute the centered-in-mode residual $\hat{\epsilon}^*_i = \hat{\epsilon}_i - Mode(\hat{\epsilon}_i)$, where $Mode(\hat{\epsilon}_i)$ is achieved via kernel density estimation, and generate the bootstrapped residuals $\{\hat{\epsilon}^*_{b,i}\}_{i=1}^n$ with replacement from the empirical distribution function of $\hat{\epsilon}^*_i$.

S-3 Calculate $Y^*_{b,i} = \hat{m}(X_i) + \hat{\epsilon}^*_{b,i}$ and estimate $\hat{\tau}_{b,RD}$ using the new samples $\{Y^*_{b,i}, X_i, D_i\}_{i=1}^n$ with the same bandwidths as **S-1**, that is,

$$\hat{\tau}_{b,RD} = \hat{m}^b_{Y_1}(\bar{X}) - \hat{m}^b_{Y_0}(\bar{X}).$$

S-4 Repeat **S-1**-**S-3** for B times (e.g., $B = 200$) with the new samples $\{Y^*_{b,i}, X_i, D_i\}_{i=1}^n$ for $b = 1, 2, \cdots, B$, and construct the α percentile bootstrapped confidence interval based on $\{\hat{\tau}_{b,RD}\}_{b=1}^B$ to find z^*_α such that

$$P^*\left(\frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_{b,RD} - z^*_\alpha \leq \tau_{RD} \leq \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_{b,RD} + z^*_\alpha\right) = 1 - \alpha,$$

where P^* is the probability measure induced by bootstrap sampling conditional on the estimation data. Then, the bootstrapped $1-\alpha$ confidence interval for $\hat{\tau}_{RD}$ is

$$CI^B_{RD} = [\hat{\tau}_{RD} - z^*_\alpha,\ \hat{\tau}_{RD} + z^*_\alpha].$$

4.3 Numerical Examples

We present Monte Carlo simulations and a real data analysis to demonstrate that the proposed estimator works effectively in moderate sample sizes. Throughout this section, the previously specified bandwidth selection procedure is implemented, and the kernel functions for $\phi(\cdot)$ and $K(\cdot)$ are fixed as normal kernels, i.e., $\frac{1}{h\sqrt{2\pi}}\exp(-(\cdot)^2/2h^2)$ in which h is a bandwidth.

4.3.1 Monte Carlo Experiments

We carry out simulation experiments to illustrate the finite sample performance of the proposed estimator in this part, including two Monte Carlo experiments with a skewed error distribution. We use DGP to represent the data generating process, in what follows, and compare the mode treatment effect estimates to those of the mean treatment effect from local linear mean regression,[10] which is standard in the RD design literature. The optimal bandwidth for local linear mean regression is determined using the R package *rdrobust* with the option *bwselect=mserd*. The sample sizes we consider are $n \in \{200, 400, 600, 1000\}$, with the number of observations on either side of the cutoff being different. A total of M =200 simulation replications are conducted in all experiments, and the data are *i.i.d.* draws in each replication. We compute the average value of the treatment effect, the standard error (SE), and the MSE for all estimators considered to assess the performance of estimators for each simulation, where

[10]Due to time consuming and space limitations, we do not compare the CMTE to the quantile treatment effect. However, we can apply the method proposed in Frandsen et al. (2012) directly to calculate the quantile treatment effect; see Figure 4.1 for an illustration of the differences between treatment effects.

$$MSE(\hat{\tau}) = \frac{1}{M}\sum_{l=1}^{M}\left(\hat{\tau}_l - \tau\right)^2,$$

$\hat{\tau}_l$ represents the lth estimate, and τ denotes the true value of the treatment effect.

DGP 1 We first generate random samples from the following DGP

$$Y_i = m(X_i) + D_i\tau + X_i\epsilon_i,\ i = 1, \cdots, n,$$

where $m(X_i) = X_i^2$, $D_i = \mathbf{1}\left(X_i \geq \bar{X}\right)$, τ is chosen to be 2, and X_i follows the uniform distribution on $[-2, 2]$. To accommodate different structures between mean regression and modal regression, we set $\epsilon_i \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ with $E(\epsilon_i) = 0$ and $Mode(\epsilon_i) = 1$ (Yao and Li, 2014; Ullah et al., 2022). The model has a jump at $\bar{X} = 0.5$, which is assumed to be known in advance. We thus have the conditional modal function

$$Mode(Y_i \mid X_i) = X_i^2 + X_i + \mathbf{1}\left(X_i \geq 0.5\right)\tau,$$

and the conditional mean function with heteroskedasticity

$$E(Y_i \mid X_i) = X_i^2 + \mathbf{1}\left(X_i \geq 0.5\right)\tau,$$

where $\lim_{X\uparrow 0.5} Mode(Y_0 \mid X_i = \bar{X}) = 0.75$, $\lim_{X\downarrow 0.5} Mode(Y_1 \mid X_i = \bar{X}) = 2.75$, $\lim_{X\uparrow 0.5} E(Y_0 \mid X_i = \bar{X}) = 0.25$, and $\lim_{X\downarrow 0.5} E(Y_1 \mid X_i = \bar{X}) = 2.25$. We can therefore have the direct causal effect of interest $\tau_{RD} = \tau_{mean} = 2$. We are not comparing the efficiency of estimators since the modal regression function differs from the mean regression function.[11]

Table 4.1 presents the simulation results of the studied estimators for DGP 1, which are in qualitative agreement with the main theoretical results from the current paper, as well as the existing modal regression and RD design studies. The bold numbers indicate

[11] We also run the simulation for the homoskedasticity case of DGP 1, where $Y_i = m(X_i) + D_i\tau + \epsilon_i$. Other settings are identical. The results do not show much difference with the observations in DGP 1, except that the modal estimator cannot beat the mean estimator in terms of MSE when homoskedasticity is present.

the smaller values of the results obtained from the mean and modal regressions, while the values in brackets represent standard errors. One can see at a glance that both mean and modal regressions are capable of capturing the true values of treatment effects with a reasonable bias that decreases with the increase in sample size. In addition, when compared to modal regression, mean regression can estimate treatment effect with less bias but with a larger standard error. From another perspective, the estimation results indicate that modal regression can obtain the CMTE estimator with a smaller standard error when heteroskedasticity is present. Therefore, it is not surprising that the CMTE estimator has a lower MSE in all cases for the DGP 1 settings. We also report the coverage rates of the bootstrapped confidence interval for the CMTE parameter with $B = 200$, showing that the coverage rates are close to the nominal converge probabilities. These results indicate that the proposed bootstrap algorithm works well in practice. According to the asymptotic property, the optimal rate of convergence for the modal estimator is $O(n^{-1/4})$, which is slower than that of the mean estimator. We would then expect the modal estimator to be less accurate (in terms of MSE) than the mean estimator as sample size grows. This observation, however, does not occur in this simulation because the underlying models for the modal and mean regressions are different, as well as the usage of undersmoothing for modal regression. Table 4.1 also shows that undersmoothing may not completely eliminate the estimator's bias—the bias of the CMTE estimator remains manageable (though very small) when the number of observations becomes large.

Following the practice in the RD designs, we present the visual results for one set of simulated observations in Figure 4.4 for various choices of sample size. The plots clearly

indicate that there is a "jump" at point 0.5 in the outcome variable. The developed estimation procedure captures the modal regression lines well, confirming the good performance of the newly proposed modal regression for treatment effect analysis. Although the magnitudes of treatment effects are the same for mean and modal regressions, the underlying data generating mechanisms are completely different, as demonstrated by varied fitted lines.

Table 4.1: Results of Simulations—DGP 1

Sample Size	Mean Treatment Effect (SE)	MSE	CMTE (SE)	MSE	1-$\alpha = 0.99$	1-$\alpha = 0.95$	1-$\alpha = 0.90$
n=200	1.9534 (0.5280)	0.2795	2.0514 (**0.4523**)	**0.2061**	0.9743	0.9327	0.8893
n=400	2.0186 (0.3424)	0.1170	2.0446 (**0.3089**)	**0.0969**	0.9806	0.9388	0.8901
n=600	2.0179 (0.2822)	0.0796	2.0400 (**0.2401**)	**0.0590**	0.9843	0.9416	0.8923
n=1000	2.0146 (0.2228)	0.0496	2.0241 (**0.1092**)	**0.0125**	0.9889	0.9443	0.8927
True Value	$\tau_{mean} = 2$		$\tau_{RD} = 2$				

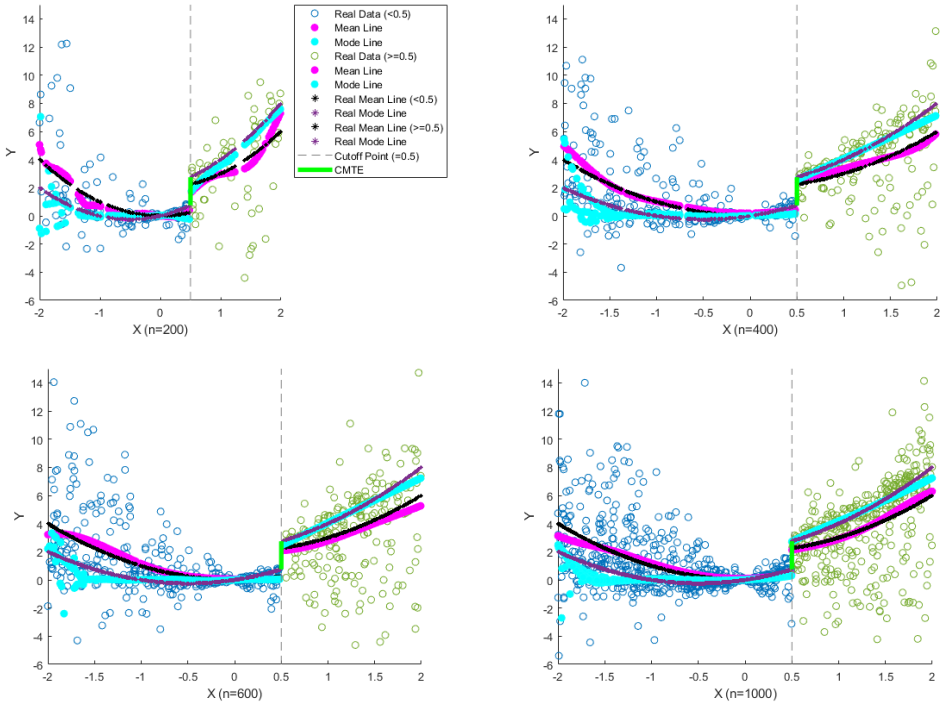

Figure 4.4: Visual Results of DGP 1 for One Set of Simulated Observations

In order to evaluate the asymptotic normality of the CMTE estimator presented in Section 4.2, we compare the shape of the empirical density of the standardized CMTE estimator (and mean estimator) to that of the standard normal density. We plot the approximated distribution of $(\hat{\tau} - E(\hat{\tau}))/\sqrt{\text{Var}(\hat{\tau})}$ instead of $(\hat{\tau} - \tau)/\sqrt{\text{Var}(\hat{\tau})}$ to account for any bias caused by the slow convergence rate and sensitivity to bandwidth. Due to space limitations, we only present the results for $n = 400$ and $n = 1000$ in Figure 4.5. The results for the other sample size schemes are comparable. Figure 4.5 shows that the sample distribution has a similar bell shape to the standard normal distribution and becomes closer when the sample size n increases, which is in accordance with the asymptotic theory in Section 4.2.

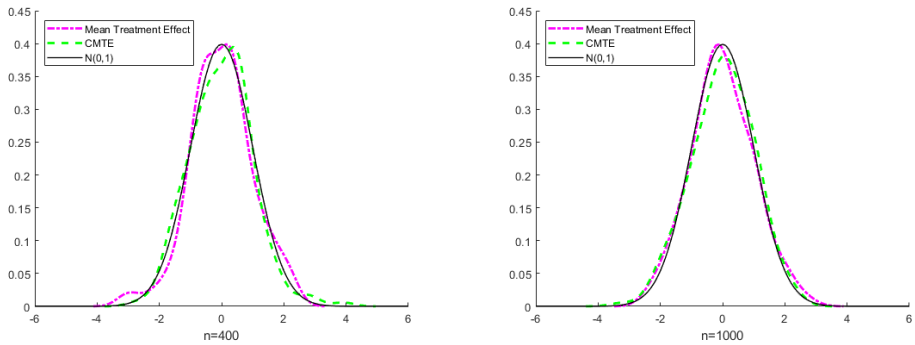

Figure 4.5: Distributions of Standardized Treatment Effects—DGP 1

DGP 2 To further illustrate the newly proposed mode treatment effect with a skewed dataset, we consider the following DGP

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i, \ i = 1, \cdots, n,$$

where $X_i \sim i.i.d.U[-2, 2]$ and $\epsilon_i \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ with $E(\epsilon_i) = 0$ and $Mode(\epsilon_i) = 1$ (Yao and Li, 2014; Ullah et al., 2022). To show that the CMTE can be different from the existing mean treatment effect, we set

$$m(X_i) = \begin{cases} 2 + X_i + 2X_i^2 & \text{if } X_i \geq \bar{X}, \\ 1 + X_i + X_i^2 & \text{if } X_i < \bar{X}, \end{cases} \quad \text{and} \quad \sigma(X_i) = \begin{cases} X_i & \text{if } X_i \geq \bar{X}, \\ 1 & \text{if } X_i < \bar{X}. \end{cases}$$

Thus, we have the conditional modal function

$$Mode(Y_i \mid X_i) = \begin{cases} 2 + 2X_i + 2X_i^2 & \text{if } X_i \geq \bar{X}, \\ 2 + X_i + X_i^2 & \text{if } X_i < \bar{X}, \end{cases}$$

and the conditional mean function

$$E(Y_i \mid X_i) = \begin{cases} 2 + X_i + 2X_i^2 & \text{if } X_i \geq \bar{X}, \\ 1 + X_i + X_i^2 & \text{if } X_i < \bar{X}. \end{cases}$$

We allow treatment to be assigned at the cutoff 0.5, resulting in a jump in the model at $\bar{X} = 0.5$. We then obtain $\lim_{X \uparrow 0.5} Mode(Y_0 \mid X_i = \bar{X}) = 2.75$, $\lim_{X \downarrow 0.5} Mode(Y_1 \mid X_i = \bar{X}) = 3.5$, $\lim_{X \uparrow 0.5} E(Y_0 \mid X_i = \bar{X}) = 1.75$, and $\lim_{X \downarrow 0.5} E(Y_1 \mid X_i = \bar{X}) = 3$. These values suggest that the mode treatment effect is $\tau_{RD} = 0.75$, which is different from the traditional mean treatment effect with $\tau_{mean} = 1.25$.

Table 4.2: Results of Simulations—DGP 2

Sample Size	Mean Treatment Effect (SE)	MSE	CMTE (SE)	MSE	1-α = 0.99	1-α = 0.95	1-α = 0.90
n=200	1.2160 (0.7020)	0.4916	0.8910 (**0.5770**)	**0.3512**	0.9724	0.9385	0.8735
n=400	1.2243 (0.4922)	0.2418	0.8388 (**0.3278**)	**0.1148**	0.9781	0.9390	0.8806
n=600	1.2359 (0.4271)	0.1817	0.8040 (**0.2480**)	**0.0641**	0.9822	0.9427	0.8913
n=1000	1.2426 (0.3251)	0.1052	0.8009 (**0.2086**)	**0.0459**	0.9858	0.9436	0.8955
True Value	$\tau_{mean} = 1.25$		$\tau_{RD} = 0.75$				

The simulation results are summarized in Table 4.2, which corroborates the good performance revealed by the theoretical investigation in Section 4.2. The bold numbers in

the table represent the smaller values between the mean and modal estimates. The same comments for Table 4.1 apply here as well. The mean treatment effect estimator performs slightly better in terms of bias.[12] The CMTE estimator, on the other hand, consistently has smaller variance and MSE that decrease with the number of observations, which is expected due to the use of the mode. These observations suggest that in situations where centers of location are needed but distributions are not symmetrically normal, the mode treatment effect should be considered as a supplement to the existing mean treatment effect. Also, the bootstrap inference for the CMTE parameter performs well across all sample sizes considered.

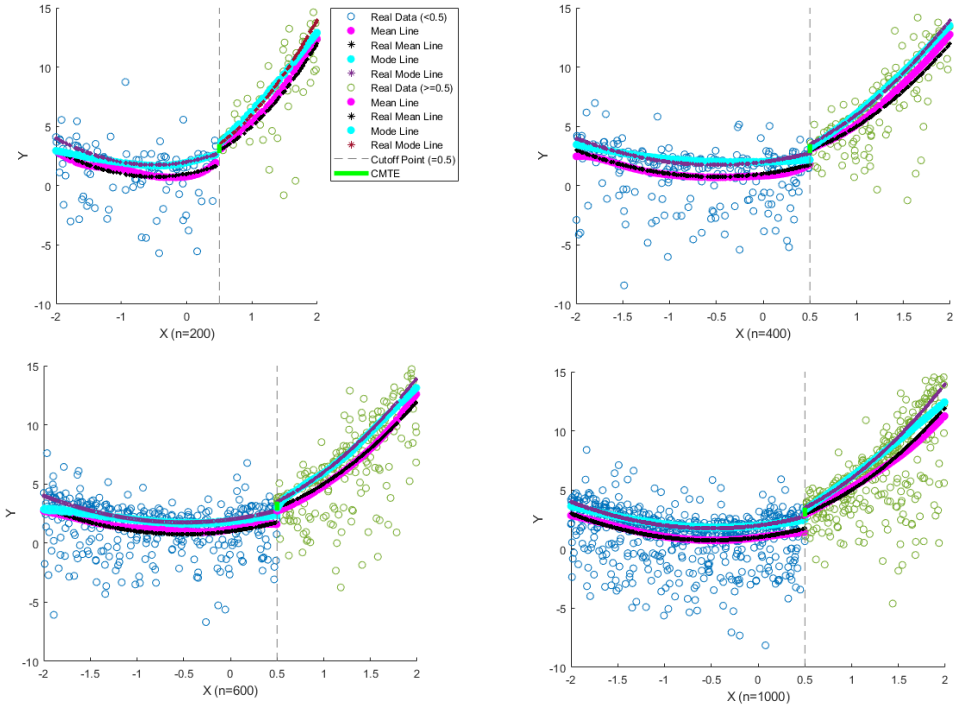

Figure 4.6: Visual Results of DGP 2 for One Set of Simulated Observations

[12]We attribute the bias of CMTE to the slow convergence rate and sensitivity to bandwidth. The performance of the modal estimator may be improved with a more efficient bandwidth selection procedure.

For further graphical illustration, we present the visual results for one set of simulated observations in Figure 4.6 across different choices of sample size n, from which we can see that both mean and modal regressions can identify the treatment effects well. Because of the different magnitudes of treatment effects and the diverse underlying mechanisms of data generation, these plots also suggest that the CMTE can serve as a complement to the existing treatment effects to disclose the whole treatment features.

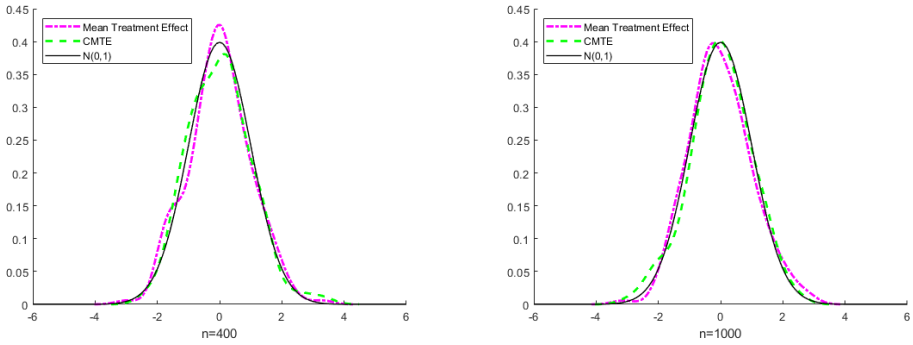

Figure 4.7: Distributions of Standardized Treatment Effects—DGP 2

Similar to DGP 1, Figure 4.7 depicts the approximated distributions of the standardized treatment effect estimates, which is consistent with the asymptotic result presented in Section 4.2. Particularly, with a larger sample size, the estimator approaches a normal distribution. Taken in conjunction with the estimation findings presented in Table 4.2, the plots indicate the necessity of undersmoothing with bootstrap in practice at certain places.

4.3.2 Empirical Analysis: JOBS Act

We follow Mitts (2014) to study the effect of section 601(a)(2) of the Jumpstart Our Small Business (JOBS) Act of 2012, which modified the cutoff for unlisted banks and bank holding companies (BHCs) to deregister a class of securities from 300 to 1200 shareholders of record

under the Securities Exchange Act. This imposed cutoff provides researchers with a natural quasi-experiment to identify the effect of deregistration on the performance of banks and BHCs, because the location of banks and BHCs around the cutoff is as good as randomly assigned (Mitts, 2014), which can be analyzed by applying RD designs.

Table 4.3: Results of Treatment Effects of JOBS Act

	CMTE	90% CI^B_{RD}	95% CI^B_{RD}	99% CI^B_{RD}	Mean Treatment Effect
Total Other Expenses	-0.0014	[-0.0357, 0.0085]	[-0.0412, 0.0087]	[-0.0490, 0.0088]	-0.0012
Net Income	0.4073	[0.4022, 0.6275]	[0.3892, 0.6287]	[0.3804, 0.6298]	0.4231
Total Noninterest Expenses	-0.6822	[-0.7371, -0.5468]	[-0.7383, -0.5448]	[-0.7394, -0.5440]	-0.7721
Total Pretax Expenses	-0.7057	[-0.6161, -0.5778]	[-0.7385, -0.5759]	[-0.8058, -0.5731]	-0.6843

Different from Mitts (2014) who used the FRD design to analyze the mean treatment effect, we employ nonparametric modal regression to investigate the CMTE of this JOBS act on the financial performance of banks and BHCs through the SRD design. The dependent variables we focus on are Total Other Expenses, Net Income, Total Noninterest Expenses, and Total Pretax Expenses, which are ratios from the Uniform Bank Performance Report. The running variable is Shareholders of Record, which consists of the number of shareholders at the time of deregistration, and the cutoff is 1200. We refer to Mitts (2014) for the detailed explanation of those variables. The dataset used in this paper was downloaded from Harvard Dataverse, which includes quarterly data of 187 banks and BHCs from January 1, 2003 to December 31, 2013.[13] Compared to the original dataset, we exclude firms that have missing points for any of those five variables and end up with 5733 observations for each variable.

[13]https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/Q2OHRH.
The detailed sample summary statistics for the full sample are omitted in this paper for space consideration.

The bandwidth selection for mean regression is carried out in the same way as in simulation studies. Table 4.3 summarizes the numerical estimation results of treatment effects and reports the values of confidence intervals calculated by the bootstrap Algorithm 5 with $B = 200$. It shows that the CMTEs for Net Income and Total Noninterest Expenses are significant at the 10%, 5%, and 1% significance levels, respectively, while the CMTE for Total Other Expenses is not significant at any significance level. It is interesting to observe that the CMTE for Total Pretax Expenses is not included in the 90% confidence interval. For all four outcome variables we consider, the CMTEs have consistent signs with the mean treatment effects, indicating that the JOBS Act can reduce expenses and increase the net income of banks and BHCs in either the mean or mode sense. However, the magnitudes are different between CMTEs and mean treatment effects. In particular, the JOBS Act can reduce Total Other Expenses and Total Pretax Expenses more but Total Noninterest Expenses less and increase Net Income less in terms of mode value. This difference demonstrates that the proposed modal RD designs can provide valuable information in practice that is not available from the conventional approach.

The visual results with the conditional mean and modal functions are reported in Figure 4.8 to graphically illustrate the difference between CMTEs and mean treatment effects. In accordance with the findings in Table 4.3, the "jump" of the variable Total Other Expenses at the cutoff is not apparent in the plot, indicating that the treatment effect is somewhat insignificant. In addition, we can find from those plots that the CMTE is different from the mean treatment effect when we have a skewed dataset, such as Net Income, Total Noninterest Expenses, and Total Pretax Expenses, and thus can serve as a complement to

the existing treatment effects to reveal the "most likely" effect. On the other hand, when the dependent variable is Total Other Expenses, the modal regressions (estimates) are not significantly different from the mean regressions (estimates) because the data are nearly symmetrical in distribution.

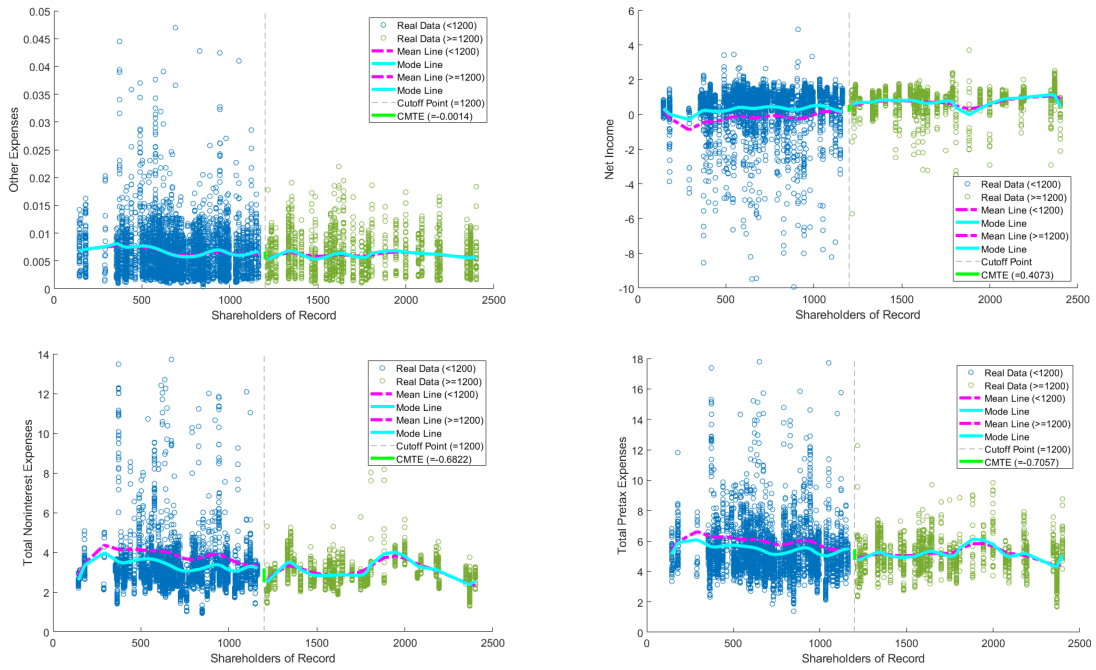

Figure 4.8: Visual Results of Empirical Analysis of JOBS Act

4.4 Extensions

This paper has concentrated exclusively on the modal SRD design. To broaden the application of CMTE, we in this section extend the proposed model to several related topics that are of either practical or theoretical importance.

4.4.1 Including Additional Covariates

Although controlling for additional "pre-intervention" covariates $Z \in R^k$ may not be necessary if we are only interested in local treatment effects, in practice, we can have the benefit of adding Z into regression to control the sample bias (or different observed characteristics) and reduce the sampling variability (improve efficiency); see Calonico et al. (2019) for the related discussion on the RD designs using additional covariates for mean regression. In accordance with the phenomenon of the mean or quantile treatment effect, the presence of Z will not change the identification and estimation strategies in this paper as long as the listed assumptions continue to hold conditionally on X and Z or the distribution of Z is balanced at the cutoff. With a nonseparable function $m(\cdot)$, including Z allows us to improve the convergence rate of the CMTE estimator to the rate for one-dimensional nonparametric modal regression regardless of the dimension of Z, because we can have

$$\tau_{RD} = E_Z \left(\lim_{X \downarrow \bar{X}} Mode(Y \mid X, Z) - \lim_{X \uparrow \bar{X}} Mode(Y \mid X, Z) \mid X = \bar{X} \right) \tag{4.18}$$

by taking the mean difference over all possible values of Z, that is, the CMTE on all treated individuals without conditioning on Z. Such an unconditional CMTE can also be used to examine the robustness of the results to the set of control variables. Furthermore, under mild regularity conditions, we can avoid fully nonparametric estimation over $(X, Z^T)^T \in R^{1+k}$, where T represents the transpose of a vector or matrix, and consider an additively separable linear modal regression if k is large. Therefore, the "curse of dimensionality" does not apply. We provide a parametric version of CMTE with covariates in the appendix.

4.4.2 Multiple Running Variables

The proposed model structure can also cope with situations in which people are assigned to more than two treatments depending on the values of the multiple running variables (Papay et al., 2011). This is quite common in practice. For example, students may need to pass multiple subject exams in order to advance to the next grade level, or geography units are assigned to treatment on the basis of latitude and longitude. When multiple running variables are present, the discontinuity becomes a boundary, and the CMTE can be measured at every point along the treatment boundary. Suppose that we have two running variables S_1 and S_2. A unit will receive the treatment if $S_1 \geq \bar{S}_1$ and $S_2 \geq \bar{S}_2$, where $\bar{S}_1$ and $\bar{S}_2$ are two cutoffs. On the other hand, a unit will receive no treatment in the following three situations $S_1 < \bar{S}_1$ & $S_2 \geq \bar{S}_2$, $S_1 \geq \bar{S}_1$ & $S_2 < \bar{S}_2$, and $S_1 < \bar{S}_1$ & $S_2 < \bar{S}_2$. Define D_{S_1} and D_{S_2} in the same way as that D_i. With mode rank invariance, we have

$$\begin{aligned}\tau_{RD} =& Mode(Y_{11}^{D_{S_1}=1,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2) - Mode(Y_{00}^{D_{S_1}=0,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2) \\ &- \Big[Mode(Y_{10}^{D_{S_1}=1,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2) - Mode(Y_{00}^{D_{S_1}=0,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2)\Big] \\ &- \Big[Mode(Y_{01}^{D_{S_1}=0,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2) - Mode(Y_{00}^{D_{S_1}=0,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2)\Big], \qquad (4.19)\end{aligned}$$

where the two partial effects relative to $Y_{00}^{D_{S_1}=0,D_{S_2}=0}$ are subtracted due to each running variable crossing its own cutoff. For identification, we assume that the density function $f_{S_1,S_2}(S_1, S_2)$ is strictly positive on a neighborhood of $(\bar{S}_1, \bar{S}_2)$ and impose the following assumptions

$$(1)\ Mode(Y_{00}^{D_{S_1}=0,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2) = Mode(Y_{00}^{D_{S_1}=1,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2);$$

$$(2)\ Mode(Y_{10}^{D_{S_1}=1,D_{S_2}=0} \mid \bar{S}_1, \bar{S}_2) = Mode(Y_{10}^{D_{S_1}=1,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2);$$

(3) $Mode(Y_{01}^{D_{S_1}=0,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2) = Mode(Y_{01}^{D_{S_1}=1,D_{S_2}=1} \mid \bar{S}_1, \bar{S}_2)$.

The above assumptions demonstrate how counterfactuals for the treatment group with $(\bar{S}_1, \bar{S}_2)$ can be identified. After that, τ_{RD} can be rewritten as

$$\begin{aligned}\tau_{RD} = &\lim_{(S_1,S_2)\to(S_{1,+},S_{2,+})} Mode(Y_i \mid S_1, S_2) - \lim_{(S_1,S_2)\to(S_{1,-},S_{2,-})} Mode(Y_i \mid S_1, S_2) \\ &- \big[\lim_{(S_1,S_2)\to(S_{1,+},S_{2,-})} Mode(Y_i \mid S_1, S_2) - \lim_{(S_1,S_2)\to(S_{1,-},S_{2,-})} Mode(Y_i \mid S_1, S_2)\big] \\ &- \big[\lim_{(S_1,S_2)\to(S_{1,-},S_{2,+})} Mode(Y_i \mid S_1, S_2) - \lim_{(S_1,S_2)\to(S_{1,-},S_{2,-})} Mode(Y_i \mid S_1, S_2)\big].\end{aligned} \tag{4.20}$$

If there are no partial effects, we obtain

$$\tau_{RD} = \lim_{(S_1,S_2)\to(S_{1,+},S_{2,+})} Mode(Y_i \mid S_1, S_2) - \lim_{(S_1,S_2)\to(S_{1,-},S_{2,-})} Mode(Y_i \mid S_1, S_2). \tag{4.21}$$

By substituting the identified elements with sample versions, we can estimate the CMTE.

4.4.3 Multiple Cutoffs

In reality, it is common to observe the presence of multiple cutoffs. For instance, a school test score cutoff may vary across geographic regions. The existence of multiple cutoffs allows the researcher to learn about the causal effect at various levels of the running variable. Following Cattaneo et al. (2016), we in this case can pool the data from multiple cutoffs to produce a single CMTE estimate. Assume that the cutoff has finite support and there is a random variable C_i denoting the cutoff of each unit with support $\{c_1, c_2, \cdots, c_J\}$ and $P(C_i = c) = P_c \in [0, 1]$. Let $f_{X|C}(x \mid c)$ represent a continuous conditional density of $X_i|C_i = c$ (for the previous analysis, C_i is a fixed value with $P(C_i = \bar{X}) = 1$). We have $\lim_{\varepsilon\to 0^+} E[D_i \mid X_i = c + \varepsilon, C_i = c] = 1$, $\lim_{\varepsilon\to 0^+} E[D_i \mid X_i = c - \varepsilon, C_i = c] = 0$, and

$Y_{di}(C_i) = \sum_{c \in C} \mathbf{1}(C_i = c) Y_{di}(c)$ for $d = 0, 1$ ($2J$ potential outcomes). With mode rank invariance, by assuming that the outcomes under treatment and control are continuous functions of the running variable at all possible cutoffs, we have the following weighted average of local mode treatment effects

$$\begin{aligned}\tau_{RD} &= \lim_{\varepsilon \to 0^+} Mode[Y_i \mid \tilde{X}_i = \varepsilon] - \lim_{\varepsilon \to 0^+} Mode[Y_i \mid \tilde{X}_i = -\varepsilon] \\ &= \sum_{c \in \mathcal{C}} \{Mode\left[Y_{1i}(c) \mid X_i = c, C_i = c\right] - Mode\left[Y_{0i}(c) \mid X_i = c, C_i = c\right]\} \omega(c), \end{aligned} \tag{4.22}$$

where $\tilde{X}_i = X_i - C_i$ and $\omega(c) = f_{X|C}(c \mid c) P[C_i = c] / \sum_{c \in \mathcal{C}} f_{X|C}(c \mid c) P[C_i = c]$ determines the effects that are included in the pooled estimate and how much each effect contributes to the pooled estimate. Particularly, for any given $\varepsilon > 0$, we have

$$\begin{aligned} Mode[Y_i \mid \tilde{X}_i = \varepsilon] &= E\left\{ Mode\left[Y_i \mid X_i - C_i = \varepsilon, C_i\right] \mid \tilde{X}_i = \varepsilon \right\} \\ &= \sum_{c \in \mathcal{C}} Mode\left[Y_i \mid X_i - C_i = \varepsilon, C_i = c\right] P\left[C_i = c \mid \tilde{X}_i = \varepsilon\right] \\ &= \sum_{c \in \mathcal{C}} Mode\left[Y_{1i}(c) \mid X_i = c + \varepsilon, C_i = c\right] P\left[C_i = c \mid \tilde{X}_i = \varepsilon\right], \end{aligned} \tag{4.23}$$

and

$$Mode\left[Y_i \mid \tilde{X}_i = -\varepsilon\right] = \sum_{c \in \mathcal{C}} Mode\left[Y_{0i}(c) \mid X_i = c - \varepsilon, C_i = c\right] P\left[C_i = c \mid \tilde{X}_i = -\varepsilon\right]. \tag{4.24}$$

Meanwhile, we know that

$$P\left[C_i = c \mid \tilde{X}_i = x\right] = \frac{f_{\tilde{X}|C}(x \mid c) P\left[C_i = c\right]}{f_{\tilde{X}}(x)} = \frac{f_{X|C}(c + x \mid c) P\left[C_i = c\right]}{\sum_{c \in \mathcal{C}} f_{X|C}(c + x \mid c) P\left[C_i = c\right]}. \tag{4.25}$$

Under the assumptions of continuity and finite support, we can get the result by interchanging limits and sums. The pooled CMTE is then a weighted average of the mode treatment effects at each cutoff when there exist multiple cutoffs,

4.4.4 Modal FRD Design

The discussion of CMTE in this paper is based on the SRD design. However, it is common for practical applications of RD designs to be fuzzy rather than sharp. We can then relax the sharp assignment mechanism from the previous sections and study the FRD case. In the FRD design, the treatment assignment rule remains the same as that in the SRD design, but D_i is not necessarily equal to 1 when $X_i \geq \bar{X}$ or 0 when $X_i < \bar{X}$ (Figure 4.9), since X_i is not informative enough to determine the treatment. Instead, it can affect the probability of treatment in a discontinuous way when it exceeds a certain cutoff $\bar{X}$, i.e., $\lim_{X \to X_+} P(D = 1 \mid X) > \lim_{X \to X_-} P(D = 1 \mid X)$. This indicates that treatment is not solely determined by the strict cutoff rule in the FRD design. Some individuals who are above the cutoff may not receive treatment, while some individuals who are below the cutoff may receive treatment. Hahn et al. (2001) showed that the mean treatment effect in the FRD design can be obtained by taking the ratio of the difference in outcomes and the difference in treatment probabilities at the cutoff $\bar{X}$. Frandsen et al. (2012) introduced a nonparametric estimator for local quantile treatment effect in the RD designs, including the fuzzy case. However, because the mode lacks the additive property, it is not straightforward to extend the results from the mean or quantile FRD design to the modal FRD design. Providing that the modal regression line coincides with a quantile regression line, we can propose a simple approach considering from density function to recover the estimate of CMTE in the FRD design.

Especially, instead of investigating the CMTE on the basis of nonparametric modal regression at the boundary point, we define the CMTE in the FRD design as

$$\tau_{RD} = Mode\left(Y_1 \mid i \text{ is complier}, \bar{X}\right) - Mode\left(Y_0 \mid i \text{ is complier}, \bar{X}\right)$$

$$= \arg\max_{Y_1}\{f_{Y_1|D,X}(Y_1)\} - \arg\max_{Y_0}\{f_{Y_0|D,X}(Y_0)\}. \tag{4.26}$$

Because the CMTE in the FRD design is acquired from the quantile treatment effect, all conditions imposed for the quantile FRD design should be satisfied here as well; see Frandsen et al. (2012). As a result, the mode rank invariance condition is met automatically.

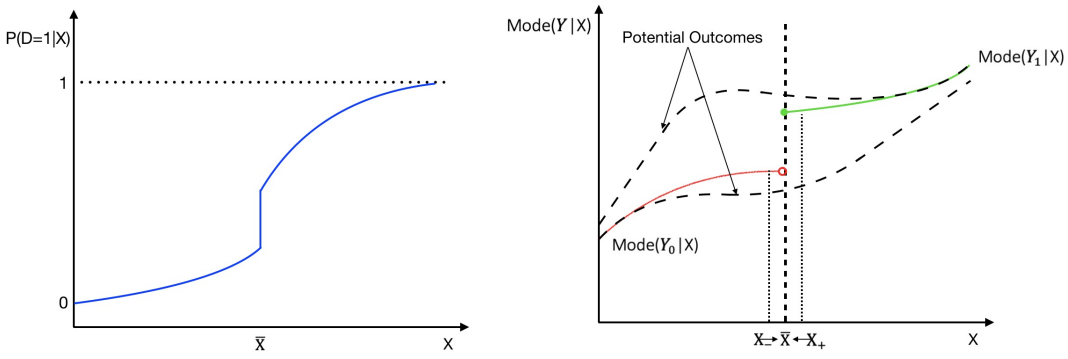

Figure 4.9: Modal Fuzzy Regression Discontinuity

The conditional distributions of the potential outcomes in (4.26) are identified by using two separate Wald representations for $\mathbf{1}\left(Y_i \leq y\right)$ interacted with the treatment state

$$F_{Y_1|D,X=\bar{X}}(y) = \frac{\lim_{X_i \downarrow \bar{X}} E\left[\mathbf{1}\left(Y_i \leq y\right) D_i \mid X_i = \bar{X}\right] - \lim_{X_i \uparrow \bar{X}} E\left[\mathbf{1}\left(Y_i \leq y\right) D_i \mid X_i = \bar{X}\right]}{\lim_{X_i \downarrow \bar{X}} E\left[D_i \mid X_i = \bar{X}\right] - \lim_{X_i \uparrow \bar{X}} E\left[D_i \mid X_i = \bar{X}\right]}, \tag{4.27}$$

$$F_{Y_0|D,X=\bar{X}}(y)$$
$$= \frac{\lim_{X_i \downarrow \bar{X}} E\left[\mathbf{1}\left(Y_i \leq y\right)\left(1 - D_i\right) \mid X_i = \bar{X}\right] - \lim_{X_i \uparrow \bar{X}} E\left[\mathbf{1}\left(Y_i \leq y\right)\left(1 - D_i\right) \mid X_i = \bar{X}\right]}{\lim_{X_i \downarrow \bar{X}} E\left[\left(1 - D_i\right) \mid X_i = \bar{X}\right] - \lim_{X_i \uparrow \bar{X}} E\left[\left(1 - D_i\right) \mid X_i = \bar{X}\right]}, \tag{4.28}$$

and the sample analogues can be estimated by local linear weighted two-stage least squares. After obtaining estimates $\hat{F}_{Y_1|D,X=\bar{X}}(y)$ and $\hat{F}_{Y_0|D,X=\bar{X}}(y)$, we can calculate the mode values using

$$Mode\left(Y_1 \mid \bar{X}\right) = \arg\max_{Y_1} \frac{\hat{F}_{Y_1|D,X=\bar{X}}(Y_1+\lambda_1) - \hat{F}_{Y_1|D,X=\bar{X}}(Y_1-\lambda_1)}{2\lambda_1}, \tag{4.29}$$

$$Mode\left(Y_0 \mid \bar{X}\right) = \arg\max_{Y_0} \frac{\hat{F}_{Y_0|D,X=\bar{X}}(Y_0+\lambda_0) - \hat{F}_{Y_0|D,X=\bar{X}}(Y_0-\lambda_0)}{2\lambda_0}, \tag{4.30}$$

where λ_1 and λ_0 are two chosen tuning parameters. In light of those estimates, the value of CMTE in the FRD design can be achieved straightforwardly. Such a quantile-based CMTE estimator includes the estimator in the SRD design as a special case. The detailed asymptotic properties and inferences will be investigated in the future.

4.5 Concluding Remarks

The increasing popularity of RD methods for causal inference has led to a large number of different models and estimating strategies, under which treatment effects are expressed as comparisons between features of the distributions of both potential outcomes, such as their means or quantiles. In this paper, we develop an alternative model for estimation and statistical inference in the RD designs. Particularly, we propose a valuable measure, CMTE, to complement the existing mean and quantile treatment effects under the assumption of mode rank invariance, as in certain situations policy makers may be interested especially in the effects at the highest point and their primary concern is likely with a larger group of people. We show that when the data are asymmetrically distributed, econometricians should pay careful attention to interpreting modal RD estimates. We nonparametrically estimate CMTE in the SRD design with local linear modal regression and provide asymptotic normality for the proposed estimator at the cutoff under some mild assumptions. We numerically estimate the developed model by virtue of a modified MEM algorithm. With

the optimal bandwidths, the resultant CMTE estimator is consistent with a $n^{-1/4}$ rate, which is the same as the optimal convergence rate for nonparametric modal regression with local linear approximation for the interior points but slower than that for nonparametric mean regression. The bootstrap algorithm relying on undersmoothing is presented for constructing the confidence interval. We in the end show a simple method based on quantile regression to derive the CMTE in the FRD design. Several other extensions are investigated as well to enlarge the applicability of the suggested CMTE.

The novel mode treatment effect suggested in this paper has a wide range of applications in economics, statistics, social science, and other related fields, because it can capture the "most likely" effect and be robust to outliers. While the present paper focuses on the modal RD designs in the classical sharp and fuzzy settings, there are many other directions related to CMTE that deserve further research. For example, we focus on the continuous case of a running variable with local linear approximation in this paper, as the bandwidth cannot be shrunk beyond a certain point. However, it is of interest to extend to the discrete distribution with a modest number of points of support, which often arises from various scenarios, including many social and economic studies. In such a case, it is no longer possible to find treated and control units with values of the running variable that are arbitrarily close. We can then solve such an issue by the combination of nonparametric estimation with product kernel functions for smoothing discrete data, which is to effectively utilize data information from some nearby neighborhoods that might share similar characteristics with the target. The other issue that deserves to be researched further is bandwidth choice. Choosing a bandwidth has long been a challenging issue in

the nonparametric and RD design literature. In this paper, we require the undersmoothed bandwidths to avoid nuisance bias terms in the limiting distributions of local linear modal estimators. It would be interesting to research other bandwidth selection criteria for the CMTE estimator with a rigorous justification, such as using a kernel-based cross-validation method to fit the curve over the support of the data. Finally, we assume in this paper that the discontinuity point is known, which may not always be the case in practice. It is well understood that the identification of RD designs is a time-consuming manual process that involves human judgment and construction and is therefore subject to human bias. Investigating CMTE by allowing for an unknown discontinuity point will be an interesting research direction, where we can adopt a structural-break detection method for the detection of an unknown cutoff. All of these will be saved for future research.

Chapter 5

Conclusions

Skewed or heavy-tailed data (e.g., wages, prices, scores on a difficult exam, movie ticket sales, and expenditures) appear in a broad variety of practical applications, including economic, statistical, social, and educational research studies, among others. In such instances, the mean estimate may not adequately disclose the data's characteristics, and the mode estimate (one of the center measures) should be considered as a supplemental measure to capture the "most likely" element of the data. Nevertheless, for a long time, mode has not received much attention from researchers. With more available datasets and powerful computation tools, it is important for (applied) econometricians to be aware of the application of modal regression, which focuses on modeling how the conditional mode of the response variable depends on the covariates. This dissertation proposes three new models based on mode value that can broaden the scope of existing modal regressions.

Chapter 2, to the best of our knowledge, is the first work that analyzes the endogeneity issue in modal regression and systematically studies its statistical properties with

the conditional mode independence restriction. In particular, we introduce a computationally efficient two-step estimation procedure based on control function to estimate parametric modal regression with endogeneity, followed by a three-stage estimation method for semiparametric partially linear modal regression with the estimated modal residual from the reduced form equation in the second step.

Chapter 3 novelly introduces a nonparametric modal regression estimator of volatility functions in a general framework that includes the nonlinear time series model as a special case. It shows that the modal regression estimator of conditional volatility could be obtained asymptotically as well as if the mean regression were given. Moreover, Chapter 3 proposes a variance reduction technique in terms of modal volatility estimator to achieve asymptotic relative efficiency and keep the asymptotic bias unchanged. To avoid negativity, a local exponential modal modal volatility is also introduced

Chapter 4 proposes an innovative and valuable measure, CMTE, to complement the existing mean and quantile treatment effects under the assumption of mode rank invariance. It nonparametrically estimates CMTE in the SRD design with local linear modal regression and provides asymptotic normality for the proposed estimator at the cutoff under some mild assumptions. The bootstrap algorithm relying on undersmoothing is presented for constructing the confidence interval. Chapter 4 also shows a simple method based on quantile regression to derive the CMTE in the FRD design.

Bibliography

Abbring, J. H. and Heckman, J. J. (2007). Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation. *in J. J. Heckman & E.E. Leamer (ed.), Handbook of Econometrics*, 6 (72).

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *The American Economic Review*, 91 (5), 1369-1401.

Akritas, M. G. and Van Keilegom, I. (2001). ANCOVA Methods for Heteroscedastic Nonparametric Regression Models. *Journal of the American Statistical Association*, 96 (453), 220-232.

Belloni, A. and Chernozhukov, V. (2013). Least Squares After Model Selection in High-Dimensional Sparse Models. *Bernoulli*, 19, 521-547.

Bhattacharya, P. K. and Zhao, P. L. (1997). Semiparametric Inference in A Partial Linear Model. *The Annals of Statistics*, 25 (1), 244-262.

Blundell, R. W. and Powell, J. L. (2003). Endogeneity in Nonparametric and Semiparametric Response Models. *In:Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress.Cambridge University Press, Cambridge.*

Blundell, R. W. and Powell, J. L. (2004). Endogeneity in Semiparametric Binary Response Models. *Review of Economic Studies*, 71, 655-670.

Borkovec, M. and Klüppelberg, C. (2001). The Tail of the Stationary Distribution of An Autoregressive Process with ARCH(1) Errors. *Annals of Applied Probability*, 11, 1220-1241.

Cai, Z. and Ould-Said, E. (2003). Local M-Estimator for Nonparametric Time Series. *Statistics & Probability Letters*, 65, 433-449.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression Discontinuity Designs Using Covariates. *Review of Economics and Statistics*, 101 (3), 442-451.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression Discontinuity Designs. *Econometrica*, 82 (6), 2295-2326.

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). Optimal Data-Driven Regression Discontinuity Plots. *Journal of the American Statistical Association*, 110, 1753-1769.

Campbell, J. Y. (2003). Consumption-Based Asset Pricing. *In: G. M. Constantinides, M. Harris, R. M. Stulz, North Holland (Eds.), Handbook of the Economics of Finance: Financial Markets and Asset Pricing*, 1 (B), 803-887.

Card, D. (1995). Using Geographic Variation in College Proximity to Estimate the Return to Schooling. *In: Christofides, L., Grant, E.K., Swindinsky, R. (Eds.), Aspects of Labour Economics: Essays in Honour of John Vanderkamp, University of Toronto Press.*

Card, D. (2001). Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 69 (5), 1127-1160.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., and Keele, L. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *The Journal of Politics*, 78, 1229-1248.

Chang, N. C. (2020). Mode Treatment Effect. *arXiv:2007.11606.*

Chen, X., Chernozhukov, V., Lee, S., and Newey, W. K. (2014). Local Identification of Nonparametric and Semiparametric Models. *Econometrica*, 82, 785-809.

Chen, L. H., Cheng, M. Y., and Peng, L. (2009). Conditional Variance Estimation in Heteroscedastic Regression Models. *Journal of Statistical Planning and Inference*, 139 (2), 236-245.

Chen, Y. C. (2018). Modal Regression using Kernel Density Estimation: A Review. *Wiley Interdisciplinary Reviewers: Computational Statistics*, 10: e1431.

Chen, Y. C., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric Modal Regression. *The Annals of Statistics*, 44 (2), 489-514.

Cheng, M., Y. and Peng, L. (2007). Variance Reduction in Multiparameter Likelihood Models. *Journal of the American Statistical Association*, 102 (477), 293-304.

Cheng, M. Y., Peng, L., and Wu, J. (2007). Reducing Variance in Univariate Smoothing. *The Annals of Statistics*, 35 (2), 522-542.

Chernozhukov, V., Fernandez-Val, I., and Kowalski, A. E. (2015). Quantile Regression with Censoring and Endogeneity. *Journal of Econometrics*, 186 (1), 201-221.

Choi, E. and Hall, P. (1998). On Bias Reduction in Local Linear Smoothing. *Biometrika*, 85 (2), 333-345.

de Castro, L. and Galvao, A. F. (2019). Dynamic Quantile Models of Rational Behavior. *Econometrica*, 87, 1893-1939.

de Castro, L., Galvao, A. F., Kaplan, D. M., and Liu, X. (2019). Smoothed GMM for Quantile Models. *Journal of Econometrics*, 213 (1), 121-144.

Dimitriadis, T., Patton, A. J., and Schmidt, P. W. (2020). Testing Forecast Rationality for Measures of Central Tendency. *arXiv:1910.12545v2*.

Dong, Y. and Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *The Review of Economics and Statistics*, 97 (5), 1081-1092.

Engle, R. F. (1982) Autoregressive Conditional Heteroscedasticity with Estimates of Variance of U.K. Inflation. *Econometrica*, 50, 987-1008.

Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric Estimates of the Relation between Weather and Electricity Sales. *Journal of the American Statistical Association*, 81, 310-320.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modelling and Its Application. *London: Chapman and Hall*.

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 Yao, Q. (1998). Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika*, 85 (3), 645-660.

Fan, Q. (2012). The Adaptive Lasso Method for Instrumental Variable Selection. *North Carolina State University*.

Feng, Y., Fan, J., and Suykens, J. A. K. (2020). A Statistical Learning Approach to Modal Regression. *Journal of Machine Learning Research*, 21 (2), 1-35.

Frandsen, B. R., Frölich, M., and Melly, B. (2012). Quantile Treatment Effects in the Regression Discontinuity Design. *Journal of Econometrics*, 168 (2), 382-395.

Fu, W. (1998). Penalized Regressions: The Bridge Versus the Lasso. *Journal of Computational and Graphical Statistics*, 7 (3), 397-416.

Gelman, A. and Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs? *Journal of Business & Economic Statistics*, 37 (3), 447-456.

Giovannetti, B. C. (2013). Asset Pricing under Quantile Utility Maximization. *Review of Financial Economics*, 22, 169-179.

Gouriéroux, Ch. and Monfort, A. (1992). Qualitative Threshold ARCH models. *Journal of Econometrics*, 52, 159-199.

Hahn, J., Todd. P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69, 201-209.

Hall, P. (1993). On Edgeworth Expansion and Bootstrap Confidence Bands in Nonparametric Curve Estimation. *Journal of the Royal Statistical Society Series B*, 291-304.

Havranek, T. (2015). Measuring Intertemporal Substitution: The Importance of Method Choices and Selective Reporting. *Journal of the European Economic Association*, 13, 1180-1204.

He, Y. and Bartalotti, O. (2020). Wild Bootstrap for Fuzzy Regression Discontinuity Designs: Obtaining Robust Bias-Corrected Confidence Intervals. *The Econometrics Journal*, 23 (2), 211-231.

Heckman, N. (1986). Spline Smoothing in A Partly Linear Model. *Journal of the Royal Statistical Society, Series B*, 48, 244-248.

Horowitz, J. L. (1998). Bootstrap Methods for Median Regression Models. *Econometrica*, 66 (6), 1327-1351.

Hsu, Y. C., Kuan, C. M., and Lo, T. Y. (2017). Quantile Treatment Effects in Regression Discontinuity Designs with Covariates. *IEAS Working Paper No. 17-A009, Academia Sinica*.

Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive Lasso for Sparse High-Dimensional Regression Models. *Statistica Sinica*, 18 (4), 1603-1618.

Imbens, G. W. and Lemieux, T. (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*, 142, 615-635.

Imbens, G. W. and Newey, W. K. (2009). Identification and Estimation of Triangular Simultaneous Equations Models without Additivity. *Econometrica*, 77 (5), 1481-1512.

Kai, B., Li, R., and Zou, H. (2011). New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models. *The Annals of Statistics*, 39 (1), 305-332.

Kaplan, D. M. (2020). sivqr: Smoothed IV quantile regression (in Stata). *Working Papers 2009, Department of Economics, University of Missouri*.

Kemp, G. C. R., Parente, P. M. D. C., and Santos Silva, J. M. C. (2020). Dynamic Vector Mode Regression. *Journal of Business & Economic Statistics*, 38 (3), 647-661.

Kemp, G. C. R. and Santos Silva, J. M. C. (2012). Regression towards the Mode. *Journal of Econometrics*, 170 (1), 92-101.

Kim, K. and Petrin, A. (2011). A New Control Function Approach for Nonparametric Regressions with Endogenous Variables. *NBER Working Paper 16679*.

Koenker, R. and Zhao, Q. (1996). Conditional Quantile Estimation and Inference for ARCH Models. *Econometric Theory*, 12, 793-813.

Krief, J. M. (2017). Semi-Linear Mode Regression. *Econometrics Journal*, 20, 149-167.

Lee, D. S. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections. *Journal of Econometrics*, 142 (2), 675-697.

Lee, D. S. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*, 48 (2), 281-355.

Lee, M. (1989). Mode Regression. *Journal of Econometrics*, 42, 337-349.

Lee, M. (1993). Quadratic Mode Regression. *Journal of Econometrics*, 57, 1-19.

Li, J., Ray, S., and Lindsay, B. G. (2007). A Nonparametric Statistical Approach to Clustering via Mode Identification. *Journal of Machine Learning Research*, 8 (8), 1687-1723.

Li, X. and Huang, X. (2019). Linear Mode Regression with Covariate Measurement Error. *Canadian Journal of Statistics*, 47, 262-280.

Ma, L. and Koenker, R. (2006). Quantile Regression Methods for Recursive Structural Equation Models. *Journal of Econometrics*, 134, 471-506.

Mishra, S., Su, L., and Ullah, A. (2010). Semiparametric Estimator of Time Series Conditional variance. *Journal of Business and Economic Statistics*, 28 (2), 256-274.

Mitts, J. (2014). Did the JOBS Act Benefit Community Banks? A Regression Discontinuity Study. *A Regression Discontinuity Study (February 19, 2014)*.

Newey, W. K., Powell, J. L., and Vella, F. (1999). Nonparametric Estimation of Triangular Simultaneous Equations Models. *Econometrica*, 67, 565-603.

Ota, H., Kato, K., and Hara, S. (2019). Quantile Regression Approach to Conditional Mode Estimation. *Electronic Journal of Statistics*, 13, 3120-3160.

Pagan, A. and Ullah, A. (1998). The Econometric Analysis of Models with Risk Terms. *Journal of Applied Econometrics*, 3 (2), 87-105.

Pagan, A. and Ullah, A. (1999). Nonparametric Econometrics. *Cambridge University Press, Cambridge*.

Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the Regression-Discontinuity Approach to Multiple Assignment Variables. *Journal of Econometrics*, 161, 203-207.

Parzen, M. (1962). On Estimation of a Probability Density Function and Mode. *Philos. Trans. Roy. Soc. London Ser. A*, 186, 343-414.

Politis, D. N. and Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, 89, 1303-1313.

Psacharopoulos, G. and Patrinos, H. A. (2018). Returns to Investment in Education. *World Bank Group Policy Research Working Paper 8402*.

Qu, Z. and Yoon, J. (2019). Uniform Inference on Quantile Effects under Sharp Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 37 (4), 625-647.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56 (4), 931-954.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66 (5), 688.

Silverman, B. W. (1985). Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting (with Discussion). *Journal of Royal Statistical Society B*, 47, 1-52.

Silverman, B. (1986). Density Estimation for Statistics and Data Analysis. *Chapman & Hall*.

Smith, R. and Blundell, R. (1986). An Exogeneity Test for A Simultaneous Equation Tobit Model with An Application to Labor Supply. *Econometrica*, 54 (3), 679-685.

Su, L. and Ullah, A. (2008). Local Polynomial Estimation of Nonparametric Simultaneous Equations Models. *Journal of Econometrics*, 144, 193-218.

Su, L., Ullah, A., Mishra, S., and Wang, Y. (2012). Nonparametric and Semiparametric Volatility Models: Specification, Estimation, and Testing. *In Handbook of Volatility Models and Their Applications (eds L. Bauwens, C. Hafner and S. Laurent)*.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-Discontinuity Analysis: An Alternative to the Ex-post Facto Experiment. *Journal of Educational Psychology*, 51, 309-317.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

Ullah, A., Wang, T., and Yao, W. (2021). Modal Regression for Fixed Effects Panel Data. *Empirical Economics*, 60, 261-308.

Ullah, A., Wang, T., and Yao, W. (2022). Nonlinear Modal Regression for Dependent Data with Application for Predicting COVID-19. *Journal of the Royal Statistical Society Series A*, forthcoming.

Van Der Klaauw, W. (2008). Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics. *Labour*, 22, 219-245.

Wang, L., Li, R., and Tsai, C. L. (2007). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, 94 (3), 553-568.

Wang, Y. and Tang, M. (2016). Local M-Estimation for Conditional Variance Function with Dependent Data. *Rocky Mountain Journal of Mathematics*, 46 (1), 333-356.

Wu, Y. and Liu, Y. (2019). Variable Selection in Quantile Regression. *Statistica Sinica*, 19 (2), 801-817.

Xu, K. and Philips, P. C. B. (2011). Tilted Nonparametric Estimation of Volatility Functions With Empirical Applications. *Journal of Business & Economic Statistics*, 29 (4), 518-528.

Yao, Q. and Tong, H. (1994). Quantifying the Influence of Initial Values on Nonlinear Prediction. *Journal of Royal Statistical Society B*, 56, 701-725.

Yao, W. (2013). A Note on EM Algorithm for Mixture Models. *Statistics and Probability Letters*, 83, 519-526.

Yao, W. and Li, L. (2014). A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, 41, 656-671.

Yao, W., Lindsay, B. G., and Li, R. (2012). Local Modal Regression. *Journal of Nonparametric Statistics*, 24 (3), 647-663.

Yao, W. and Xiang, S. (2016). Nonparametric and Varying Coefficient Modal Regression. *arXiv:1602.06609*.

Yogo, M. (2004). Estimating the Elasticity of Intertemporal Substitution When Instruments are Weak. *Review of Economics and Statistics*, 86, 797-810

Yu, K. and Jones, M. C. (2004). Likelihood-Based Local Linear Estimation of the Conditional Variance Function. *Journal of the American Statistical Association*, 99 (465), 139-144.

Zhang, H. and Lu, W. (2007). Adaptive Lasso for Cox's Proportional Hazards Model. *Biometrika*, 94 (3), 691-703.

Zhang, T., Kato, K., and Ruppert, D. (2020). Pivotal Bootstrap for Quantile-Based Modal Regression. *arXiv:2006.00952v1*.

Zhang, R., Zhao, W., and Liu, J. (2013). Robust Estimation and Variable Selection for Semiparametric Partially Linear Varying Coefficient Model Based on Modal Regression. *Journal of Nonparametric Statistics*, 25 (2), 523-544.

Ziegelmann, F. A. (2002). Nonparametric Estimation of Volatility Functions: The Local Exponential Estimator. *Econometric Theory*, 18 (4), 985-991.

Ziegelmann, F. A. (2008). A Local Linear Least-Absolute-Deviations Estimator of Volatility. *Communications in Statistics-Simulation and Computation*, 81 (6), 707-728.

Zhou, H. and Huang, X. (2016). Nonparametric Modal Regression in the Presence of Measurement Error. *Electronic Journal of Statistics*, 10, 3579-3620.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101, 1418-1429.

Appendix A

Appendix for Chapter 2

A.1 Modal Asset Pricing Model

The consumption-based mean asset pricing model, widely regarded as one of the most significant models in the asset pricing literature, states that the conditional expected return on a risky asset should be proportional to its conditional consumption beta. However, it has long been recognized that several fundamental principles of expected utility as a risk preference measure are often broken in reality. We in this part develop a modal asset pricing model to further illustrate the applicability of the proposed modal regression with endogeneity and to enlarge the asset pricing literature. The model framework follows Giovannetti (2013), which utilized quantile maximization decision theory to the standard intertemporal problem of a consumer-investor agent, leading to quantile-based Euler equations that the agent must satisfy in equilibrium. We solve the standard intertemporal problem of a consumer-investor agent and consider a two-period economy with two assets, one risky and one risk-free. Assume that the value of the risky asset at $t+1$ is $X_{t+1} = P_{t+1} + D_{t+1}$, where P_{t+1} is the

price of the asset at $t+1$ and D_{t+1} is the value of some cash flow the investor received between t and $t+1$. Define the value of the risk-free asset at $t+1$ as X_{t+1}^f and the price at t as P_t^f. The quantities for these two assets at time t are θ and θ^f. Assume the agent's consumption at t is C_t and the initial wealth is W_t. Then, under the modal maximization decision and time-separability, the agent maximizes

$$\max_{\theta,\theta^f} Mode_t(U(C_t) + \beta U(C_{t+1})), \tag{S.1}$$

where β is the time discount factor, $U(\cdot)$ is a strictly increasing utility function, and $Mode_t(\cdot)$ represents the mode of the conditional distribution of a random variable conditional on the information set available at t. The budget constraint is

$$\begin{aligned} C_t &= W_t - P_t\theta - P_t^f\theta^f, \\ C_{t+1} &= X_{t+1}\theta + X_{t+1}^f\theta^f. \end{aligned} \tag{S.2}$$

As shown in the Modal Euler Equation example, for a strict increasing function $U(\cdot)$, we have $Mode(U(X)) = U(Mode(X))$. Because the model is set for two periods, the objective function can be expressed as

$$Mode_t(U(C_t) + \beta U(C_{t+1})) = U(C_t) + \beta U(Mode_t(C_{t+1})). \tag{S.3}$$

Replacing C_t and C_{t+1} with the corresponding budget constraint and taking the first order conditions with respect to θ and θ^f, we obtain

$$P_t = \beta \frac{U^{(1)}\left(Mode_t\left(C_{t+1}\right)\right)}{U^{(1)}\left(C_t\right)} Mode_t\left(X_{t+1}\right), \tag{S.4}$$

$$P_t^f = \beta \frac{U^{(1)}\left(Mode_t\left(C_{t+1}\right)\right)}{U^{(1)}\left(C_t\right)} X_{t+1}^f. \tag{S.5}$$

It is worth noting that the above two equations adhere to the Law of One Price and the no-arbitrage condition, that is, price is linear. In accordance with Giovannetti (2013), define $\Theta_t = (\theta_t, \theta_t^f)$ as a portfolio formed at t with a price P_t^{Θ}. Then, we have

$$\begin{aligned} P_t^{\Theta} &= \beta \frac{U^{(1)}\left(Mode_t\left(C_{t+1}\right)\right)}{U^{(1)}\left(C_t\right)} Mode_t(X_{t+1}\theta_t + X_{t+1}^f \theta_t^f) \\ &= \beta \frac{U^{(1)}\left(Mode_t\left(C_{t+1}\right)\right)}{U^{(1)}\left(C_t\right)} Mode_t(X_{t+1}\theta_t) + \beta \frac{U^{(1)}\left(Mode_t\left(C_{t+1}\right)\right)}{U^{(1)}\left(C_t\right)} X_{t+1}^f \theta_t^f \\ &= P_t \theta_t + P_t^f \theta_t^f . \end{aligned} \tag{S.6}$$

If there is an arbitrage opportunity occurring, it implies that $P_t\theta_t + P_t^f\theta_t^f = 0$ and $X_{t+1}\theta_t + X_{t+1}^f\theta_t^f \geq 0$. Rearranging the aforementioned equations yields $X_{t+1}^f\theta_t^f = -\theta_t Mode_t(X_{t+1})$. As a consequence, a necessary and sufficient condition for arbitrage is $\theta_t(X_{t+1} - Mode_t(X_{t+1})) \geq 0$. Therefore, in order to rule out arbitrage, we need to impose the conditions that

$$Mode_t(X_{t+1}) \in \left(\min\left\{\text{supp}\left(X_{t+1}\right)\right\}, \max\left\{\text{supp}\left(X_{t+1}\right)\right\}\right). \tag{S.7}$$

If we let $U(C) = C^{1-\gamma}/(1-\gamma)$, the modal Euler equations are given by

$$\begin{aligned} P_t &= \beta \left(Mode_t \left(\frac{C_{t+1}}{C_t} \right) \right)^{-\gamma} Mode_t\left(X_{t+1}\right), \\ P_t^f &= \beta \left(Mode_t \left(\frac{C_{t+1}}{C_t} \right) \right)^{-\gamma} X_{t+1}^f . \end{aligned} \tag{S.8}$$

In equilibrium, we can obtain $Mode(X_{t+1} \mid \Omega_t) = X_{t+1}^f$. Rearranging equation, we achieve

$$Mode\left(\frac{C_{t+1}}{C_t} \mid \Omega_t \right) = (\beta X_{t+1}^f)^{1/\gamma}. \tag{S.9}$$

Such a modal Euler equation must be satisfied in equilibrium, suggesting a set of population orthogonality conditions. Using the mode's invariance property and the fact that the logarithm function is monotonically increasing, we get

$$Mode_t(ln(X_{t+1}^f)) = -ln(\beta) + \gamma Mode\left(ln\left(\frac{C_{t+1}}{C_t} \right) \mid \Omega_t \right), \tag{S.10}$$

where $1/\gamma$ is the EIS parameter defining the degree of substitutability-complementarity between consumption today and the certainty equivalent of consumption tomorrow. For the numerical estimation, we can follow the same procedure as in the Modal Euler Equation example to apply the proposed estimation method in this paper.

Remark S.1. (**Constant Economic Uncertainty**) Define $r_{t+1} = ln(X_{t+1}/P_t)$, followed by $r_{t+1} = \mu_r + u_{t+1}$, where $u_{t+1} \sim i.i.d.N(0, \sigma_r^2)$. Assume that the consumption growth rate $g_{t+1} = ln(C_{t+1}/C_t)$ is followed by $g_{t+1} = \mu_c + \eta_{t+1}$, where $\eta_{t+1} \sim i.i.d.N(0, \sigma_c^2)$. Then, we have[1]

$$r_{t+1} = -ln(\beta) + \gamma\mu_c - \gamma\sigma_c^2 + \sigma_r^2 + u_{t+1},$$
$$r_{t+1}^f = -ln(\beta) + \gamma\mu_c - \gamma\sigma_c^2,$$
$$E_t(r_{t+1} - r_{t+1}^f) = \sigma_r^2,$$

where r_{t+1}^f refers to the risk-free asset return and $E_t(\cdot)$ represents the expectation conditional on the information set available at t. It can be seen that the risk-free rate is linear in expected consumption growth, with the slope equal to the inverse of the EIS. The higher the desire for consumption smoothing, the higher the risk-free rate. Furthermore, the higher the rate at which the agent discounts future utility, the higher the risk-free rate required by the agent in order to save across time. Different from the results of the mean model, the higher variability of consumption growth may have a less negative effect on the level of the risk-free rate under the modal model (the third component for r_{t+1}^f is $-\gamma^2\sigma_c^2/2$ with expected utility). It has been observed that the risk premium does not depend on the covariance between consumption and stock returns for the modal model, but on the standard deviation of the stock return. Also, the risk premium is always positive.

[1]If $ln(x) \sim N(\mu, \sigma^2)$, then $Mode(x) = exp(\mu - \sigma^2)$. According to the model settings, we know $Mode_t(C_{t+1}/C_t) = exp(\mu_c - \sigma_c^2)$ and $Mode_t(X_{t+1}/P_t) = exp(\mu_r - \sigma_r^2)$. Under expected utility, $E_t(r_{t+1} - r_{t+1}^f) = -\sigma_r^2/2 + \gamma\sigma_{cr}$, where $Cov(\eta_{t+1}, u_{t+1}) = \sigma_{cr}$.

A.2 Return to Schooling

There has been a large number of empirical research in labor economics focusing on the causal links between education and labor market success (return to schooling); see the relevant literature summarized in Card (2001) and Psacharopoulos and Patrinos (2018). However, all research associated with the return to schooling is conducted based on the mean or quantile regression, regardless of whether endogeneity is considered. In this part, we further investigate the finite sample performance of the proposed procedure by researching the estimation of return to schooling presented in Card (1995) to enhance our understanding of the education-return relationship, which can be regarded as a contribution to the literature on education. The data are from the National Longitudinal Survey of Young Men (NLSYM), which began in 1966 with 5525 men aged 14-24 and continued with follow-up surveys through 1981. It contains many variables that we can use directly, such as the education background of parents, dummies for family structure, and dummies for living near a college. The motivation for a control function approach stems from the endogeneity of schooling, which could be due to ability bias. Furthermore, it is quite plausible that growing up near a four-year college is independent of ability factors. For the purposes of illustration, we in this part focus on the results of Table 3 Column 5A in Card (1995) and adopt the same dummy variable for whether someone grew up near a four-year college as an instrumental variable for education.

The total number of data used in this paper is 3010. The dependent variable Y is log wages and the endogenous variable X is years of schooling (ed76). The instrumental variable Z_2 is living near a four-year college. The number of exogenous variables Z_1 is

14, which includes a linear experience term (exp76), a quadratic function of experience (exp762), a race indicator (black), dummies for residence in the south (reg76r) and in a metropolitan area in 1976 (smsa76r), and indicators for region of residence in 1966 (reg661-reg668) and for residence in a metropolitan area in 1966 (smsa66r). Although we lack a formal test to verify the conditional mode independence restriction, we argue that it is quite plausible that living near a four-year college (Z_2) is independent of any ability factors (U_i) that may affect the return to schooling; see Card (2001). The descriptive statistics for the sample as well as the arguments for the validity of the instrument can be found in Card (1995). We thus leave them out for brevity.

Table A.1: Estimates of Return to Schooling

Variables	Two-Step Modal	Naive Linear Modal	Mean-2SLS	Quantile (0.3)	Quantile (0.5)	Quantile (0.7)
ed76	0.1331*** (0.0010)	0.0772*** (0.0005)	0.1315** (0.0548)	0.1652*** (0.0561)	0.1351*** (0.0790)	0.0945** (0.0391)
black	-0.1471*** (0.0009)	-0.2132*** (0.0012)	-0.1468*** (0.0538)	-0.1346** (0.0595)	-0.1431** (0.0660)	-0.1568*** (0.0387)
smsa76r	0.1133*** (0.0006)	0.1786*** (0.0014)	0.1118*** (0.0316)	0.0789* (0.0442)	0.1180*** (0.0428)	0.1211*** (0.0360)
reg76r	-0.1440*** (0.0004)	-0.1758*** (0.0017)	-0.1447*** (0.0272)	-0.1573*** (0.0418)	-0.1330*** (0.0308)	-0.1197*** (0.0292)
reg661	-0.1065*** (0.0007)	-0.0177*** (0.0040)	-0.1078*** (0.0417)	-0.0601 (0.0889)	-0.0638 (0.0542)	-0.1094*** (0.0359)
reg662	-0.0059*** (0.0006)	-0.0817*** (0.0027)	-0.0070 (0.0328)	-0.0046 (0.0491)	-0.0164 (0.0491)	-0.0071 (0.0381)
reg663	0.0418*** (0.0006)	0.0427*** (0.0026)	0.0404 (0.0317)	0.0516 (0.0457)	0.0281 (0.0431)	0.0285 (0.0387)
reg664	-0.0574*** (0.0007)	-0.0835*** (0.0046)	-0.0579 (0.0375)	-0.0932 (0.0566)	-0.0810* (0.0445)	-0.0485 (0.0545)
reg665	0.0396*** (0.0008)	0.0335*** (0.0030)	0.0385 (0.0468)	0.0581 (0.0769)	0.0165 (0.0677)	-0.0008 (0.0545)
reg666	0.0560*** (0.0010)	0.0187*** (0.0045)	0.0551 (0.0525)	0.1040 (0.0782)	0.0348 (0.0669)	0.0019 (0.0524)
reg667	0.0285*** (0.0009)	0.0074** (0.0031)	0.0268 (0.0487)	0.0565 (0.0822)	0.0293 (0.0866)	-0.0048 (0.0560)
reg668	-0.1894*** (0.0009))	-0.0877*** (0.0040)	-0.1909*** (0.0506)	-0.2225*** (0.0635)	-0.1791*** (0.0618)	-0.1548*** (0.0491)
smsa66r	0.0189*** (0.0004)	0.0830*** (0.0013)	0.0185 (0.0216)	0.0211 (0.0253)	0.0207 (0.0238)	0.0395 (0.0264)
exp76	0.1089*** (0.0005)	0.0868*** (0.0008)	0.1083*** (0.0236)	0.1200*** (0.0240)	0.1083*** (0.0385)	0.0879*** (0.0189)
exp762	-0.0023*** (0.0006)	-0.0029*** (0.00003)	-0.0023*** (0.0003)	-0.0022*** (0.0004)	-0.0024*** (0.0005)	-0.0021*** (0.0004)

We then use the proposed method in this paper to estimate the return to schooling, which is defined as the derivative of the primary wage equation with respect to education. The bandwidths are chosen based on the selection procedure introduced in Section 2.4

with $1.6MADn^{-\lambda_{h_j}}$ rather than the grid search method. For comparison purposes, we also report the results of the naive linear modal regression (without endogeneity adjustment), the mean regression with 2SLS estimation, and the smoothed instrumental variable quantile regression (Kaplan, 2020). To obtain the standard errors of modal estimates, we follow Ullah et al. (2021) to apply the bootstrap method with 200 replications. The estimation results are summarized in Table A.1, with the values in brackets representing standard errors. We use *** to denote the 1% significance level, ** to indicate the 5% significance level, and * to represent the 10% significance level.

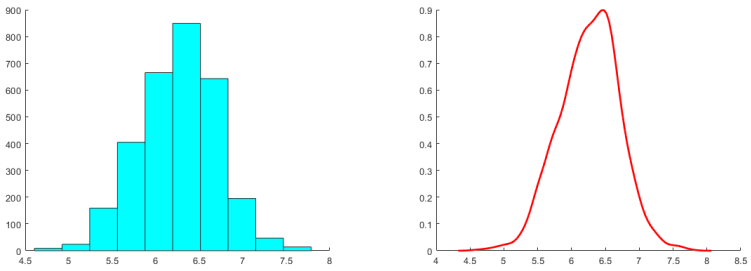

Figure A.1: Histogram and Kernel Density for Dependent Variable

It can be observed that all three methods considered produce estimates with the same signs for the three major variables (ed76, exp76, and exp762). All estimates, whether based on the mean, mode, or quantile, show that education level is positively associated with wage level. The quantile instrumental variable estimates of the return to schooling exhibit considerable heterogeneity, ranging from 0.0945 to 0.1652. The magnitude of the estimate of the return to schooling obtained from the naive linear modal regression is quietly different from that obtained from the proposed modal regression. Particularly, the naive modal estimate underestimates the return to schooling as a result of the presence of endogeneity,

which indicates that the naive estimate is inconsistent. This fact is further confirmed by the result of the mean or median (0.5 quantile) regression with instrumental variables, which is significantly different from the naive modal estimate. Compared to the traditional mean or median regression with instrumental variables, the estimate of the return to schooling from modal regression with control function is slightly higher, although the difference is not substantial.[2] Particularly, it is 0.0016 higher than the mean estimate and 0.0020 lower than the median estimate, suggesting that the effect of education is considerably stronger for the majority of individuals. This provides important implications for policy makers from the perspective of the "most likely" effect. Finally, most of the bootstrapped standard errors of the modal estimates are smaller than those of the mean estimates, which implies that the proposed modal regression method is more efficient.

A.3 Additional Numerical Results

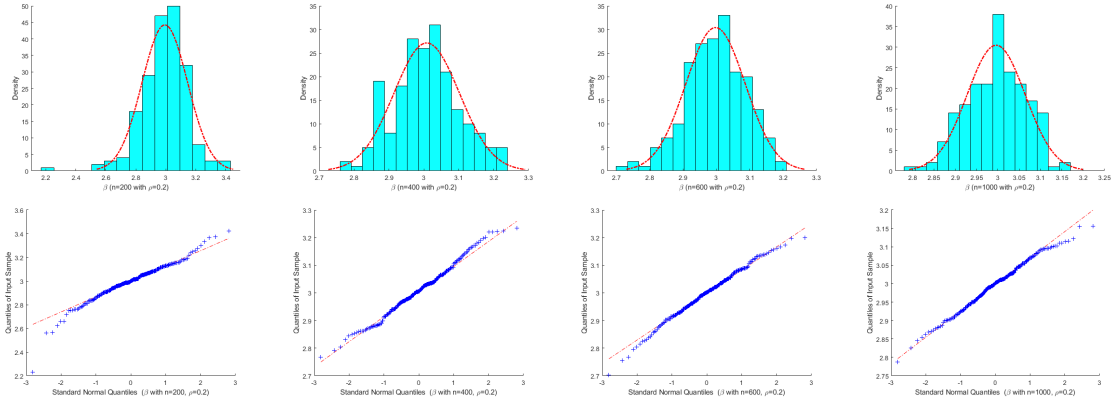

Figure A.2: Histograms and QQ Plots for Estimates (β with $\rho = 0.2$)—DGP 2

[2]We attribute this to the almost symmetrically distributed response variable (Figure A.1).

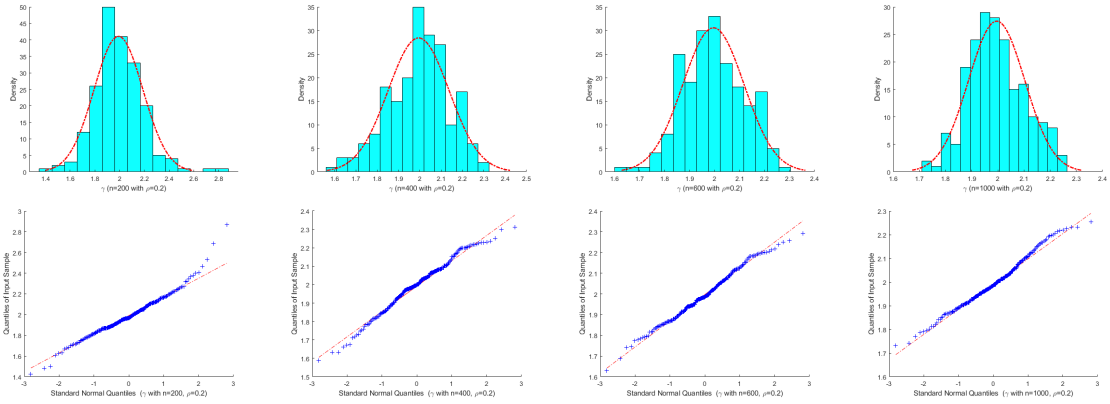

Figure A.3: Histograms and QQ Plots for Estimates (γ with $\rho = 0.2$)—DGP 2

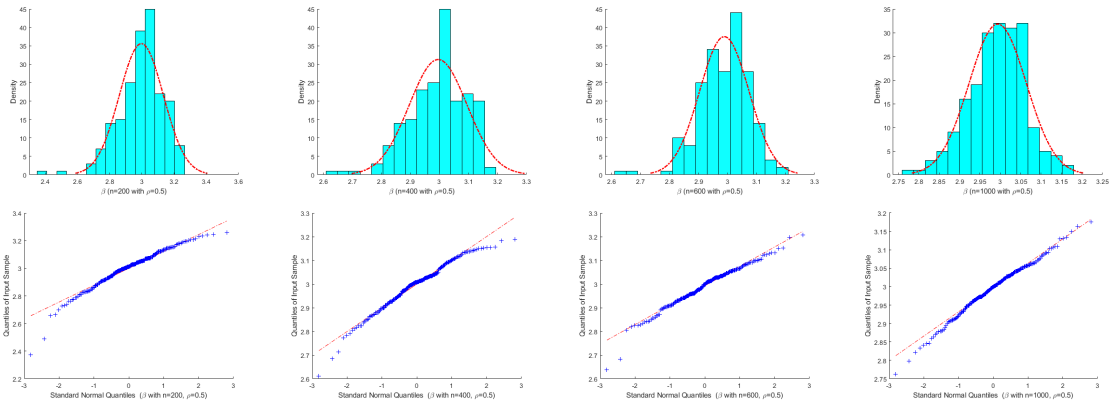

Figure A.4: Histograms and QQ Plots for Estimates (β with $\rho = 0.5$)—DGP 2

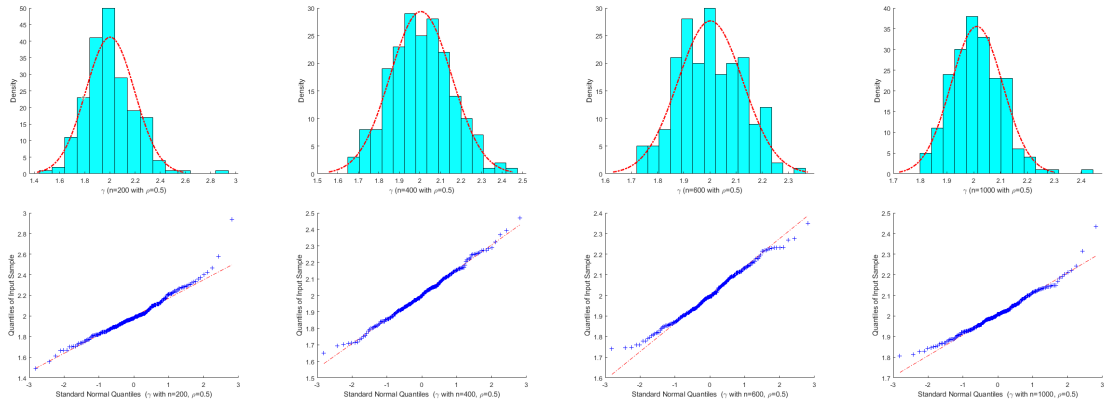

Figure A.5: Histograms and QQ Plots for Estimates (γ with $\rho = 0.5$)—DGP 2

Table A.2: Regression with Endogeneity of Log GDP Per Capita (Additional Controls)

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
				Two-Step Modal					
Avexpr	1.0389	0.9645	1.0643	1.0693	1.0101	0.9675	0.8945	0.8382	0.9073
	(0.0133)	(0.0163)	(0.0308)	(0.0366)	(0.0113)	(0.0151)	(0.0100)	(0.0125)	(0.0139)
Lat		0.7698		-0.0078		0.5046		0.5986	0.4081
		(0.0758)		(0.2130)		(0.0852)		(0.0760)	(0.0722)
British dummy	-0.3473	-0.3715							
	(0.0211)	(0.0231)							
French dummy	-0.0841	-0.1509							0.2652
	(0.0249)	(0.0279)							(0.0257)
French legal dummy					0.4336	0.4249			-0.0102
					(0.0192)	(0.0201)			(0.0273)
				Naive Linear Modal					
Avexpr	0.5276	0.4647	0.6044	0.5475	0.5624	0.5078	0.5332	0.4724	0.4717
	(0.0084)	(0.0082)	(0.0264)	(0.0296)	(0.0085)	(0.0090)	(0.0071)	(0.0070)	(0.0071)
Lat		1.8410		1.1456		1.5098		1.5973	1.5937
		(0.0977)		(0.3056)		(0.0827)		(0.0729)	(0.0765)
British dummy	-0.3084	-0.3702							
	(0.0257)	(0.0269)							
French dummy	-0.3802	-0.4611							0.0186
	(0.0315)	(0.0314)							(0.0266)
French legal dummy					0.3649	0.3454			-0.0223
					(0.0249)	(0.0250)			(0.0295)
				Mean-2SLS					
Avexpr	1.0779	1.1552	1.0662	1.2118	1.0800	1.1811	0.9174	1.0062	1.2122
	(0.2176)	(0.3372)	(0.2443)	(0.5164)	(0.1911)	(0.2910)	(0.1467)	(0.2517)	(0.3949)
Lat		-0.7512		-2.9863		-1.1258		-0.9378	-1.7936
		(1.3351)		(3.2136)		(1.5597)		(1.5034)	(2.1328)
British dummy	-0.7777	-0.7955							
	(0.3543)	(0.3930)							
French dummy	-0.1170	-0.0578							0.3960
	(0.3548)	(0.4188)							(0.4953)
French legal dummy					0.8865	0.9624			0.2942
					(0.3242)	(0.3935)			(0.5187)

Note: Model 1 and Model 2 are for base samples. Model 3 and Model 4 are for base samples wit British colonies only. Model 5 and Model 6 are for base samples with control for French legal origin. Model 7, Model 8, and Model 9 are for base samples with additional religion variables which are fractions of the populations that are Catholic, Muslim, and of other religions.

A.4 Monte Carlo Experiment

To illustrate the proposed modal regression for robustly estimating coefficients, we generate data according to the following model

$$\begin{cases} Y_i = X_i\beta + Z_{1,i}\gamma + U_i, \\ U_i = V_i + cos(V_i) + \tilde{U}_i, \quad i = 1, \cdots, n, \\ X_i = \alpha + Z_{1,i}\pi_1 + Z_{2,i}\pi_2 + V_i, \end{cases}$$

where $Z_{1,i}$ and $Z_{2,i}$ are drawn from the following multivariate normal distribution

$$\begin{pmatrix} Z_{1,i} \\ Z_{2,i} \end{pmatrix} \sim \text{i.i.d.}\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

The following three different error distributions are considered: (1) $\tilde{U}_i \sim$ mixture Laplace $0.8L_p(0,1) + 0.2L_p(0,5)$ and $V_i \sim t$ with three degrees of freedom; (2) $\tilde{U}_i \sim$ mixture normal $0.9N(0,1) + 0.1N(0,9)$ and $V_i \sim$ mixture Laplace $0.8L_p(0,1) + 0.2L_p(0,5)$; (3) $\tilde{U}_i \sim$ t with three degrees of freedom and $V_i \sim$ mixture normal $0.9N(0,1) + 0.1N(0,9)$. Without sacrificing generality, we follow DGP 1 to only consider the case of one instrumental variable $Z_{2,i}$ and denote the set of all instruments as $Z_i = [Z_{1,i}\ Z_{2,i}]^T$. The parameter values are set as $(\beta, \gamma, \alpha, \pi_1, \pi_2) = (1.5, 2, 2, 2, 2)$. We then have $E(V_i \mid Z_i) = Mode(V_i \mid Z_i) = 0$, $E(\tilde{U}_i \mid Z_i) = Mode(\tilde{U}_i \mid Z_i) = 0$, as well as the control function $E(U_i \mid V_i, Z_i) = Mode(U_i \mid V_i, Z_i) = V_i + cos(V_i)$.

The sample sizes n are set to be 200, 400, 600, and 1000. A total of 200 simulation replications are conducted for each model setting. To assess the robustness of the proposed modal-based estimation, we compare its performance to that of mean estimation (control function). For simplicity, we use the rule of thumb ($1.06\ \hat{\sigma}(\hat{V}_i)n^{-1/5}$) for bandwidth choice

in mean estimation. The estimates, standard errors, and MSEs are reported in Table A.3, which shows that the performance of modal-based estimation is better than mean estimation for non-normal distributions, with significant gains in efficiency and robustness. In addition, given error distributions, both estimation methods become better as the sample size n increases.

To demonstrate the asymptotic behavior of the proposed modal-based estimators, we present the histograms and QQ plots for the estimates of β and γ (Figures A.6-A.11). The plots all support our theoretical results of the asymptotic distribution of the proposed estimators. With the sample size increasing, the points in QQ align more closely to a straight line, indicating higher accuracy in the asymptotic standard approximation for these two estimators.

Table A.3: Results of Simulations

	Modal-Based Estimation				Mean Estimation			
Sample Size	β (SE)	MSE(β)	γ (SE)	MSE(γ)	β (SE)	MSE(β)	γ (SE)	MSE(γ)
				$V_i \sim t(3)$				
n=200	1.4901 (0.1192)	0.0142	2.0404 (0.3239)	0.1060	1.4756 (0.1783)	0.0322	2.0499 (0.4800)	0.2317
n=400	1.4931 (0.0805)	0.0065	2.0256 (0.2303)	0.0534	1.4959 (0.1244)	0.0154	2.0229 (0.3439)	0.1182
n=600	1.5036 (0.0684)	0.0047	1.9918 (0.1827)	0.0333	1.5004 (0.0970)	0.0094	1.9742 (0.2673)	0.0718
n=1000	1.4948 (0.0516)	0.0027	1.9977 (0.1434)	0.0205	1.4914 (0.0827)	0.0069	2.0082 (0.2383)	0.0565
				$V_i \sim 0.8L_p(0,1) + 0.2L_p(0,5)$				
n=200	1.5042 (0.0849)	0.0072	1.9930 (0.2428)	0.0587	1.5007 (0.1158)	0.0133	2.0027 (0.3069)	0.0937
n=400	1.5016 (0.0589)	0.0035	1.9900 (0.1671)	0.0279	1.5022 (0.0746)	0.0055	1.9677 (0.2195)	0.0490
n=600	1.5019 (0.0497)	0.0025	2.0035 (0.1302)	0.0169	1.5123 (0.0659)	0.0045	1.9683 (0.1877)	0.0361
n=1000	1.4981 (0.0349)	0.0012	1.9966 (0.0950)	0.0090	1.4958 (0.0470)	0.0022	1.9986 (0.1287)	0.0165
				$V_i \sim 0.9N(0,1) + 0.1N(0,9)$				
n=200	1.4989 (0.1324)	0.0174	2.0303 (0.3465)	0.1204	1.4971 (0.1712)	0.0292	2.0309 (0.4331)	0.1876
n=400	1.4954 (0.0909)	0.0082	2.0164 (0.2488)	0.0619	1.5011 (0.1063)	0.0112	2.0213 (0.2851)	0.0813
n=600	1.4982 (0.0755)	0.0057	2.0014 (0.2073)	0.0428	1.4987 (0.0833)	0.0069	1.9989 (0.2194)	0.0479
n=1000	1.4969 (0.0400)	0.0016	2.0085 (0.1136)	0.0129	1.4975 (0.0518)	0.0027	2.0202 (0.1479)	0.0222

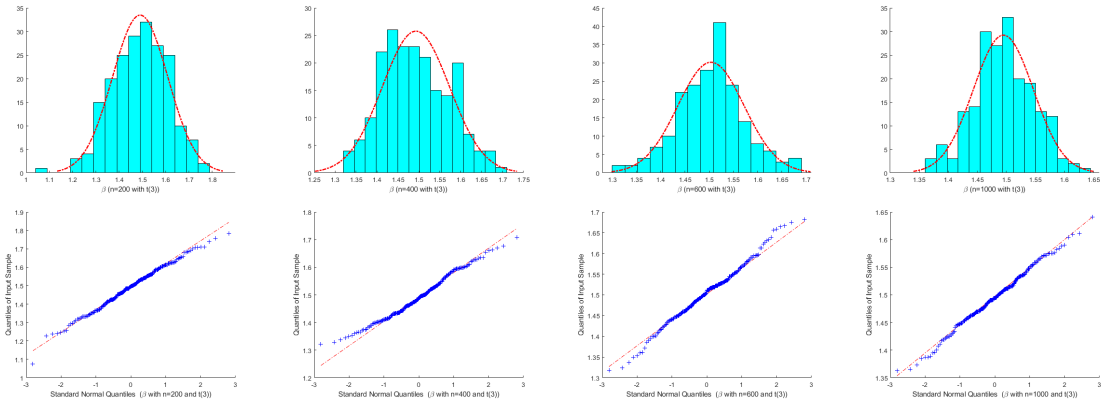

Figure A.6: Histograms and QQ Plots for Estimates (β with $V_i \sim t(3)$)

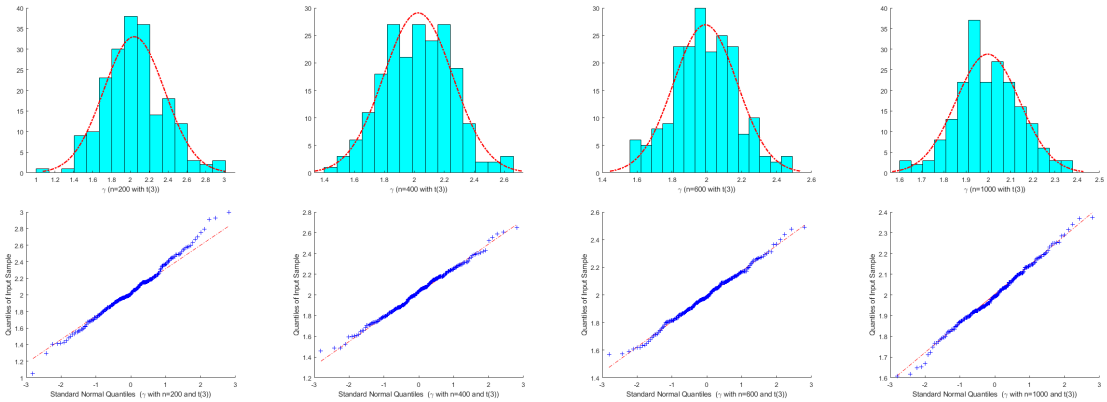

Figure A.7: Histograms and QQ Plots for Estimates (γ with $V_i \sim t(3)$)

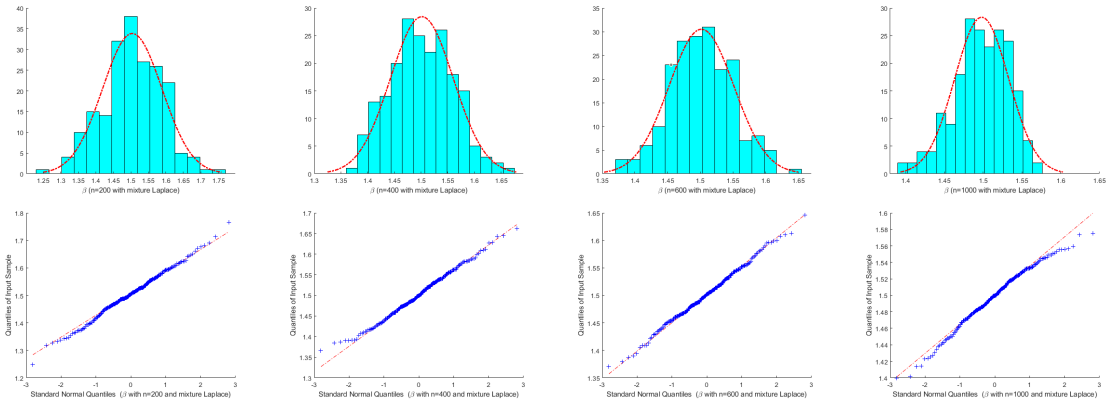

Figure A.8: Histograms and QQ Plots for Estimates (β with $V_i \sim 0.8L_p(0,1) + 0.2L_p(0,5)$)

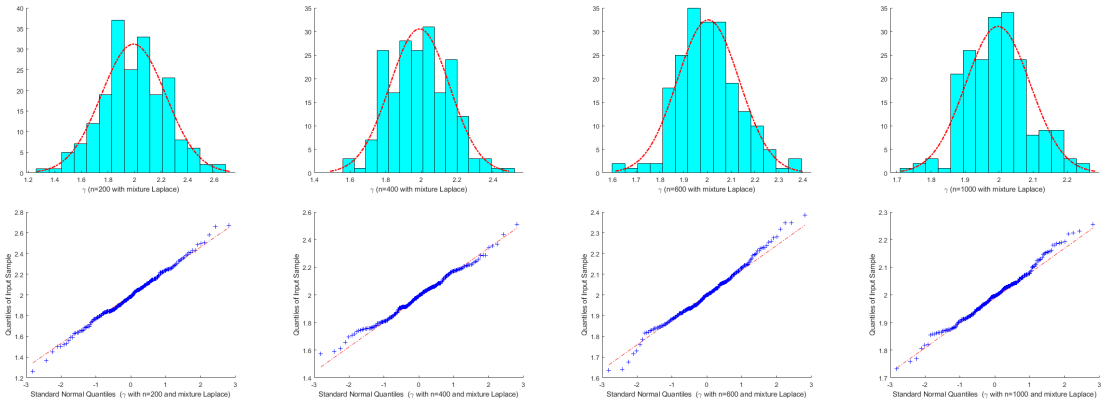

Figure A.9: Histograms and QQ Plots for Estimates (γ with $V_i \sim 0.8L_p(0,1) + 0.2L_p(0,5)$)

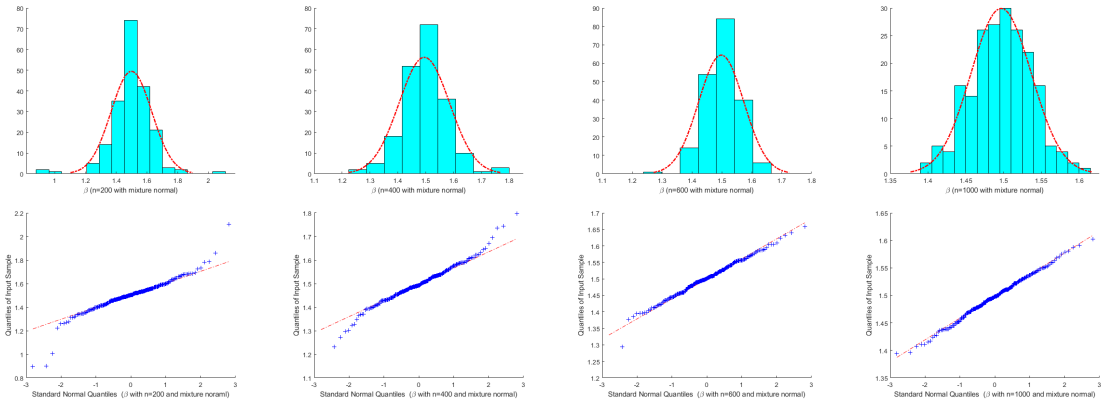

Figure A.10: Histograms and QQ Plots for Estimates (β with $V_i \sim 0.8L_p(0,1) + 0.2L_p(0,5)$)

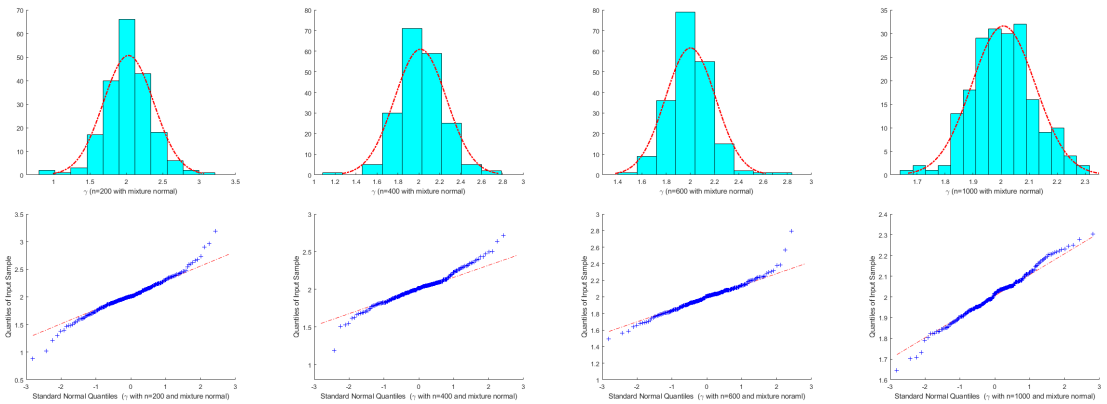

Figure A.11: Histograms and QQ Plots for Estimates (γ with $V_i \sim 0.8L_p(0,1) + 0.2L_p(0,5)$)

A.5 Technical Proofs

Proof of Theorem 2.4.3

Recall that

$$\frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{Y_i - X_i\beta - Z_{1,i}^T\gamma - m(V_i) - (m(\hat{V}_i) - m(V_i))}{h_1}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right), \tag{A.1}$$

where $m(\hat{V}_i) - m(V_i) = m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)$ and $\bar{V}_i$ is between $\hat{V}_i$ and V_i. From Theorem 2.4.1, we know $|\hat{V}_i - V_i| = O_p((nh^3)^{-1/2} + h^2) = O_p(n^{-2/7})$ with the MSE-optimal bandwidth. Define $\theta = (\beta, \gamma^T, m(v), h_2m^{(1)}(v))^T$, $X_i^* = [X_i, Z_{1,i}^T, 1, h_2^{-1}(V_i - v)]^T$, and $R(V_i) = \sum_{j=2}(m^{(j)}(v)/j!)(V_i - v)^j$, we have

$$Q_n(\theta) = \frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{Y_i - X_i^{*T}\theta + R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)}{h_1}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right). \tag{A.2}$$

Define $\delta_n = h_1^2 + h_2^2 + \sqrt{(nh_1^3h_2)^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_0\right)\right\} \geq 1 - \eta, \tag{A.3}$$

where $\boldsymbol{\theta}_0$ is the true parameter and $\|.\|$ represents the Euclidean distance. (A.3) implies that with probability tending to 1, there is a local maximum in the ball $\{\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu} : \|\mu\| \leq c\}$. Using Taylor expansion, it follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\theta}_0\right) \\
=&\frac{1}{nh_1h_2}\sum_{i=1}^{n}\left[\phi\left(\frac{R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i) + \epsilon_i - \delta_n\mu^T X_i^*}{h_1}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\right. \\
&\left. - \phi\left(\frac{R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i) + \epsilon_i}{h_1}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{nh_1h_2}\sum_{i=1}^{n}\Bigg[-\phi^{(1)}\left(\frac{R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i}{h_1}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\\
&+\frac{1}{2}\phi^{(2)}\left(\frac{R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i}{h_1}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_1}\right)^2K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\\
&-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{h_1}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_1}\right)^3K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\Bigg]\\
=&I_1+I_2+I_3,
\end{aligned}
\tag{A.4}
$$

where ϵ_i^* is between $R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i$ and $R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i-\delta_n\mu^TX_i^*$. Based on the result $T_n=E\left(T_n\right)+O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is $I_1=\frac{-1}{nh_1h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i}{h_1}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)$, by Taylor expansion, we can rewrite it as

$$
\begin{aligned}
E(I_1)=&\frac{-\delta_n}{h_1h_2}E\left(\phi^{(1)}\left(\frac{R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_1}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)\\
=&\frac{-\delta_n}{h_1h_2}E\Bigg(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_1}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\\
&+\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))\boldsymbol{\mu}^TX_i^*}{h_1^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\\
&+\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{**}}{h_1}\right)\frac{(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))^2\boldsymbol{\mu}^TX_i^*}{h_1^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\Bigg)\\
=&I_{11}+I_{12}+I_{13},
\end{aligned}
\tag{A.5}
$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i+R(X_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)$. Notice that as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{13}, we instead use ϵ_i directly. By some direct calculations for each part, we can get

$$
I_{11}=\frac{-\delta_n}{h_1h_2}E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_1}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)
$$

$$
\begin{aligned}
=&\frac{-\delta_n}{h_1 h_2}\iiint \phi^{(1)}\left(\frac{\epsilon}{h_1}\right)\frac{\boldsymbol{\mu}^T X^*}{h_1} g_\epsilon(\epsilon \mid X^*) K\left(\frac{V-v+\hat{V}-V}{h_2}\right) f_V(V) dV d\epsilon dF(X^*) \\
=&\frac{\delta_n}{h_1}\iiint \phi(\tau)\tau \boldsymbol{\mu}^T X^* g_\epsilon(\tau h_1 \mid X^*) K(w+w') f_V(wh_2+v) dw d\tau dF(X^*) \\
=&O_p(\delta_n c h_1^2), \qquad \text{(A.6)}
\end{aligned}
$$

where $w' = \frac{\hat{V}-V}{h_2} = O_p(h^2/h_2) = o_p(1)$ according to the definition of δ_n, and $K(w+w') = K(w) + K^{(1)}(\bar{w})w' = K(w) + o_p(1)$ where $\bar{w}$ is between w and $w+w'$.

$$
\begin{aligned}
I_{12} =& \frac{-\delta_n}{h_1 h_2} E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_1} K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\frac{R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)}{h_1}\right) \\
=&\frac{-\delta_n}{h_1 h_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\frac{\boldsymbol{\mu}^T X^*}{h_1} g_\epsilon(\epsilon \mid X^*) K\left(\frac{V-v+\hat{V}-V}{h_2}\right)\frac{R(V) - m^{(1)}(\bar{V})(\hat{V}-V)}{h_1} \\
& f_V(V) dV d\epsilon dF(X^*) \\
=&\frac{-\delta_n}{h_1}\iiint \phi(\tau)(\tau^2-1)\boldsymbol{\mu}^T X^* g_\epsilon(\tau h_1 \mid X^*) K\left(w+w'\right)\frac{R(V) - m^{(1)}(\bar{V})(\hat{V}-V)}{h_1} \\
& f_V(wh_2+v) dw d\tau dF(X^*) = O_p(\delta_n c h_2^2) \qquad \text{(A.7)}
\end{aligned}
$$

with the condition that $h/h_2 = o_p(1)$.

$$
\begin{aligned}
I_{13} \approx& \frac{-\delta_n}{h_1 h_2} E\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i))^2 \boldsymbol{\mu}^T X_i^*}{h_1^3} K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\right) \\
\leq& \frac{-\delta_n h_2^4}{2}\iiint \phi(\tau)(3\tau - \tau^3)\frac{(m^{(2)}(v))^2 \boldsymbol{\mu}^T X^*}{4h_1^3} g_\epsilon(\tau h_1 \mid X^*) K(w) w^4 f_V(wh_2+v) \\
& dw d\tau dF(X^*)\{1 + o_p(1)\} = o_p(\delta_n h_2^2). \qquad \text{(A.8)}
\end{aligned}
$$

Meanwhile, with the condition $h_2^2/h_1 \to 0$ held, we obtain

$$
\frac{\delta_n^2}{h_1^2 h_2^2} E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_1} K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\right)^2 = O_p(\delta_n^2 c^2 (h_1^3 h_2)^{-1}). \qquad \text{(A.9)}
$$

$$
\frac{\delta_n^2}{h_1^2 h_2^2} E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_1} K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right)\frac{R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)}{h_1}\right)^2
$$

$$\leq \frac{\delta_n^2 h_2^3}{h_1^5} \iiint \phi^2(\tau)(\tau^2-1)^2(\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1 \mid X^*) w^4 K^2(w) \frac{(m^{(2)}(v))^2}{4} f_V(wh_2+v)$$

$$dw d\tau dF(X^*)\{1+o_p(1)\} = o_p(\delta_n^2(h_1^3 h_2)^{-1}). \tag{A.10}$$

The above equations show that $I_1 = O_p(\delta_n c(h_1^2+h_2^2)) + O_p(\sqrt{\delta_n^2 c^2 (nh_1^3 h_2)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_2 = \frac{1}{nh_1h_2}\sum_{i=1}^n \left(\frac{1}{2}\phi^{(2)}\left(\frac{R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)+\epsilon_i}{h_1}\right)\left(\frac{\delta_n \mu^T X_i^*}{h_1}\right)^2 K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)$, we can rewrite it as

$$\begin{aligned}
E(I_2) =& \frac{\delta_n^2}{2h_2h_1} E\left(\phi^{(2)}\left(\frac{\epsilon_i + R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_1^2} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right) \\
=& \frac{\delta_n^2}{2h_2h_1} E\Bigg(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_1^2} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right) \\
&+ \phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))(\boldsymbol{\mu}^T X_i^*)^2}{h_1^3} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right) \\
&+ \frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{**}}{h_1}\right)\frac{(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))^2(\boldsymbol{\mu}^T X_t^*)^2}{h_1^4} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\Bigg) \\
=& I_{21} + I_{22} + I_{23},
\end{aligned} \tag{A.11}$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i + R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)$. Similarly, as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{23}, we instead use ϵ_i directly. By some calculations for each part, we can get

$$\begin{aligned}
I_{21} =& \frac{\delta_n^2}{2h_2h_1} E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_1^2} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right) \\
=& \frac{\delta_n^2}{2h_2h_1}\iiint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T X^*)^2}{h_1^2} g_\epsilon(\epsilon \mid X^*) K\left(\frac{V-v+\hat{V}-V}{h_2}\right) f_V(V) d\epsilon dV dF(X^*) \\
=& \frac{\delta_n^2}{2h_1^2}\iiint \phi(\tau)(\tau^2-1)(\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1 \mid X^*) K(w+w') f_V(wh_2+v) dw d\tau dF(X^*) \\
=& O_p((\delta_n c)^2).
\end{aligned} \tag{A.12}$$

$$I_{22} = \frac{\delta_n^2}{2h_2h_1} E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))(\boldsymbol{\mu}^T X_i^*)^2}{h_1^3} K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)$$

$$
\begin{aligned}
=&\frac{\delta_n^2}{2h_2h_1}\iiint \phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{(R(V)-m^{(1)}(\bar{V})(\hat{V}-V))(\boldsymbol{\mu}^T X^*)^2}{h_1^3}g_\epsilon(\epsilon \mid X^*)\\
&K\left(\frac{V-v+\hat{V}-V}{h_2}\right)f_V(V)d\epsilon dV dF(X^*)\\
\leq&\frac{\delta_n^2h_2^2}{2h_1^3}\iiint \phi(\tau)(3\tau-\tau^3)\frac{m^{(2)}(v)}{2}(\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1 \mid X^*)w^2K(w)f_V(wh_2+v)\\
&dwd\tau dF(X^*)\{1+o_p(1)\}=o_p((\delta_n c)^2). \qquad \text{(A.13)}
\end{aligned}
$$

Meanwhile, we can prove that $I_{23}=o_p((\delta_n c)^2)$ as well. Following the same steps in (i), we obtain the following result

$$
\begin{aligned}
&\frac{\delta_n^4}{4h_2^2h_1^2}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_1^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)^2\\
=&\frac{\delta_n^4}{4h_2^2h_1^2}\iiint \phi^{(2)2}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T X^*)^4}{h_1^4}g_\epsilon(\epsilon \mid X^*)K^2\left(\frac{V-v+\hat{V}-V}{h_2}\right)f_V(V)d\epsilon dV dF(X^*)\\
=&\frac{\delta_n^4}{4h_2h_1^2}\iiint \phi^2(\tau)(\tau^2-1)^2\frac{(\boldsymbol{\mu}^T X^*)^4}{h_1^4}g_\epsilon(\tau h_1 \mid X^*)K^2(w+w')f_V(wh_2+v)dwd\tau dF(X^*)\\
=&O_p((\delta_n c)^4(h_2h_1^5)^{-1}). \qquad \text{(A.14)}
\end{aligned}
$$

With the condition $nh_1^5h_2\to\infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) Following the same way, we can calculate the third part. As the order of ϵ_i^* is the same as the order of ϵ_i, by direct calculations, we have

$$
\begin{aligned}
&\frac{\delta_n^3}{6h_2h_1}E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\left(\boldsymbol{\mu}^T X_i^*\right)^3}{h_1^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)\\
=&\frac{\delta_n^3}{6h_2h_1}\iiint \phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu}^T X^*\right)^3}{h_1^3}g_\epsilon(\epsilon \mid X^*)K\left(\frac{V-v+\hat{V}-V}{h_2}\right)f_V(V)d\epsilon dV dF(X^*)\\
=&\frac{\delta_n^3}{6}\iiint \phi(\tau)(3\tau-\tau^3)\frac{\left(\boldsymbol{\mu}^T X^*\right)^3}{h_1^3}g_\epsilon(\tau h_1 \mid X^*)K(w+w')f_V(wh_2+v)dwd\tau dF(X^*)\\
=&O_p(\delta_n^3). \qquad \text{(A.15)}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\delta_n^6}{36h_2^2h_1^2}E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\left(\boldsymbol{\mu}^T X_i^*\right)^3}{h_1^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\right)^2\\
=&\frac{\delta_n^6}{36h_2^2h_1^2}\iiint\phi^{(3)2}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu}^T X^*\right)^6}{h_1^6}g_\epsilon(\epsilon\mid X^*)K^2\left(\frac{V-v+\hat{V}-V}{h_2}\right)f_V(V)d\epsilon dV dF(X^*)\\
=&\frac{\delta_n^6}{36h_2h_1}\iiint\phi^2(\tau)(3\tau-\tau^3)^2\frac{\left(\boldsymbol{\mu}^T X^*\right)^6}{h_1^6}g_\epsilon(\tau h_1\mid X^*)K^2(w+w')f_V(wh_2+v)dwd\tau dF(X^*)\\
=&O_p(\delta_n^6(h_2h_1^7)^{-1}).
\end{aligned}
\tag{A.16}
$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, thus $P\{\sup_{\|\boldsymbol{\mu}\|=c}Q_n\left(\boldsymbol{\theta}_0+\delta_n\boldsymbol{\mu}\right)$ $<Q_n\left(\boldsymbol{\theta}_0\right)\}\geq 1-\eta$ holds.

Proof of Theorem 2.4.4

Based on (A.2), the estimator $\tilde{\theta}$ must satisfy the following equation

$$
\begin{aligned}
&-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i+R(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*\\
&\qquad=0.
\end{aligned}
\tag{A.17}
$$

By taking Taylor expansion, we could obtain

$$
\begin{aligned}
&-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*+\\
&\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*(R(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))-\\
&\frac{1}{nh_1^4h_2}\sum_{i=1}^n\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^*}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*\left(R(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)\right)^2\\
&=0,
\end{aligned}
\tag{A.18}
$$

where $\tilde{\epsilon}_i^*$ is between ϵ_i and $\epsilon_i + R(V_i) - X_i^{*T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. According to Theorem 2.4.3, we know $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\delta_n)$, which indicates that

$$\begin{aligned}
&\sup_{i:|V_i - v|/h_2 \leq 1} |R(V_i) - X_i^{*T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)| \\
&\leq \sup_{i:|V_i - v|/h_2 \leq 1} \{|R(V_i)| + |X_i^{*T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| \\
&+ |m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)|\} = O_p(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|) = O_p(\delta_n), \qquad \text{(A.19)}
\end{aligned}$$

with the condition $h/h_2 \to 0$. Combining (A.19) with the Proof of Theorem 2.4.3, we can see that the third part which is associated with $\boldsymbol{X}_i^* \left(R(V_i) - X_i^{*T}(\tilde{\theta} - \theta_0) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)\right)^2$ is dominated by the second part which is associated with $\boldsymbol{X}_i^*(R(V_i) - X_i^{*T}(\tilde{\theta} - \theta_0) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i))$. We then mainly focus on the first two parts of the left side of (A.18).

Considering $-\frac{1}{nh_1^2h_2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right) \boldsymbol{X}_i^* + \frac{1}{nh_1^3h_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right) \boldsymbol{X}_i^*(R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i))$, by some direct calculations, we can obtain

$$\begin{aligned}
&E\Bigg(- \frac{1}{nh_1^2h_2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right) \boldsymbol{X}_i^* \\
&+ \frac{1}{nh_1^3h_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_2}\right) \boldsymbol{X}_i^*(R(V_i) - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i))\Bigg) \\
=& - \frac{1}{h_1^2h_2}\iiint \phi^{(1)}\left(\frac{\epsilon}{h_1}\right) \boldsymbol{X}^* g_\epsilon(\epsilon \mid X^*) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) f_V(V) d\epsilon dV dF(X^*) \\
&+ \frac{1}{h_1^3h_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right) \boldsymbol{X}^* g_\epsilon(\epsilon \mid X^*) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) (R(V) - m^{(1)}(\bar{V})(\hat{V} - V)) \\
&f_V(V) d\epsilon dV dF(X^*) \\
=&\frac{1}{h_1}\iiint \phi(\tau)\, \tau \boldsymbol{X}^* g_\epsilon(\tau h_1 \mid X^*) K\left(w + w'\right) f_V(wh_2 + v) dw d\tau dF(X^*) - \frac{1}{h_1^2}\iiint \phi(\tau) \\
&(\tau^2 - 1)\boldsymbol{X}^* g_\epsilon(\tau h_1 \mid X^*) K\left(w + w'\right) (R(V) - m^{(1)}(\bar{V})(\hat{V} - V)) f_V(wh_2 + v) dw d\tau dF(X^*)
\end{aligned}$$

$$=\left\{ \frac{h_1^2}{2} f_V(v) E \begin{bmatrix} \mu_0 Z_X g_\epsilon^{(3)}(0 \mid X^*) \\ \mu_0 g_\epsilon^{(3)}(0 \mid X^*) \\ \mu_1 g_\epsilon^{(3)}(0 \mid X^*) \end{bmatrix} - \left(\frac{h_2^2 m^{(2)}(v)}{2} f_V(v) E \begin{bmatrix} \mu_2 Z_X g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_2 g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_3 g_\epsilon^{(2)}(0 \mid X^*) \end{bmatrix} \right) \right\}$$

$$\{1 + o_p(1)\}, \tag{A.20}$$

where $Z_X = [X \ Z_1^T]^T$, $w' = h_2^{-1}(\hat{V} - V)$, $h/h_2 \to 0$, $\int \tau^4 \phi(\tau) d\tau = 3$, $\int \tau^2 \phi(\tau) d\tau = 1$, and $\int w^j K(w) dw = \mu_j$.

Considering $\frac{1}{nh_1^3 h_2} \sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) \boldsymbol{X}_i^* X_i^{*T}$, by direct calculations, we have

$$\begin{aligned}
& E\left(\frac{1}{nh_1^3 h_2} \sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) \boldsymbol{X}_i^* X_i^{*T} \right) \\
=& E\left(\frac{1}{h_1^3 h_2} \phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) \boldsymbol{X}_i^* X_i^{*T} \right) \\
=& \frac{1}{h_1^3 h_2} \iiint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right) \boldsymbol{X}^* \boldsymbol{X}^{*T} g_\epsilon(\epsilon \mid X^*) K\left(\frac{V - v + \hat{V} - V}{h_2}\right) f_V(V) d\epsilon dV dF(X^*) \\
=& \frac{1}{h_1^2} \iiint \phi(\tau)(\tau^2 - 1) \boldsymbol{X}^* \boldsymbol{X}^{*T} g_\epsilon(\tau h_1 \mid X^*) K(w) f_V(w h_2 + v) dw d\tau dF(X^*)(1 + o_p(1)) \\
=& f_V(v) E \begin{bmatrix} \mu_0 Z_X Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_0 Z_X g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 Z_X g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_0 Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_0 g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_1 Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 g_\epsilon^{(2)}(0 \mid X^*) & \mu_2 g_\epsilon^{(2)}(0 \mid X^*) \end{bmatrix}.
\end{aligned} \tag{A.21}$$

Based on the above two equations (A.20) and (A.21), we can achieve

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = E \begin{bmatrix} \mu_0 Z_X Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_0 Z_X g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 Z_X g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_0 Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_0 g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 g_\epsilon^{(2)}(0 \mid X^*) \\ \mu_1 Z_X^T g_\epsilon^{(2)}(0 \mid X^*) & \mu_1 g_\epsilon^{(2)}(0 \mid X^*) & \mu_2 g_\epsilon^{(2)}(0 \mid X^*) \end{bmatrix}^{-1}$$

$$\left\{\frac{h_1^2}{2}f_V(v)E\begin{bmatrix}\mu_0 Z_X g_\epsilon^{(3)}(0\mid X^*)\\ \mu_0 g_\epsilon^{(3)}(0\mid X^*)\\ \mu_1 g_\epsilon^{(3)}(0\mid X^*)\end{bmatrix}-\left(\frac{h_2^2 m^{(2)}(v)}{2}f_V(v)E\begin{bmatrix}\mu_2 Z_X g_\epsilon^{(2)}(0\mid X^*)\\ \mu_2 g_\epsilon^{(2)}(0\mid X^*)\\ \mu_3 g_\epsilon^{(2)}(0\mid X^*)\end{bmatrix}\right)\right\}\{1+o_p(1)\}. \tag{A.22}$$

Note that we can ignore the effect of the first step estimator on the asymptotic variance of the first stage estimator due to the faster convergence of the first step estimator. It is calculated that

$$\begin{aligned}&Var\Big(\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))\Big)\\ &=O_p\left(\frac{1}{nh_1^5h_2}\right)O_p(Var(\hat{V}_i-V_i))=O_p\left(\frac{1}{n^2h^3h_1^5h_2}\right)=o_p\left(\frac{1}{nh_1^3h_2}\right).\end{aligned} \tag{A.23}$$

Meanwhile, with the condition $h_2^2/h_1\to 0$ held, we could obtain

$$\begin{aligned}&\mathrm{Var}\left(-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*\right.\\ &\left.+\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*(R(V_i)-m^{(1)}(\bar{V}_i)(\hat{V}_i-V_i))\right)\\ =&E\left(-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*\right)\\ &\left(-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*\right)^T(1+o_p(1))\\ =&\frac{1}{nh_1^4h_2^2}\iiint\phi^{(1)2}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon\mid X^*)K^2\left(\frac{V-v+\hat{V}-V}{h_2}\right)f_V(V)d\epsilon dVdF(X^*)\\ &(1+o_p(1))\\ =&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_1^3h_2}f_V(v)E\begin{bmatrix}v_0Z_XZ_X^Tg_\epsilon(0\mid X^*) & v_0Z_Xg_\epsilon(0\mid X^*) & v_1Z_Xg_\epsilon(0\mid X^*)\\ v_0Z_X^Tg_\epsilon(0\mid X^*) & v_0g_\epsilon(0\mid X^*) & v_1g_\epsilon(0\mid X^*)\\ v_1Z_X^Tg_\epsilon(0\mid X^*) & v_1g_\epsilon(0\mid X^*) & v_2g_\epsilon(0\mid X^*)\end{bmatrix}(1+o_p(1)),\end{aligned} \tag{A.24}$$

where $v_j = \int w^j K^2(w)dw$. Define $W_n = \frac{1}{nh_1^2h_2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right) K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_2}\right)\boldsymbol{X}_i^*$. To show Theorem 2.4.4, it is sufficient to show that

$$T_n = \sqrt{nh_2h_1^3}W_n \xrightarrow{d} \mathcal{N}(0,T), \tag{A.25}$$

where $T = \int \tau^2\phi^2(\tau)d\tau f_V(v)E\begin{bmatrix} v_0Z_XZ_X^Tg_\epsilon(0\mid X^*) & v_0Z_Xg_\epsilon(0\mid X^*) & v_1Z_Xg_\epsilon(0\mid X^*) \\ v_0Z_X^Tg_\epsilon(0\mid X^*) & v_0g_\epsilon(0\mid X^*) & v_1g_\epsilon(0\mid X^*) \\ v_1Z_X^Tg_\epsilon(0\mid X^*) & v_1g_\epsilon(0\mid X^*) & v_2g_\epsilon(0\mid X^*) \end{bmatrix}$.

By Slutsky's theorem and the above two equations, we can obtain Theorem 2.4.4. To show the above equation, we prove that for any unit vector $\boldsymbol{d}\in R^2$,

$$\left\{\boldsymbol{d}^T\operatorname{Cov}(T_n)\,\boldsymbol{d}\right\}^{-1/2}\left\{\boldsymbol{d}^TT_n - \boldsymbol{d}^TE(T_n)\right\} \xrightarrow{d} N(0,1). \tag{A.26}$$

Then, we check Lyapunov's condition. Let

$$\xi_i = \sqrt{h_2h_1^3/n}K\left(\frac{V_i-v}{h_2}\right)\frac{1}{h_1h_2}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\boldsymbol{d}^T\boldsymbol{X}_i^*, \tag{A.27}$$

we need to prove $nE|\xi_1|^3 \to 0$. As $\left(\boldsymbol{d}^T\boldsymbol{X}_i^*\right)^2 \le \|\boldsymbol{d}\|^2\|\boldsymbol{X}_i^*\|^2, \phi^{(1)}(\cdot)$ is bounded, and $K(\cdot)$ has compact support, we have

$$nE|\xi|^3 \le O\left(nn^{-3/2}h_2^{-3/2}h_1^{3/2}\right)E\left|K^3\left(\frac{V_i-v}{h_2}\right)\phi^{(1)3}\left(\frac{\epsilon_i}{h_1}\right)\boldsymbol{d}^T\boldsymbol{X}_i^*\right| \to 0. \tag{A.28}$$

Thus, the asymptotic normality for T_n holds.

Proof of Theorem 2.4.5

The critical steps of the proof in this part is similar to these of Theorem 2.4.3, we thus outline the main steps here. Notice that the meanings of notations in this part are independent of other parts. Recall that

$$\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i-\tilde{m}(\hat{V}_i)-X_i\beta-Z_{1,i}^T\gamma}{h_3}\right),$$
$$=\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i-m(V_i)-X_i\beta-Z_{1,i}^T\gamma+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right), \quad \text{(A.29)}$$

where $m(V_i)-\tilde{m}(V_i)=O_p\left(\left(nh_2h_1^3\right)^{-1/2}+h_1^2+h_2^2\right)=O_p(n^{-2/8})$ and $\tilde{m}(\hat{V}_i)-\tilde{m}(V_i)=\tilde{m}^{(1)}(\bar{V}_i)(\hat{V}_i-V_i)$ in which $\bar{V}_i$ is between $\hat{V}_i$ and V_i. Based on Theorem 2.4.1, we know $|\hat{V}_i-V_i|=O_p((nh^3)^{-1/2}+h^2)=O_p(n^{-2/7})$. Define $\theta=(\beta,\gamma^T)^T$ and $X_i^*=[X_i,Z_{1,i}^T]^T$, we have

$$Q_n(\theta)=\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i-m(V_i)-X_i^{*T}\theta+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right). \quad \text{(A.30)}$$

Define $\delta_n=h_3^2+\sqrt{(nh_3^3)^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c}Q_n\left(\boldsymbol{\theta}_0+\delta_n\boldsymbol{\mu}\right)<Q_n\left(\boldsymbol{\theta}_0\right)\right\}\geq 1-\eta,$$

where $\boldsymbol{\theta}_0$is the true parameter. Using Taylor expansion, it follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\theta}_0+\delta_n\boldsymbol{\mu}\right)-Q_n\left(\boldsymbol{\theta}_0\right)\\
=&\frac{1}{nh_3}\sum_{i=1}^{n}\left[\phi\left(\frac{\epsilon_i-\delta_n\mu^TX_i^*+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right)\right.\\
&\left.-\phi\left(\frac{\epsilon_i+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right)\right]\\
=&\frac{1}{nh_3}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\epsilon_i+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_3}\right)\right.\\
&+\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i+m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i)}{h_3}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_3}\right)^2\\
&\left.-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{h_3}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_3}\right)^3\right]\\
=&I_1+I_2+I_3, \quad \text{(A.31)}
\end{aligned}$$

where ϵ_i^* is between $\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)$ and $\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - \delta_n \mu^T X_i^*$. Based on the result $T_n = E(T_n) + O_p(\sqrt{\mathrm{Var}(T_n)})$, we consider each part of above Taylor expansion.

(i) For the first part, which is $I_1 = \frac{1}{nh_3}\sum_{i=1}^n -\phi^{(1)}\left(\frac{\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right)\left(\frac{\delta_n \mu^T X_i^*}{h_3}\right)$, by Taylor expansion, we can rewrite it as

$$\begin{aligned}
E(I_1) =& \frac{-\delta_n}{h_3} E\left(\phi^{(1)}\left(\frac{\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right)\left(\frac{\mu^T X_i^*}{h_3}\right)\right) \\
=& \frac{-\delta_n}{h_3} E\Bigg(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_3} + \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))\boldsymbol{\mu}^T X_i^*}{h_3^2} \\
& + \frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{**}}{h_3}\right)\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))^2 \boldsymbol{\mu}^T X_i^*}{h_3^3}\Bigg) \\
=& I_{11} + I_{12} + I_{13}, \qquad \text{(A.32)}
\end{aligned}$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)$. Notice that as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{13}, we instead use ϵ_i directly. By some direct calculations for each part, we can get

$$I_{11} = \frac{-\delta_n}{h_3} E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_3}\right) = O_p(\delta_n c h_3^2). \qquad \text{(A.33)}$$

$$\begin{aligned}
I_{12} =& \frac{-\delta_n}{h_3} E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_3}\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))}{h_3}\right) \\
=& \frac{-\delta_n}{h_3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\frac{\boldsymbol{\mu}^T X^*}{h_3} g_\epsilon(\epsilon \mid X^*)\frac{(m(V) - \tilde{m}(V) + \tilde{m}(V) - \tilde{m}(\hat{V}))}{h_3} d\epsilon dF(X^*) \\
=& \frac{-\delta_n}{h_3}\iint \phi(\tau)(\tau^2 - 1)\boldsymbol{\mu}^T X^* g_\epsilon(\tau h_3 \mid X^*)\frac{m(V) - \tilde{m}(V) + \tilde{m}(V) - \tilde{m}(\hat{V})}{h_3} d\tau dF(X^*) \\
=& O_p(\delta_n c(h_1^2 + h_2^2)), \qquad \text{(A.34)}
\end{aligned}$$

as h converges faster than h_1 and h_2. With the condition that $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$, it can be seen that I_{11} dominates I_{12} and I_{13}. Meanwhile, we obtain

$$\frac{\delta_n^2}{h_3^2}E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_3}\right)^2 = O_p(\delta_n^2 c^2 (h_3^3)^{-1}). \tag{A.35}$$

The above equations show that $I_1 = O_p(\delta_n c h_3^2) + O_p(\sqrt{\delta_n^2 c^2 (n h_3^3)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_2 = \frac{1}{nh_3}\sum_{i=1}^n \left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right)\left(\frac{\delta_n \mu^T X_i^*}{h_3}\right)^2\right)$, we can rewrite it as

$$\begin{aligned}
E(I_2) =& \frac{\delta_n^2}{2h_3}E\left(\phi^{(2)}\left(\frac{\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_3^2}\right) \\
=& \frac{\delta_n^2}{2h_3}E\Bigg(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_3^2} + \phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))(\boldsymbol{\mu}^T X_i^*)^2}{h_3^3} \\
&+ \frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{**}}{h_3}\right)\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))^2(\boldsymbol{\mu}^T X_i^*)^2}{h_3^4}\Bigg) \\
=& I_{21} + I_{22} + I_{23},
\end{aligned} \tag{A.36}$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)$. Notice that as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{23}, we instead use ϵ_i directly. By some calculations for each part, we can get

$$I_{21} = \frac{\delta_n^2}{2h_3}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_3^2}\right) = O_p((\delta_n c)^2). \tag{A.37}$$

$$I_{22} = \frac{\delta_n^2}{2h_3}E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))(\boldsymbol{\mu}^T X_i^*)^2}{h_3^3}\right) = o_p((\delta_n c)^2). \tag{A.38}$$

Meanwhile, we can prove that $I_{23} = o_p((\delta_n c)^2)$ as well. Following the same steps in (i), we obtain the following result

$$\frac{\delta_n^4}{4h_3^2}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_3^2}\right)^2 = O_p((\delta_n c)^4 (h_3^5)^{-1}). \tag{A.39}$$

With the condition $nh_3^5 \to \infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) Following the same way, we can calculate the third part. As the order of ϵ_i^* is the same as the order of ϵ_i, by direct calculations, we have

$$\frac{\delta_n^3}{6h_3} E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\left(\boldsymbol{\mu}^T X_i^*\right)^3}{h_3^3}\right) = O_p(\delta_n^3). \tag{A.40}$$

$$\frac{\delta_n^6}{36h_3^2} E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\left(\boldsymbol{\mu}^T X_i^*\right)^3}{h_3^3}\right)^2 = O_p(\delta_n^6 (h_3^7)^{-1}). \tag{A.41}$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, thus $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_0\right)\} \geq 1-\eta$ holds.

Proof of Theorem 2.4.6

Following the same steps as proving Theorem 2.4.2, the estimator $\tilde{\theta}$ must satisfy

$$-\frac{1}{nh_3^2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i - X_i^{*T}(\tilde{\theta}-\theta_0) + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i)}{h_3}\right)\boldsymbol{X}_i^* = 0. \tag{A.42}$$

By taking Taylor expansion, we could obtain

$$\begin{aligned}
&-\frac{1}{nh_3^2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* + \frac{1}{nh_3^3}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T} \\
&(\tilde{\theta}-\theta_0)) - \frac{1}{nh_3^4}\sum_{i=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^*}{h_3}\right)\boldsymbol{X}_i^*\left(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T}(\tilde{\theta}-\theta_0)\right)^2 = 0,
\end{aligned} \tag{A.43}$$

where $\tilde{\epsilon}_i^*$ is between ϵ_i and $\epsilon_i + m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Assume h converges faster than h_1 and h_2, $h_1/h_3 \to 0$, and $h_2/h_3 \to 0$, from Theorem 2.4.5,

we know $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\delta_n)$, which indicates that $|m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)| = O_p(\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\|) = O_p(\delta_n)$. Thus, the third part which is associated with $\boldsymbol{X}_i^* \left(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T}(\tilde{\theta} - \theta_0)\right)^2$ is dominated by the second part which is associated with $\boldsymbol{X}_i^* \left(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i) - X_i^{*T}(\tilde{\theta} - \theta_0)\right)$. We then mainly focus on the first two parts of the left side of (A.43).

Considering $-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* + \frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))$, by some direct calculations, we can obtain

$$
\begin{aligned}
&E\bigg(- \frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* + \frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))\bigg) \\
=& - \frac{1}{h_3^2}\iint \phi^{(1)}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{X}^* g_\epsilon(\epsilon \mid X^*)d\epsilon dF(X^*) \\
&+ \frac{1}{h_3^3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{X}^* g_\epsilon(\epsilon \mid X^*)(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))d\epsilon dF(X^*) \\
=&\frac{1}{h_3}\iint \phi(\tau)\,\tau \boldsymbol{X}^* g_\epsilon(\tau h_3 \mid X^*)d\tau dF(X^*) \\
&-\frac{1}{h_3^2}\iint \phi(\tau)\,(\tau^2 - 1)\boldsymbol{X}^* g_\epsilon(\tau h_3 \mid X^*)(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))d\tau dF(X^*) \\
=&\frac{h_3^2}{2}E(X^* g_\epsilon^{(3)}(0 \mid X^*))\{1 + o_p(1)\}.
\end{aligned} \tag{A.44}
$$

Considering $\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* X_i^{*T}$, by direct calculations, we have

$$
E\left(\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* X_i^{*T}\right) = E(X^* X^{*T} g_\epsilon^{(3)}(0 \mid X^*)). \tag{A.45}
$$

Based on the above two equations (A.44) and (A.45), we can achieve

$$
\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \frac{h_3^2}{2}(E(X^* X^{*T} g_\epsilon^{(3)}(0 \mid X^*)))^{-1} E(X^* g_\epsilon^{(3)}(0 \mid X^*))\{1 + o_p(1)\}. \tag{A.46}
$$

We can calculate

$$
Var\Big(\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i) - \tilde{m}(V_i) + \tilde{m}(V_i) - \tilde{m}(\hat{V}_i))\Big) = O_p\left(\frac{1}{n^2 h_3^5 h_2 h_1^3}\right).
$$

With the condition that $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$, it can be seen that the variance of $-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*$ dominates the variance of $\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i))$.

Meanwhile, with the condition $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$ held, we could obtain

$$\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^* + \frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{X}_i^*(m(V_i)-\tilde{m}(V_i)+\tilde{m}(V_i)-\tilde{m}(\hat{V}_i))\right)\\
=&\frac{1}{nh_3^4}\iint \phi^{(1)2}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon \mid X^*)d\epsilon dF(X^*)(1+o_p(1))\\
=&\frac{\int \tau^2\phi^2(\tau)d\tau}{nh_3^3}E(X^*X^{*T}g_\epsilon(0 \mid X^*))(1+o_p(1)).
\end{aligned} \tag{A.47}$$

For the remaining part, we could follow the same idea in the Proof of Theorem 2.4.4 to easily obtain the results.

Proof of Theorem 2.4.7

The proof is similar to those of Theorem 2.4.3, we thus mainly outline the main steps here. Notice that the meanings of notations in this part are independent of other parts. Recall that

$$\frac{1}{nh_4h_5}\sum_{i=1}^n \phi\left(\frac{Y_i - X_i\hat{\beta} - Z_{1,i}^T\hat{\gamma} - m(V_i) - (m(\hat{V}_i)-m(V_i))}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right), \tag{A.48}$$

where $m(\hat{V}_i)-m(V_i) = m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)$ and $\bar{V}_i$ is between $\hat{V}_i$ and V_i. From Theorem 2.4.1, we know $|\hat{V}_i - V_i| = O_p((nh^3)^{-1/2}+h^2) = O_p(n^{-2/7})$ with the MSE-optimal bandwidth. Define $\theta = (m(v), h_2m^{(1)}(v))^T$ and $X^* = [1\ h_2^{-1}(V-v)]^T$, we have

$$Q_n(\theta) = \frac{1}{nh_4h_5}\sum_{i=1}^n \phi\left(\frac{Y_i - X_i\hat{\beta} - Z_{1,i}^T\hat{\gamma} - X_i^{*T}\theta - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i)}{h_4}\right)$$

$$K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right). \tag{A.49}$$

Define $\delta_n = h_4^2 + h_5^2 + \sqrt{(nh_4^3h_5)^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that $P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_0\right)\right\} \geq 1-\eta$, where $\boldsymbol{\theta}_0$ is the true parameter. Using Taylor expansion, it follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\theta}_0\right) \\
=&\frac{1}{nh_4h_5}\sum_{i=1}^{n}\left[\phi\left(\frac{\tilde{R}(V_i) + \epsilon_i - \delta_n\mu^T X_i^*}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right)\right. \\
&\left. - \phi\left(\frac{\tilde{R}(V_i) + \epsilon_i}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right)\right] \\
=&\frac{1}{nh_4h_5}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\tilde{R}(V_i) + \epsilon_i}{h_4}\right)\left(\frac{\delta_n\mu^T X_i^*}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right)\right. \\
&+ \frac{1}{2}\phi^{(2)}\left(\frac{\tilde{R}(V_i) + \epsilon_i}{h_4}\right)\left(\frac{\delta_n\mu^T X_i^*}{h_4}\right)^2 K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right) \\
&\left. - \frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{h_4}\right)\left(\frac{\delta_n\mu^T X_i^*}{h_4}\right)^3 K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right)\right] \\
=&I_1 + I_2 + I_3,
\end{aligned} \tag{A.50}$$

where ϵ_i^* is between $\tilde{R}(V_i) + \epsilon_i$ and $\tilde{R}(V_i) + \epsilon_i - \delta_n\mu^T X_i^*$, and $\tilde{R}(V_i) = \sum_{j=2}(m^{(j)}(v)/j!)(V_i - v)^j - m^{(1)}(\bar{V}_i)(\hat{V}_i - V_i) - X_i(\hat{\beta} - \beta_0) - Z_{1,i}^T(\hat{\gamma} - \gamma_0)$. Based on the result $T_n = E\left(T_n\right) + O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is

$$I_1 = \frac{1}{nh_4h_5}\sum_{i=1}^{n} -\phi^{(1)}\left(\frac{\tilde{R}(V_i) + \epsilon_i}{h_4}\right)\left(\frac{\delta_n\mu^T X_i^*}{h_4}\right)K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right),$$

by Taylor expansion, we can rewrite it as

$$E(I_1) = \frac{-\delta_n}{h_4h_5}E\left(\phi^{(1)}\left(\frac{\tilde{R}(V_i) + \epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^T X_i^*}{h_4}K\left(\frac{V_i - v + \hat{V}_i - V_i}{h_5}\right)\right)$$

$$
\begin{aligned}
=&\frac{-\delta_n}{h_4h_5}E\Bigg(\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)+\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\tilde{R}(V_i)\boldsymbol{\mu}^TX_i^*}{h_4^2}\\
&K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)+\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{**}}{h_4}\right)\frac{(\tilde{R}(V_i))^2\boldsymbol{\mu}^TX_i^*}{h_4^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\Bigg)\\
=&I_{11}+I_{12}+I_{13},
\end{aligned}
\tag{A.51}
$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i+\tilde{R}(X_i)$. Notice that as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{13}, we instead use ϵ_i directly. By some direct calculations for each part, we can get

$$
I_{11}=\frac{-\delta_n}{h_4h_5}E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)=O_p(\delta_nch_4^2). \tag{A.52}
$$

$$
I_{12}=\frac{-\delta_n}{h_4h_5}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\frac{\tilde{R}(V_i)}{h_4}\right)=O_p(\delta_nch_5^2). \tag{A.53}
$$

$$
I_{13}\approx\frac{-\delta_n}{h_4h_5}E\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\tilde{R}(V_i))^2\boldsymbol{\mu}^TX_i^*}{h_4^3}K\left(\frac{V-v+\hat{V}-V}{h_5}\right)\right)=o_p(\delta_nh_5^2), \tag{A.54}
$$

where $h/h_5\to 0$ and $h_3/h_5\to 0$. Meanwhile, with the condition $h_5^2/h_4\to 0$ held, we obtain

$$
\frac{\delta_n^2}{h_4^2h_5^2}E\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)^2=O_p(\delta_n^2c^2(h_4^3h_5)^{-1}). \tag{A.55}
$$

$$
\frac{\delta_n^2}{h_4^2h_5^2}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\boldsymbol{\mu}^TX_i^*}{h_4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\frac{(\tilde{R}(V_i))}{h_4}\right)^2=o_p(\delta_n^2(h_4^3h_5)^{-1}). \tag{A.56}
$$

The above equations show that $I_1=O_p(\delta_nc(h_4^2+h_5^2))+O_p(\sqrt{\delta_n^2c^2(nh_4^3h_5)^{-1}})=O_p(\delta_n^2c)$.

(ii) For the second part, which is $I_2=\frac{1}{nh_4h_5}\sum_{i=1}^n\left(\frac{1}{2}\phi^{(2)}\left(\frac{\tilde{R}(V_i)+\epsilon_i}{h_4}\right)\left(\frac{\delta_n\mu^TX_i^*}{h_4}\right)^2\right.$ $\left.K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)$, we can rewrite it as

$$
E(I_2)=\frac{\delta_n^2}{2h_5h_4}E\left(\phi^{(2)}\left(\frac{\epsilon_i+\tilde{R}(V_i)}{h_4}\right)\frac{(\boldsymbol{\mu}^TX_i^*)^2}{h_4^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)
$$

$$
\begin{aligned}
=&\frac{\delta_n^2}{2h_5h_4}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_4^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right.\\
&+\phi^{(3)}\left(\frac{\epsilon_i}{h_4}\right)\frac{\tilde{R}(V_i)(\boldsymbol{\mu}^T X_i^*)^2}{h_4^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\\
&\left.+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{**}}{h_4}\right)\frac{(\tilde{R}(V_i))^2(\boldsymbol{\mu}^T X_t^*)^2}{h_4^4}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)\\
=&I_{21}+I_{22}+I_{23},
\end{aligned}
\tag{A.57}
$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i+\tilde{R}(V_i)$. Notice that as the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{23}, we instead use ϵ_i directly. By some calculations for each part, we can get

$$
I_{21}=\frac{\delta_n^2}{2h_5h_4}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_4^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)=O_p((\delta_n c)^2). \tag{A.58}
$$

Meanwhile, we can prove that $I_{22}=o_p((\delta_n c)^2)$ and $I_{23}=o_p((\delta_n c)^2)$. Following the same steps in (i), we obtain the following result

$$
\frac{\delta_n^4}{4h_5^2h_4^2}E\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^2}{h_4^2}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)^2=O_p((\delta_n c)^4(h_5h_4^5)^{-1}). \tag{A.59}
$$

With the condition $nh_4^5h_5\to\infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) Following the same steps as the proofs in Theorem 2.4.3, we can calculate the third part. As the order of ϵ_i^* is the same as the order of ϵ_i, by direct calculations, we have

$$
\frac{\delta_n^3}{6h_5h_4}E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^3}{h_4^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)=O_p(\delta_n^3). \tag{A.60}
$$

$$
\frac{\delta_n^6}{36h_5^2h_4^2}E\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_4}\right)\frac{(\boldsymbol{\mu}^T X_i^*)^3}{h_4^3}K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\right)^2=O_p(\delta_n^6(h_5h_4^7)^{-1}). \tag{A.61}
$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, thus $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_0+\delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_0\right)\} \geq 1-\eta$ holds.

Proof of Theorem 2.4.8

Based on (A.50), the estimator $\tilde{\theta}$ must satisfy the following equation

$$-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i+\tilde{R}(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*=0. \tag{A.62}$$

By taking Taylor expansion, we could obtain

$$\begin{aligned}
&-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^* \\
&+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*(\tilde{R}(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)) \\
&-\frac{1}{nh_4^4h_5}\sum_{i=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^*}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*\left(\tilde{R}(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)\right)^2=0,
\end{aligned} \tag{A.63}$$

where $\tilde{\epsilon}_i^*$ is between ϵ_i and $\epsilon_i+\tilde{R}(V_i)-X_i^{*T}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$. From Theorem 2.4.3, we know $\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|=O_p(\delta_n)$ with $h/h_5\to 0$ and $h_3/h_5\to 0$, which indicates that

$$\sup_{i:|V_i-v|/h_2\leq 1}|\tilde{R}(V_i)-X_i^{*T}(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)|=O_p(\|\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|)=O_p(\delta_n). \tag{A.64}$$

Combining (A.64) with the Proof of Theorem 2.4.7, we can see that the third part which is associated with $\boldsymbol{X}_i^*\left(\tilde{R}(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)\right)^2$ is dominated by the second part which is associated with $\boldsymbol{X}_i^*\left(\tilde{R}(V_i)-X_i^{*T}(\tilde{\theta}-\theta_0)\right)$. We then mainly focus on the first two parts of the left side of (A.63).

Considering $-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*\tilde{R}(V_i)$, by some direct calculations, we can obtain

$$
\begin{aligned}
&E\bigg(-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^* \\
&+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*\tilde{R}(V_i)\bigg) \\
=&\left\{\frac{h_1^2}{2}f_V(v)E\begin{bmatrix}\mu_0 g_\epsilon^{(3)}(0\mid X^*)\\ \mu_1 g_\epsilon^{(3)}(0\mid X^*)\end{bmatrix}-\left(\frac{h_2^2m^{(2)}(v)}{2}f_V(v)E\begin{bmatrix}\mu_2 g_\epsilon^{(2)}(0\mid X^*)\\ \mu_3 g_\epsilon^{(2)}(0\mid X^*)\end{bmatrix}\right)\right\}\{1+o_p(1)\}
\end{aligned}
\tag{A.65}
$$

with $h/h_5\to 0$ and $h_3/h_5\to 0$. Considering $\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V-v+\hat{V}-V}{h_5}\right)\boldsymbol{X}_i^*X_i^{*T}$, by direct calculations, we have

$$
\begin{aligned}
&E\left(\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V-v+\hat{V}-V}{h_5}\right)\boldsymbol{X}_i^*X_i^{*T}\right) \\
=&f_V(v)E\begin{bmatrix}\mu_0 g_\epsilon^{(2)}(0\mid X^*) & \mu_1 g_\epsilon^{(2)}(0\mid X^*)\\ \mu_1 g_\epsilon^{(2)}(0\mid X^*) & \mu_2 g_\epsilon^{(2)}(0\mid X^*)\end{bmatrix}.
\end{aligned}
\tag{A.66}
$$

Based on the above two equations (A.65) and (A.66), we can achieve

$$
\begin{aligned}
\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0=&E\begin{bmatrix}\mu_0 g_\epsilon^{(2)}(0\mid X^*) & \mu_1 g_\epsilon^{(2)}(0\mid X^*)\\ \mu_1 g_\epsilon^{(2)}(0\mid X^*) & \mu_2 g_\epsilon^{(2)}(0\mid X^*)\end{bmatrix}^{-1} \\
&\left\{\frac{h_1^2}{2}f_V(v)E\begin{bmatrix}\mu_0 g_\epsilon^{(3)}(0\mid X^*)\\ \mu_1 g_\epsilon^{(3)}(0\mid X^*)\end{bmatrix}-\left(\frac{h_2^2m^{(2)}(v)}{2}f_V(v)E\begin{bmatrix}\mu_2 g_\epsilon^{(2)}(0\mid X^*)\\ \mu_3 g_\epsilon^{(2)}(0\mid X^*)\end{bmatrix}\right)\right\}\{1+o_p(1)\}.
\end{aligned}
\tag{A.67}
$$

Meanwhile, with the condition $h_5^2/h_4\to 0$ held, we could obtain

$$
\mathrm{Var}\left(-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*\right.
$$

$$+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{V_i-v+\hat{V}_i-V_i}{h_5}\right)\boldsymbol{X}_i^*\tilde{R}(V_i)\Bigg)$$

$$=\frac{1}{nh_4^4h_5^2}\iint\phi^{(1)2}\left(\frac{\epsilon}{h_4}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon\mid X^*)K^2\left(\frac{V-v+\hat{V}-V}{h_5}\right)f_V(V)d\epsilon dV$$

$$(1+o_p(1))$$

$$=\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_4^3h_5}f_V(v)E\begin{bmatrix}v_0g_\epsilon(0\mid X^*) & v_1g_\epsilon(0\mid X^*)\\ v_1g_\epsilon(0\mid X^*) & v_2g_\epsilon(0\mid X^*)\end{bmatrix}(1+o_p(1)). \tag{A.68}$$

For the remaining part, we could follow the same idea in the Proof of Theorem 2.4.4 to easily obtain the results.

Lemma A.5.4 *Denote* $G_n(u)=\sum_{i=1}^n[\phi_h(V_i-Z_i^{*T}u/\sqrt{nh^3})-\phi_h(V_i)]$, *where* $V_i=X_i-Z_i^{*T}\theta_0$. *Under the same conditions as those in Theorem 2.4.2, for any fixed* u,

$$G_n(u)=\frac{g^{(2)}(0\mid Z)}{2}u^T\frac{\sum_{i=1}^n Z_i^*Z_i^{*T}}{nh^3}u+W_n^Tu+o_p(1),$$

where $W_n=\sum_{i=1}^n\phi_h^{(1)}(V_i)Z_i^*/\sqrt{nh^3}$.

Proof. The proof of this lemma should be followed by the arguments in Wu and Liu (2009). To save space, we skip the details of the proof. ■

Proof of Theorem 2.6.9

For any $\theta_1-\theta_{10}=O_p(h^2+(nh^3)^{-1/2})$, $0<\|\theta_2\|<C\delta_n$, where $\delta_n=h^2+(nh^3)^{-1/2}$, we have

$$Q((\theta_1^T,0)^T)-Q((\theta_1^T,\theta_2^T)^T)$$

$$=[Q((\theta_1^T,0)^T)-Q((\theta_{10}^T,0)^T)]-[Q((\theta_1^T,\theta_2^T)^T)-Q((\theta_{10}^T,0)^T)]$$

$$=G_n(\sqrt{nh^3}((\theta_1-\theta_{10})^T,0)^T)-G_n(\sqrt{nh^3}((\theta_1-\theta_{10})^T,\theta_2^T)^T)-\lambda_n\sum_{j=s+1}^{d_Z+1}\hat{w}_j|\theta_j|, \tag{A.69}$$

where the first two terms are bounded. With the optimal value of bandwidth by minimizing $O_p(h^2 + (nh^3)^{-1/2})$, we have $h^2 = (nh^3)^{-1/2}$. We thus in the following proof only focus on the use of $(nh^3)^{-1/2}$. The third terms goes to $-\infty$ as $n \to \infty$ due to

$$\lambda_n \sum_{j=s+1}^{d_Z+1} \hat{w}_j |\theta_j| = \left((nh^3)^{(\gamma-1)/2}\lambda_n\right) \sqrt{nh^3} \sum_{j=s+1}^{d_Z+1} |\left(\sqrt{nh^3}|\hat{\theta}_j|\right)^{-\gamma} ||\theta_j| \to \infty. \tag{A.70}$$

Hence, the condition that $(nh^3)^{(\gamma-1)/2}\lambda_n \to \infty$ implies that $\lambda_n \sum_{j=s+1}^{d_Z+1} \hat{w}_j|\theta_j|$ is of higher order than any other terms and dominates as a result. This in turn implies that $Q((\theta_1^T, 0)^T) - Q((\theta_1^T, \theta_2^T)^T) < 0$ for large n, which proves the result.

Proof of Theorem 2.6.10

At first, we know that

$$\begin{aligned} & Q\left(\theta_0 + u/\sqrt{nh^3}\right) - Q\left(\theta_0\right) \\ =& \frac{1}{n}\sum_{i=1}^{n}\left[\phi_h\left(X_i - Z_i^{*T}\left(\theta_0 + u/\sqrt{nh^3}\right)\right) - \phi_h\left(X_i - Z_i^{*T}\theta_0\right)\right] \\ & + \lambda_n \sum_{j=1}^{d_Z+1}\left[\hat{w}_j|\theta_{j0} + u_j/\sqrt{nh^3}| - \hat{w}_j|\theta_{j0}|\right]. \end{aligned} \tag{A.71}$$

Following Wu and Liu (2009), we consider the second term first. For $j = 1, 2, \cdots, s$, we have $\theta_{j0} \neq 0$; as a result, $\hat{w}_j \xrightarrow{P} |\theta_{j0}|^{-\gamma}$. Hence

$$\lambda_n\left[\hat{w}_j|\theta_{j0} + u_j/\sqrt{nh^3}| - \hat{w}_j|\theta_{j0}|\right] \xrightarrow{P} 0 \tag{A.72}$$

as $\sqrt{nh^3}(|\theta_{j0} + u_j/\sqrt{nh^3}| - |\theta_{j0}|) \to u_j \operatorname{sign}(\theta_{j0})$ and $\sqrt{nh^3}\lambda_n \to 0$. On the other hand, for $j = s+1, \cdots, d$, the true coefficient $\theta_{j0} = 0$; so $\sqrt{nh^3}\lambda_n\hat{w}_j = (nh^3)^{(1+\gamma)/2}\lambda_n\left(\sqrt{nh^3}|\hat{\theta}_j|\right)^{-\gamma}$ with $\sqrt{nh^3}\hat{\theta}_j = O_p(1)$; so it follows that $nh^3\lambda_n\left[\hat{w}_j|\theta_{j0} + u_j/\sqrt{nh^3}| - \hat{w}_j|\theta_{j0}|\right] \xrightarrow{P} \infty$ when $u_j \neq 0$ and $= 0$ otherwise due to $\sqrt{nh^3}|u_j/\sqrt{nh^3}| = |u_j|$ for large n. These facts and the result of Lemma A.5.4 imply that

$$
\begin{aligned}
& Q\left(\theta_0+\frac{u}{\sqrt{nh^3}}\right)-Q\left(\theta_0\right) \xrightarrow{\mathcal{L}} V(u) \\
&= \begin{cases} \frac{g^{(2)}(0|Z)}{2} u_1^T E(Z_{i1}^* Z_{i1}^{*T}) u_1 + W_{n1}^T u_1 & \text{when } u_j = 0 \text{ for } j \geq s+1 \\ \infty & \text{otherwise,} \end{cases}
\end{aligned} \tag{A.73}
$$

where $u_1 = (u_1, u_2, \cdots, u_s)^T$. Noticing that $Q\left(\theta_0+\frac{u}{\sqrt{nh^3}}\right)-Q\left(\theta_0\right)$ is concave in u and V has a unique maximizer, it indicates that

$$
\operatorname{argmax} Q\left(\theta_0+\frac{u}{\sqrt{nh^3}}\right) = \sqrt{nh^3}\left(\hat{\theta}-\theta_0\right) \xrightarrow{\mathcal{L}} \operatorname{argmax} V(u), \tag{A.74}
$$

which establishes the asymptotic normality result.

Appendix B

Appendix for Chapter 3

B.1 Monte Carlo Experiment

To illustrate the applicability of the proposed variance reduction method, we generate data from the two different data generating processes (DGPs) shown below, where ϵ_t is simulated from $0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$ with $E(\epsilon) = 0$ and $Mode(\epsilon) = 1$. Thus, the modal regression line is different from the mean regression line. The sample sizes under consideration are $n = \{200, 400, 600, 1000\}$.

(**DGP 1**) $Y_t = \exp(\pi X_t) + \sigma(X_t)\epsilon_t$, where the observation X_t is i.i.d., univariate and uniformed distributed in [0,1], and $\sigma(X_t) = 1 + 2X_t$. We then have $Mode(Y_t \mid X_t) = \exp(\pi X_t) + 1 + 2X_t$ and $E(Y_t \mid X_t) = \exp(\pi X_t)$.

(**DGP 2**) $Y_t = sin(\pi X_t) + \sigma(X_t)\epsilon_t$, where the observation X_t stems from a time series $X_t = 0.5X_{t-1} + \eta_t$ with standard normal innovations η_t, and $\sigma(X_t) = X_t$. The setting indicates that $Mode(Y_t \mid X_t) = sin(\pi X_t) + X_t$ and $E(Y_t \mid X_t) = sin(\pi X_t)$.

To evaluate the performance of the variance reduced estimators, we calculate the MSE for each estimate over 200 replicates

$$\frac{1}{200n}\sum_{l=1}^{200}\sum_{t=1}^{n}\left(Mode(Y_t \mid X_t) - \hat{m}(X_t)^{(l)}\right)^2,$$

in which $\hat{m}(X_t)^{(l)}$ is the estimate at the lth replication. We compute the ratio of MSE for local linear modal regression and variance reduced modal regression for each sample. The results are summarized in Table B.1, from which we can see that the variance reduced modal estimator has a smaller MSE than the local linear modal estimator for all sample sizes. The results are consistent with the asymptotic results derived in the paper.

Table B.1: Results of Simulations

	DGP 1-Mode-Reduction	DGP 2-Mode-Reduction
n=200	0.8868	0.9396
n=400	0.8074	0.8508
n=600	0.7720	0.8379
n=1000	0.7217	0.7809

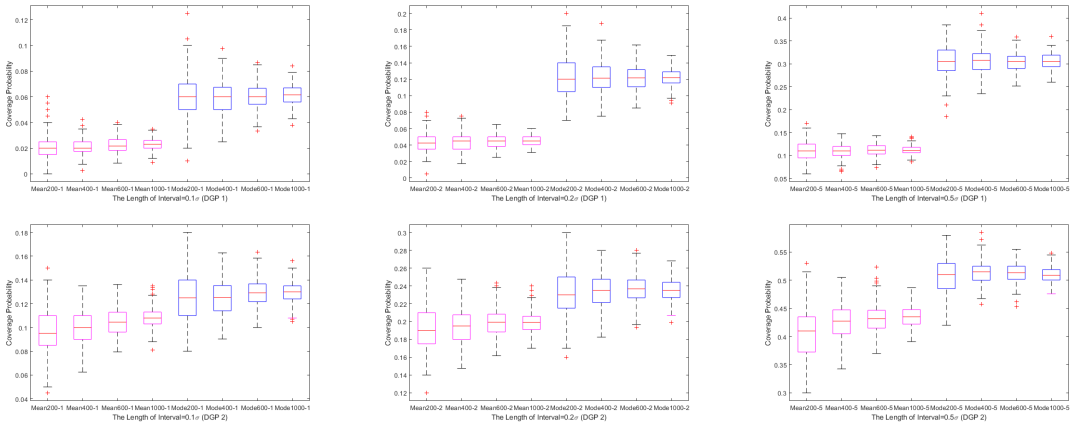

Figure B.1: Coverage Probabilities

We also report the coverage probabilities to evaluate the prediction performance of mean and modal regressions. The lengths of the intervals considered are 0.1σ, 0.2σ, and 0.5σ, respectively, where $\sigma \approx 2$. The calculation steps are identical to those in Ullah et al. (2021). Figure B.1 shows that modal regression can provide higher coverage probabilities than mean regression, indicating the prediction advantage of modal regression.

B.2 Technical Proofs

Lemma B.2.5 *Under the conditions C1-C5, with $nh_1^3h_2 \to \infty$ held, we have*

$$\begin{aligned}
&-\frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\left(\frac{X_t-x}{h_2}\right)\\
=&\frac{h_1^2}{6}\iint g_\epsilon^{(3)}(0\mid x)t^4\phi\left(t\right)s^lK(s)f_X\left(x\right)dtds(1+o(1)).
\end{aligned}$$

Proof. We proceed along the lines of the proofs in Cai and Ould-Said (2003) and Wang and Tang (2016). Define $Z_{n,t} = -\frac{1}{h_1h_2}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\left(\frac{X_t-x}{h_2}\right)$, with the assumptions of C1-C3 and conditioning on X, we obtain

$$\begin{aligned}
E\left(Z_{n,1}\right) &= \iint \frac{\epsilon}{h_1^3h_2}\phi\left(\frac{\epsilon}{h_1}\right)K\left(\frac{X-x}{h_2}\right)\left(\frac{X-x}{h_2}\right)g_\epsilon(\epsilon\mid x)f_X(x)d\epsilon dx\\
&= \frac{1}{h_1}\iint t\phi\left(t\right)sK(s)g_\epsilon(th_1\mid x)f_X\left(sh_2+x\right)dtds\\
&= \frac{h_1^2}{6}\iint g_\epsilon^{(3)}(0\mid x)t^4\phi\left(t\right)sK(s)f_X\left(x\right)dtds(1+o(1)). \qquad \text{(B.1)}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&E\left\{-\frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\left(\frac{X_t-x}{h_2}\right)\right\}\\
=&\frac{h_1^2}{6}\iint g_\epsilon^{(3)}(0\mid x)t^4\phi\left(t\right)sK(s)f_X\left(x\right)dtds(1+o(1)). \qquad \text{(B.2)}
\end{aligned}$$

Note that

$$\sum_{t=1}^{n} Z_{n,t} = E\left(\sum_{t=1}^{n} Z_{n,t}\right) + O_p\left(\sqrt{\mathrm{Var}\left(\sum_{t=1}^{n} Z_{n,t}\right)}\right), \tag{B.3}$$

and the stationary of $\{\epsilon_t\}$ gives

$$\mathrm{Var}\left(\sum_{t=1}^{n} Z_{n,t}\right) = nEZ_{n,1}^2 + 2\sum_{j=2}^{n}(n-j+1)\,\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right). \tag{B.4}$$

With the assumptions of C1-C3 and conditioning on X, we can have

$$\begin{aligned} E\left(Z_{n,1}^2\right) &= \frac{1}{h_1^2 h_2^2}\iint \frac{\epsilon^2}{h_1^4}\phi^2\left(\frac{\epsilon}{h_1}\right) K^2\left(\frac{X-x}{h_2}\right)\left(\frac{X-x}{h_2}\right)^2 g_\epsilon(\epsilon \mid x) f_X(x) d\epsilon dx \\ &= \frac{1}{h_1^3 h_2}\iint t^2\phi^2\left(t\right) s^2 K^2(s) g_\epsilon(th_1 \mid x) f_X\left(sh_2 + x\right) dtds \\ &= \frac{1}{h_1^3 h_2}\iint g_\epsilon(0 \mid x) t^2\phi^2\left(t\right) s^2 K^2(s) f_X\left(x\right) dtds(1+o(1)). \end{aligned} \tag{B.5}$$

To obtain an upper bound for the second term on the right-hand side of the above equation, let d_n be a sequence of positive integers satisfying $d_n \to \infty$ and $d_n h_1 h_2 \to 0$ as $n \to \infty$, we can split it into two terms

$$\sum_{j=2}^{n}|\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| = \sum_{j=2}^{d_n}|\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| + \sum_{j=d_n+1}^{n}|\mathrm{Cov}\left(Z_{n,t}, Z_{n,j}\right)|. \tag{B.6}$$

By the assumptions of C1-C3 and conditioning on X_t and X_j, we can have the following result, where

$$\begin{aligned} |EZ_{n,t}Z_{n,j}| \le E|Z_{n,t}Z_{n,j}| &= E\Bigg|E\left[\frac{1}{h_1^2}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)\phi^{(1)}\left(\frac{\epsilon_j}{h_1}\right)\Big|X_t, X_j\right] \\ &\quad \cdot \frac{1}{h_2^2}K\left(\frac{X_t-x}{h_2}\right)\left(\frac{X_t-x}{h_2}\right)K\left(\frac{X_j-x}{h_2}\right)\left(\frac{X_j-x}{h_2}\right)\Bigg| \\ &= E\Bigg|\iint \frac{1}{h_1^2}\frac{\epsilon}{h_1^2}\phi\left(\frac{\epsilon}{h_1}\right)\frac{\epsilon^*}{h_1^2}\phi\left(\frac{\epsilon^*}{h_1}\right) g_\epsilon(\epsilon, \epsilon^* \mid X_t, X_j) d\epsilon d\epsilon^* \\ &\quad \cdot \frac{1}{h_2^2}K\left(\frac{X_t-x}{h_2}\right)\left(\frac{X_t-x}{h_2}\right)K\left(\frac{X_j-x}{h_2}\right)\left(\frac{X_j-x}{h_2}\right)\Bigg| \end{aligned}$$

$$\leq C_1 \frac{1}{h_1^2 h_2^2} E\left| K\left(\frac{X_t - x}{h_2}\right)\left(\frac{X_t - x}{h_2}\right) K\left(\frac{X_j - x}{h_2}\right)\left(\frac{X_j - x}{h_2}\right)\right| \leq C_2 \frac{1}{h_1^2}, \tag{B.7}$$

in which C_1 and C_2 are constants. Therefore, we have

$$\sum_{j=2}^{d_n} |\mathrm{Cov}\left(Z_{n,t}, Z_{n,j}\right)| \leq C_2 \frac{1}{h_1^2} \sum_{j=2}^{d_n} 1 = o\left(n h_2^{-1} h_1^{-3}\right). \tag{B.8}$$

By applying Davydov's inequality, we have

$$|\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| \leq C_3 [\rho(j-1)]^{\delta/(2+\delta)} \left(E|Z_{n,1}|^{2+\delta}\right)^{2/(2+\delta)} \tag{B.9}$$

and obtain

$$\begin{aligned} E|Z_{n,t}|^{2+\delta} &= E\left| E\left[\frac{1}{h_1}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)|X_t\right] \frac{1}{h_2} K\left(\frac{X_t - x}{h_2}\right)\left(\frac{X_j - x}{h_2}\right)\right|^{2+\delta} \\ &\leq C_4 h_1^{(2+\delta)2} E\left|\frac{1}{h_2} K\left(\frac{X_t - x}{h_2}\right)\left(\frac{X_j - x}{h_2}\right)\right|^{2+\delta} \leq C_5 h_1^{(2+\delta)2} h_2^{-(1+\delta)}, \end{aligned} \tag{B.10}$$

where C_3, C_4, and C_5 are constants. Then, by choosing d_n such that $d_n^{\gamma} h_2^{\delta/(2+\delta)} = O(1)$ (so that $d_n h_1 h_2 \to 0$ is satisfied), we have

$$\begin{aligned} \sum_{j=d_n+1}^{n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| &\leq C_6 \sum_{j=d_n+1}^{n} C_3 [\rho(j-1)]^{\delta/(2+\delta)} \left(h_1^{(2+\delta)2} h_2^{-(1+\delta)}\right)^{2/(2+\delta)} \\ &= C_6 h_1^4 h_2^{-(2+2\delta)/(2+\delta)} \sum_{k=d_n}^{n} [\rho(k)]^{\delta/(2+\delta)} \\ &\leq C_6 d_n^{-\gamma} h_1^4 h_2^{-(2+2\delta)/(2+\delta)} \sum_{k=d_n}^{n} k^{\gamma} [\rho(k)]^{\delta/(2+\delta)} = o\left(n h_2^{-1} h_1^{-3}\right), \end{aligned} \tag{B.11}$$

where C_6 is a constant. Then, we obtain

$$\mathrm{Var}\left(\sum_{t=1}^{n} Z_{n,t}\right) = O\left(n h_2^{-1} h_1^{-3}\right). \tag{B.12}$$

With the assumption that $n h_1^3 h_2 \to \infty$, we obtain the result of Lemma B.2.5. ■

Proof of Theorem 3.2.11

Based on the result from Lemma B.2.5, we can observe that under suitable conditions, the covariance of two different error terms can be dominated by the expectation of the squared error, which is the underlying result to prove the consistency and asymptotic normality of modal estimators. Define $\theta = (\alpha_1(x), h_2\alpha_2(x))^T$, $\theta_0 = (\sigma_0^2(x), h_2\dot{\sigma}_0^2(x))^T$, and $X_t^* = (1, (X_t - x)/h_2)^T$, where θ_0 is the true value of the parameter, we have

$$\begin{aligned} Q_n(\theta) &= \frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi\left(\frac{\hat{r}_t - X_t^{*T}\theta}{h_1}\right)K\left(\frac{X_t - x}{h_2}\right) \\ &= \frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi\left(\frac{(Y_t - \hat{m}(X_t))^2 - X_t^{*T}\theta}{h_1}\right)K\left(\frac{X_t - x}{h_2}\right). \end{aligned} \tag{B.13}$$

Note that

$$\begin{aligned} \hat{r}_t &= \{Y_t - \hat{m}(X_t)\}^2 = \{\sigma(X_t)\varepsilon_t + m(X_t) - \hat{m}(X_t)\}^2 \\ &= \sigma^2(X_t)\varepsilon_t^2 + 2\sigma(X_t)\varepsilon_t\{m(X_t) - \hat{m}(X_t)\} + \{m(X_t) - \hat{m}(X_t)\}^2 \\ &= \sigma^2(X_t) + \underbrace{\sigma^2(X_t)(\varepsilon_t^2 - 1)}_{\epsilon_t} + \underbrace{2\sigma(X_t)\varepsilon_t\{m(X_t) - \hat{m}(X_t)\} + \{m(X_t) - \hat{m}(X_t)\}^2}_{U_t} \\ &= \sigma^2(X_t) + \epsilon_t + U_t, \end{aligned} \tag{B.14}$$

we then get

$$Q_n(\theta) = \frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi\left(\frac{\sigma^2(X_t) - X_t^{*T}\theta + \epsilon_t + U_t}{h_1}\right)K\left(\frac{X_t - x}{h_2}\right). \tag{B.15}$$

Define $\delta_n = h_1^2 + h_2^2 + \sqrt{(nh_1^3h_2)^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n(\boldsymbol{\theta}_0 + \delta_n\boldsymbol{\mu}) < Q_n(\boldsymbol{\theta}_0)\right\} \geq 1 - \eta, \tag{B.16}$$

where $\|\cdot\|$ represents the Euclidean distance. The above condition implies that with probability tending to one, there is a local maximum in the ball $\{\boldsymbol{\theta}_0 + \delta_n \boldsymbol{\mu} : \|\mu\| \leq c\}$. Using the Taylor expansion, it follows that

$$
\begin{aligned}
& Q_n\left(\boldsymbol{\theta}_0 + \delta_n \boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\theta}_0\right) \\
=& \frac{1}{nh_1h_2} \sum_{t=1}^{n} \left[\phi\left(\frac{R(X_t) + \epsilon_t - \delta_n \mu^T X_t^*}{h_1} \right) K\left(\frac{X_t - x}{h_2} \right) - \phi\left(\frac{R(X_t) + \epsilon_t}{h_1} \right) K\left(\frac{X_t - x}{h_2} \right) \right] \\
=& \frac{1}{nh_1h_2} \sum_{t=1}^{n} \left[-\phi^{(1)}\left(\frac{R(X_t) + \epsilon_t}{h_1} \right) \left(\frac{\delta_n \mu^T X_t^*}{h_1} \right) K\left(\frac{X_t - x}{h_2} \right) \right. \\
& + \frac{1}{2} \phi^{(2)}\left(\frac{R(X_t) + \epsilon_t}{h_1} \right) \left(\frac{\delta_n \mu^T X_t^*}{h_1} \right)^2 K\left(\frac{X_t - x}{h_2} \right) \\
& \left. - \frac{1}{6} \phi^{(3)}\left(\frac{\epsilon_t^*}{h_1} \right) \left(\frac{\delta_n \mu^T X_t^*}{h_1} \right)^3 K\left(\frac{X_t - x}{h_2} \right) \right] \\
=& I_1 + I_2 + I_3, \qquad \text{(B.17)}
\end{aligned}
$$

where ϵ_i^* is between $R(X_t)+\epsilon_t$ and $R(X_t)+\epsilon_t-\delta_n\mu^T X_t^*$, and $R(X_t) = \sum_{j=2}(\alpha_j(x)/j!)(X_t - x)^j + U_t$. Based on the result $T_n = E\left(T_n\right) + O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is

$$
I_1 = \frac{1}{nh_1h_2} \sum_{t=1}^{n} \left(-\phi^{(1)}\left(\frac{R(X_t) + \epsilon_t}{h_1} \right) \left(\frac{\delta_n \mu^T X_t^*}{h_1} \right) K\left(\frac{X_t - x}{h_2} \right) \right),
$$

by Taylor expansion, we can rewrite it as

$$
\begin{aligned}
E(I_1) &= \frac{-\delta_n}{h_1h_2} E\left(\phi^{(1)}\left(\frac{R(X_t) + \epsilon_t}{h_1} \right) \frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2} \right) \right) \\
&= \frac{-\delta_n}{h_1h_2} E\left(\phi^{(1)}\left(\frac{\epsilon_t}{h_1} \right) \frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2} \right) + \phi^{(2)}\left(\frac{\epsilon_t}{h_1} \right) \frac{R(X_t)\boldsymbol{\mu}^T X_t^*}{h_1^2} K\left(\frac{X_t - x}{h_2} \right) \right. \\
&\qquad \left. + \frac{1}{2} \phi^{(3)}\left(\frac{\epsilon_t^{**}}{h_1} \right) \frac{R^2(X_t)\boldsymbol{\mu}^T X_t^*}{h_1^3} K\left(\frac{X_t - x}{h_2} \right) \right) \\
&= I_{11} + I_{12} + I_{13}, \qquad \text{(B.18)}
\end{aligned}
$$

where ϵ_t^{**} is between ϵ_t and $\epsilon_t + R(X_t)$. Notice that as the order of ϵ_t^{**} is the same as that of ϵ_t, when we do the calculation associated with I_{13}, we instead use ϵ_t directly. By some direct calculations for each part and the results from Lemma B.2.5, we can get

$$I_{11} = \frac{-\delta_n}{h_1 h_2} E\left(\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2}\right)\right) = O_p(\delta_n c h_1^2). \tag{B.19}$$

$$\begin{aligned} I_{12} =& \frac{-\delta_n}{h_1 h_2} E\left(\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2}\right)\frac{R(X_t)}{h_1}\right) \\ =& \frac{-\delta_n}{h_1}\iint \phi(\tau)(\tau^2 - 1)\boldsymbol{\mu}^T X^* g_\epsilon(\tau h_1 \mid x) K(w)\frac{R(X_t)}{h_1} f_X(w h_2 + x) dw d\tau \\ =& O_p(\delta_n c h_2^2). \end{aligned} \tag{B.20}$$

$$\begin{aligned} I_{13} \approx& \frac{-\delta_n}{h_1 h_2} E\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\frac{R^2(X_t)\boldsymbol{\mu}^T X_t^*}{h_1^3} K\left(\frac{X_t - x}{h_2}\right)\right) \\ \leq& \frac{-\delta_n h_2^4}{2}\iint \phi(\tau)(3\tau - \tau^3)\frac{(\alpha^{(2)}(u))^2 \boldsymbol{\mu}^T X^*}{4h_1^3} g_\epsilon(\tau h_1 \mid x) K(w) w^4 \\ & f_X(w h_2 + x) dw d\tau\{1 + o_p(1)\} = o_p(\delta_n h_2^2). \end{aligned} \tag{B.21}$$

Meanwhile, with the condition $h_2^2/h_1 \to 0$ held and the results from Lemma B.2.5, we obtain

$$\frac{\delta_n^2}{h_1^2 h_2^2} E\left(\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2}\right)\right)^2 = O_p(\delta_n^2 c^2 (h_1^3 h_2)^{-1}). \tag{B.22}$$

$$\begin{aligned} & \frac{\delta_n^2}{h_1^2 h_2^2} E\left(\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\boldsymbol{\mu}^T X_t^*}{h_1} K\left(\frac{X_t - x}{h_2}\right)\frac{R(X_t)}{h_1}\right)^2 \\ \leq& \frac{\delta_n^2 h_2^3}{h_1^5}\iint \phi^2(\tau)(\tau^2 - 1)^2 (\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1 \mid x) w^4 K^2(w)\frac{(\alpha^{(2)}(u))^2}{4} \\ & f_X(w h_2 + u) dw d\tau\{1 + o_p(1)\} \\ =& o_p(\delta_n^2 (h_1^3 h_2)^{-1}). \end{aligned} \tag{B.23}$$

The above equations show that $I_1 = O_p(\delta_n c(h_1^2 + h_2^2)) + O_p(\sqrt{\delta_n^2 c^2 (n h_1^3 h_2)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_2 = \frac{1}{nh_1h_2}\sum_{t=1}^{n}\left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_t+R(X_t)}{h_1}\right)\left(\frac{\delta_n\mu^T X_t^*}{h_1}\right)^2 K\left(\frac{X_t-x}{h_2}\right)\right)$, we can rewrite it as

$$\begin{aligned}
E(I_2) =& \frac{\delta_n^2}{2h_2h_1}E\left(\phi^{(2)}\left(\frac{\epsilon_t+R(X_t)}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_t^*)^2}{h_1^2}K\left(\frac{X_t-x}{h_2}\right)\right)\\
=& \frac{\delta_n^2}{2h_2h_1}E\Bigg(\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_t^*)^2}{h_1^2}K\left(\frac{X_t-x}{h_2}\right)\\
&+\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\frac{R(X_t)(\boldsymbol{\mu}^T X_t^*)^2}{h_1^3}K\left(\frac{X_t-x}{h_2}\right)\\
&+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_t^{**}}{h_1}\right)\frac{R^2(X_t)(\boldsymbol{\mu}^T X_t^*)^2}{h_1^4}K\left(\frac{X_t-x}{h_2}\right)\Bigg)\\
=& I_{21}+I_{22}+I_{23},
\end{aligned} \tag{B.24}$$

where ϵ_t^{**} is between ϵ_t and $\epsilon_t + R(X_t)$. Notice that as the order of ϵ_t^{**} is the same as that of ϵ_t, when we do the calculation associated with I_{23}, we instead use ϵ_t directly. By some calculations for each part, we can get

$$\begin{aligned}
I_{21} =& \frac{\delta_n^2}{2h_2h_1}E\left(\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_t^*)^2}{h_1^2}K\left(\frac{X_t-x}{h_2}\right)\right)\\
=& \frac{\delta_n^2}{2h_2h_1}\iint\phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T X^*)^2}{h_1^2}g_\epsilon(\epsilon\mid X)K\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
=& \frac{\delta_n^2}{2h_1^2}\iint\phi(\tau)(\tau^2-1)(\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1\mid x)K(w)f_X(wh_2+x)dwd\tau\\
=& O_p((\delta_n c)^2).
\end{aligned} \tag{B.25}$$

$$\begin{aligned}
I_{22} =& \frac{\delta_n^2}{2h_2h_1}E\left(\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\frac{R(X_t)(\boldsymbol{\mu}^T X_t^*)^2}{h_1^3}K\left(\frac{X_t-x}{h_2}\right)\right)\\
=& \frac{\delta_n^2}{2h_2h_1}\iint\phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{R(X)(\boldsymbol{\mu}^T X^*)^2}{h_1^3}g_\epsilon(\epsilon\mid X)K\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
\leq& \frac{\delta_n^2h_2^2}{2h_1^3}\iint\phi(\tau)(3\tau-\tau^3)\frac{\alpha^{(2)}(u)}{2}(\boldsymbol{\mu}^T X^*)^2 g_\epsilon(\tau h_1\mid x)w^2K(w)\\
& f_X(wh_2+x)dwd\tau\{1+o_p(1)\} = o_p((\delta_n c)^2).
\end{aligned} \tag{B.26}$$

Meanwhile, we can prove that $I_{23} = o_p((\delta_n c)^2)$ as well. Following the same steps in (i), we obtain the following result

$$\begin{aligned}
&\frac{\delta_n^4}{4h_2^2h_1^2}E\left(\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)\frac{(\boldsymbol{\mu}^T X_t^*)^2}{h_1^2}K\left(\frac{X_t-x}{h_2}\right)\right)^2\\
=&\frac{\delta_n^4}{4h_2^2h_1^2}\iint \phi^{(2)2}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T X^*)^4}{h_1^4}g_\epsilon(\epsilon \mid X)K^2\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
=&\frac{\delta_n^4}{4h_2h_1^2}\iint \phi^2(\tau)(\tau^2-1)^2\frac{(\boldsymbol{\mu}^T X^*)^4}{h_1^4}g_\epsilon(\tau h_1 \mid x)K^2(w)f_X(wh_2+x)dwd\tau\\
=&O_p((\delta_n c)^4(h_2h_1^5)^{-1}).
\end{aligned} \tag{B.27}$$

With the condition $nh_1^5h_2 \to \infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) The same way to calculate the third part. As the order of ϵ_t^* is the same as the order of ϵ_t, which indicates that we can obtain $I_3 \approx \frac{1}{nh_1h_2}\sum_{t=1}^n\left(-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\left(\frac{\delta_n\mu^T X_t^*}{h_1}\right)^3 K\left(\frac{X_t-x}{h_2}\right)\right)$. By direct calculation, we can get

$$\begin{aligned}
&\frac{\delta_n^3}{6h_2h_1}E\left(\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\left(\boldsymbol{\mu^T X_t^*}\right)^3}{h_1^3}K\left(\frac{X_t-x}{h_2}\right)\right)\\
=&\frac{\delta_n^3}{6h_2h_1}\iint \phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu^T X^*}\right)^3}{h_1^3}g_\epsilon(\epsilon \mid X)K\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
=&\frac{\delta_n^3}{6}\iint \phi(\tau)(3\tau-\tau^3)\frac{\left(\boldsymbol{\mu^T X^*}\right)^3}{h_1^3}g_\epsilon(\tau h_1 \mid x)K(w)f_X(wh_2+x)dwd\tau = O_p(\delta_n^3).
\end{aligned} \tag{B.28}$$

$$\begin{aligned}
&\frac{\delta_n^6}{36h_2^2h_1^2}E\left(\phi^{(3)}\left(\frac{\epsilon_t}{h_1}\right)\frac{\left(\boldsymbol{\mu^T X_t^*}\right)^3}{h_1^3}K\left(\frac{X_i-x}{h_2}\right)\right)^2\\
=&\frac{\delta_n^6}{36h_2^2h_1^2}\iint \phi^{(3)2}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu^T X^*}\right)^6}{h_1^6}g_\epsilon(\epsilon \mid X)K^2\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
=&\frac{\delta_n^6}{36h_2h_1}\iint \phi^2(\tau)(3\tau-\tau^3)^2\frac{\left(\boldsymbol{\mu^T X^*}\right)^6}{h_1^6}g_\epsilon(\tau h_1 \mid x)K^2(w)f_X(wh_2+x)dwd\tau\\
=&O_p(\delta_n^6(h_2h_1^7)^{-1}).
\end{aligned} \tag{B.29}$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, thus $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_0+\delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_0\right)\} \geq 1-\eta$ holds.

Proof of Theorem 3.2.12

Following the same steps as proving Theorem 3.2.11, recall that

$$Q_n(\theta) = \frac{1}{nh_1h_2}\sum_{t=1}^{n}\phi\left(\frac{\sigma^2(X_t) - X_t^{*T}\theta + \epsilon_t + U_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right). \tag{B.30}$$

Define $\hat{\theta} = (\hat{\sigma}^2(x), h_2\hat{\dot{\sigma}}^2(x))$, then it must satisfy the following equation

$$-\frac{1}{nh_1^2h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t + \tilde{R}(X_t)}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^* = 0, \tag{B.31}$$

where

$$\tilde{R}(X_t) = \sum_{j=2}(\alpha_j(x)/j!)(X_t-x)^j + U_t - X_t^{*T}(\hat{\theta}-\theta_0) = R(X_t) - X_t^{*T}(\hat{\theta}-\theta_0).$$

We can then rewrite the above equation as

$$-\frac{1}{nh_1^2h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t + R(X_t) - X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^* = 0. \tag{B.32}$$

By taking Taylor expansion, we can obtain

$$\begin{aligned}
&-\frac{1}{nh_1^2h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^* \\
&+\frac{1}{nh_1^3h_2}\sum_{t=1}^{n}\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*(R(X_t) - X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)) \\
&-\frac{1}{nh_1^4h_2}\sum_{t=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_t^*}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\left(R(X_t) - X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^2 = 0,
\end{aligned} \tag{B.33}$$

where $\tilde{\epsilon}_t^*$ is between ϵ_t and $\epsilon_t + R(X_t) - X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$. From Theorem 3.2.11, we know $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(\delta_n)$, which indicates that

$$\sup_{t:|X_t-x|/h_2\leq 1}|R(X_t)-X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)|\leq \sup_{t:|X_t-x|/h_2\leq 1}\{|R(X_t)|+|X_i^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)|\}$$

$$=O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|)=O_p(\delta_n). \tag{B.34}$$

Combining this with the equations in the Proof of Theorem 3.2.11, we can see that the third part which is associated with $\boldsymbol{X}_t^*\left(R(X_t)-X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^2$ is dominated by the second part which is associated with $\boldsymbol{X}_t^*\left(R(X_t)-X_t^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)$. We then mainly focus on the first two parts of the left side of the above equation.

Considering $-\frac{1}{nh_1^2h_2}\sum_{t=1}^n\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*+\frac{1}{nh_1^3h_2}\sum_{t=1}^n\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*R(X_t)$, with the assumption that $h/h_2\to 0$, by some direct calculations, we can obtain

$$\begin{aligned}
&E\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^n\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*+\frac{1}{nh_1^3h_2}\sum_{t=1}^n\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*R(X_t)\right)\\
=&-\frac{1}{h_1^2h_2}\iint\phi^{(1)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*g_\epsilon(\epsilon\mid X)K\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
&+\frac{1}{h_1^3h_2}\iint\phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*g_\epsilon(\epsilon\mid X)K\left(\frac{X-x}{h_2}\right)R(X)f_X(X)d\epsilon dX\\
=&\frac{1}{h_1}\iint\phi(\tau)\tau\boldsymbol{X}^*g_\epsilon(\tau h_1\mid x)K(w)f_X(wh_2+x)dwd\tau\\
&-\frac{1}{h_1^2}\iint\phi(\tau)(\tau^2-1)\boldsymbol{X}^*g_\epsilon(\tau h_1\mid x)K(w)R(X)f_X(wh_2+x)dwd\tau\\
=&\frac{h_1^2}{6}f_X(x)\begin{bmatrix}\mu_0g_\epsilon^{(3)}(0\mid x)\\ \mu_1g_\epsilon^{(3)}(0\mid x)\end{bmatrix}-\left(\frac{h_2^2\ddot{\sigma}^2(x)}{2}f_X(x)\begin{bmatrix}\mu_2g_\epsilon^{(2)}(0\mid x)\\ \mu_3g_\epsilon^{(2)}(0\mid x)\end{bmatrix}\right)\{1+o_p(1)\},
\end{aligned}\tag{B.35}$$

where $\int\tau^4\phi(\tau)d\tau=3$, $\int\tau^2\phi(\tau)d\tau=1$, and $\int w^jK(w)dw=\mu_j$.

Considering $\frac{1}{nh_1^3h_2}\sum_{t=1}^n\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*X_t^{*T}$, by direct calculation, we have

$$\begin{aligned}
&E\left(\frac{1}{nh_1^3h_2}\sum_{t=1}^n\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\boldsymbol{X}_t^{*T}\right)\\
=&E\left(\frac{1}{h_1^3h_2}\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\boldsymbol{X}_t^{*T}\right)
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{h_1^3 h_2}\iint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon\mid X)K\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX\\
=&\frac{1}{h_1^2}\iint \phi\left(\tau\right)(\tau^2-1)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon\mid x)K\left(w\right)f_X(wh_2+x)dwd\tau(1+o_p(1))\\
=&f_X(x)\begin{bmatrix}\mu_0 g_\epsilon^{(2)}(0\mid x) & \mu_1 g_\epsilon^{(2)}(0\mid x)\\ \mu_1 g_\epsilon^{(2)}(0\mid x) & \mu_2 g_\epsilon^{(2)}(0\mid x)\end{bmatrix}.
\end{aligned}
\tag{B.36}
$$

Based on the above two equations, we can achieve

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0=&\begin{bmatrix}\mu_0 g_\epsilon^{(2)}(0\mid x) & \mu_1 g_\epsilon^{(2)}(0\mid x)\\ \mu_1 g_\epsilon^{(2)}(0\mid x) & \mu_2 g_\epsilon^{(2)}(0\mid x)\end{bmatrix}^{-1}\\
&\left(\frac{h_1^2}{6}f_X(x)\begin{bmatrix}\mu_0 g_\epsilon^{(3)}(0\mid x)\\ \mu_1 g_\epsilon^{(3)}(0\mid x)\end{bmatrix}-\left(\frac{h_2^2\ddot{\sigma}^2(x)}{2}f_X(x)\begin{bmatrix}\mu_2 g_\epsilon^{(2)}(0\mid x)\\ \mu_3 g_\epsilon^{(2)}(0\mid x)\end{bmatrix}\right)\{1+o_p(1)\}\right).
\end{aligned}
\tag{B.37}
$$

Meanwhile, with the condition $h_2^2/h_1\to 0$ held, combining the results obtained from Lemma B.2.5, we can obtain

$$
\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^n\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\right.\\
&\left.+\frac{1}{nh_1^3h_2}\sum_{t=1}^n\phi^{(2)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*R(X_t)\right)\\
=&E\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^n\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\right)\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^n\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\boldsymbol{X}_t^*\right)^T\\
&(1+o_p(1))\\
=&\frac{1}{nh_1^4h_2^2}\iint\phi^{(1)2}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}g_\epsilon(\epsilon\mid X)K^2\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX(1+o_p(1))\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_1^3h_2}f_X(x)\begin{bmatrix}v_0 g_\epsilon(0\mid x) & v_1 g_\epsilon(0\mid x)\\ v_1 g_\epsilon(0\mid x) & v_2 g_\epsilon(0\mid x)\end{bmatrix}(1+o_p(1)),
\end{aligned}
\tag{B.38}
$$

where $v_j=\int w^jK^2(w)dw$. Then, we can obtain

$$\mathrm{Var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = \frac{\int \tau^2 \phi^2(\tau) d\tau}{n h_1^3 h_2 f_X(x)} \begin{bmatrix} \mu_0 g_\epsilon^{(2)}(0 \mid x) & \mu_1 g_\epsilon^{(2)}(0 \mid x) \\ \mu_1 g_\epsilon^{(2)}(0 \mid x) & \mu_2 g_\epsilon^{(2)}(0 \mid x) \end{bmatrix}^{-1}$$

$$\begin{bmatrix} v_0 g_\epsilon(0 \mid x) & v_1 g_\epsilon(0 \mid x) \\ v_1 g_\epsilon(0 \mid x) & v_2 g_\epsilon(0 \mid x) \end{bmatrix} \begin{bmatrix} \mu_0 g_\epsilon^{(2)}(0 \mid x) & \mu_1 g_\epsilon^{(2)}(0 \mid x) \\ \mu_1 g_\epsilon^{(2)}(0 \mid x) & \mu_2 g_\epsilon^{(2)}(0 \mid x) \end{bmatrix}^{-1} (1 + o_p(1)). \tag{B.39}$$

Define $W_n = \frac{1}{n h_1^2 h_2} \sum_{t=1}^n \phi^{(1)} \left(\frac{\epsilon_t}{h_1} \right) K \left(\frac{X_t - x}{h_2} \right) \boldsymbol{X}_t^*$. To show Theorem 3.3.14, it is sufficient to show that

$$T_n = \sqrt{n h_2 h_1^3} W_n \xrightarrow{d} \mathcal{N}(0, T), \tag{B.40}$$

where $T = \int \tau^2 \phi^2(\tau) d\tau f_X(x) \begin{bmatrix} v_0 g_\epsilon(0 \mid x) & v_1 g_\epsilon(0 \mid x) \\ v_1 g_\epsilon(0 \mid x) & v_2 g_\epsilon(0 \mid x) \end{bmatrix}$.

By Slutsky's theorem and the above two equations, we can obtain Theorem 3.3.14.

To show the above equation, we prove that for any unit vector $\boldsymbol{d} \in R^2$,

$$\left\{ \boldsymbol{d}^T \operatorname{Cov}(T_n) \boldsymbol{d} \right\}^{-1/2} \left\{ \boldsymbol{d}^T T_n - \boldsymbol{d}^T E(T_n) \right\} \xrightarrow{d} N(0, 1). \tag{B.41}$$

Then, we check Lyapunov's condition. Let

$$\xi_i = \sqrt{h_2 h_1^3 / n} K \left(\frac{X_t - x}{h_2} \right) \frac{1}{h_1 h_2} \phi^{(1)} \left(\frac{\epsilon_t}{h_1} \right) \boldsymbol{d}^T \boldsymbol{X}_t^*, \tag{B.42}$$

we need to prove $n E|\xi_1|^3 \to 0$. As $\left(\boldsymbol{d}^T \boldsymbol{X}_t^* \right)^2 \leq \|\boldsymbol{d}\|^2 \|\boldsymbol{X}_t^*\|^2$, $\phi^{(1)}(\cdot)$ is bounded, and $K(\cdot)$ has compact support, we have

$$n E|\xi|^3 \leq O\left(n n^{-3/2} h_2^{-3/2} h_1^{3/2} \right) E \left| K^3 \left(\frac{X_t - x}{h_2} \right) \phi^{(1)3} \left(\frac{\epsilon_t}{h_1} \right) \boldsymbol{d}^T \boldsymbol{X}_t^* \right| \to 0. \tag{B.43}$$

Thus, the asymptotic normality for T_n holds.

Lemma B.2.6 *Let* $R(X_t) = \sigma^2(X_t) - \sigma^2(x) - \dot{\sigma}^2(x)(X_t - x)$ *and* $\epsilon_t = [Y_t - \hat{m}(X_t)]^2 - \sigma^2(X_t)$. *Under the conditions C1, C2, the first part of C4, and D1-D5, we have*

$$\sum_{t=1}^{n} \phi_{h_1}^{(2)}(\epsilon_t) K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l = nh_2^{l+1} E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x) f_X(x) \mu_l \left(1 + o_p(1)\right), \text{ and}$$

$$\sum_{t=1}^{n} \phi_{h_1}^{(2)}(\epsilon_t) R(X_t) K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l$$
$$= \frac{nh_2^{l+3}}{2} E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x) \ddot{\sigma}^2(x) f_X(x) \mu_{l+2}(1 + o_p(1)).$$

Proof. *The main steps are consistent with the ones in Lemma B.2.5 besides we treat bandwidth* h_1 *as a constant. We focus on the proof for the first equation, while the second one can be proved by the same arguments. Define* $Z_{n,t} = \phi_{h_1}^{(2)}(\epsilon_t) K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l$. *By calculating, we can have*

$$E(Z_{n,1}) = h_2^{l+1} E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x) f_X(x) \mu_l (1 + o_p(1)). \tag{B.44}$$

Therefore,

$$E\left(\sum_{t=1}^{n} \phi_{h_1}^{(2)}(\epsilon_t) K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l\right) = nh_2^{l+1} E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x) f_X(x) \mu_l \left(1 + o_p(1)\right).$$

Note that

$$\sum_{t=1}^{n} Z_{n,t} = E\left(\sum_{t=1}^{n} Z_{n,t}\right) + O_p\left(\sqrt{\operatorname{Var}\left(\sum_{t=1}^{n} Z_{n,t}\right)}\right), \tag{B.45}$$

and the stationary of $\{\epsilon_t\}$ *gives*

$$\operatorname{Var}\left(\sum_{t=1}^{n} Z_{n,t}\right) = nEZ_{n,1}^2 + 2\sum_{j=2}^{n}(n - j + 1)\operatorname{Cov}\left(Z_{n,t}, Z_{n,1}\right). \tag{B.46}$$

In addition, we can get

$$E(Z_{n,1}^2) = h_2^{2l+1} E(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x)^2 f_X(x) \int K^2(u) u^{2l} du (1 + o(1)) = O(h_2^{2l+1}). \tag{B.47}$$

To obtain an upper bound for the second term on the right-hand side of the above equation, let d_n *be a sequence of positive integers satisfying* $d_n \to \infty$ *and* $d_n h_2 \to 0$ *as* $n \to \infty$*, we can split it into two terms*

$$\sum_{j=2}^{n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| = \sum_{j=2}^{d_n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| + \sum_{j=d_n+1}^{n} |\mathrm{Cov}\left(Z_{n,t}, Z_{n,j}\right)|. \tag{B.48}$$

By calculating, we can show that

$$|EZ_{n,t}Z_{n,j}| \leq E|Z_{n,t}Z_{n,j}| = E \mid E\left[\phi_{h_1}^{(2)}\left(\epsilon_t\right)\phi_{h_1}^{(2)}\left(\epsilon_j\right) \mid X_t, X_j\right] K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l$$

$$K\left(\frac{X_j - x}{h_2}\right)(X_j - x)^l \mid \leq C h_2^{2l+2}, \tag{B.49}$$

where C *is a constant. Therefore, we have* $\sum_{j=2}^{d_n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| \leq C h_2^{2l+2} \sum_{j=2}^{d_n} 1 = o\left(n h_2^{2l+1}\right)$*. By using Davydov's inequality, we obtain*

$$|\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| \leq C[\alpha(j-1)]^{\delta/(2+\delta)} \left(E|Z_{n,1}|^{2+\delta}\right)^{2/(2+\delta)}, \tag{B.50}$$

and

$$E|Z_{n,t}|^{2+\delta} = E|E\left[\phi^{(2)}\left(\epsilon_t\right) \mid X_t\right] K\left(\frac{X_t - x}{h_2}\right)(X_t - x)^l|^{2+\delta} \leq C h_2^{(2+\delta)l+1}. \tag{B.51}$$

Then, by choosing d_n *such that* $d_n^{\gamma} h_2^{\delta/(2+\delta)} = O(1)$*, we have*

$$\begin{aligned} \sum_{j=d_n+1}^{n} |\mathrm{Cov}\left(Z_{n,1}, Z_{n,j}\right)| &\leq C \sum_{j=d_n+1}^{n} [\alpha(j-1)]^{\delta/(2+\delta)} \left(h_2^{(2+\delta)l+1}\right)^{2/(2+\delta)} \\ &\leq C d_n^{-a} h_2^{2l+2/(2+\delta)} \sum_{k=d_n}^{n} k^{\gamma} [\rho(k)]^{\delta/(2+\delta)} = o\left(n h_2^{2l+1}\right). \end{aligned} \tag{B.52}$$

Thus, $\mathrm{Var}\left(\sum_{i=1}^{n} Z_{n,i}\right) = O\left(n h_2^{2l+1}\right)$ *and we complete the proof.* ■

Lemma B.2.7 *Under the conditions C1, C2, the first part of C4, and D1-D5, we have*

$$\frac{1}{\sqrt{nh_2}}\begin{pmatrix} \sum_{t=1}^{n}\phi_{h_1}^{(1)}(\epsilon_t)K\left(\frac{X_t-x}{h_2}\right) \\ \sum_{t=1}^{n}\phi_{h_1}^{(1)}(\epsilon_t)K\left(\frac{X_t-x}{h_2}\right)\frac{X_t-x}{h_2} \end{pmatrix} \xrightarrow{D}$$
$$N\left(0, f_X(x)E((\phi_{h_1}^{(1)}(\epsilon))^2 \mid X=x)\begin{pmatrix} v_0 & v_1 \\ v_1 & v_2 \end{pmatrix}\right).$$

Proof. Let $W_n = \sum_{t=1}^{n} W_{n,t} = \sum_{t=1}^{n}\phi_{h_1}^{(1)}(\epsilon_t)K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix} 1 \\ \frac{X_t-x}{h_2}\end{pmatrix}$. Then, we have $E(W_n)=0$ and $\mathrm{Var}(W_n) = \mathrm{Var}\left(\sum_{t=1}^{n} W_{n,t}\right) = nEW_{n,1}^2 + 2\sum_{j=2}^{n}(n-j+1)\,\mathrm{Cov}\left(W_{n,t}, W_{n,j}\right)$. Following the lines of arguments as in Lemma B.2.6, we can obtain $\mathrm{Var}(W_n)$. ■

Proof of Theorem 3.2.13

At first, we have

$$\hat{r}(X_t) = [Y_t - \hat{m}(X_t)]^2 = \sigma^2(X_t) + \epsilon_t = \sigma^2(X_t) - \sigma^2(x) - \dot{\sigma}^2(x)(X_t - x)$$
$$+ \sigma^2(x) + \dot{\sigma}^2(x)(X_t - x) + \epsilon_t = \hat{\sigma}^2(x) + \hat{\dot{\sigma}}^2(x)(X_t - x) + \epsilon_t + \hat{\eta}_t, \tag{B.53}$$

where $\hat{\eta}_t = R(X_t) - (\hat{\sigma}^2(x) - \sigma^2(x)) - (\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x))(X_t - x)$. Then, taking the derivative of the objective function leads

$$\sum_{t=1}^{n}\phi_{h_1}^{(1)}(\epsilon_t + \hat{\eta}_t)K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix} 1 \\ \frac{X_t-x}{h_2}\end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{B.54}$$

By using Taylor expansion, we can obtain

$$\sum_{t=1}^{n}\left[\phi_{h_1}^{(1)}(\epsilon_t) + \phi_{h_1}^{(2)}(\epsilon_t)\hat{\eta}_t + [\phi_{h_1}^{(1)}(\epsilon_t + \hat{\eta}_t) - \phi_{h_1}^{(1)}(\epsilon_t) - \phi_{h_1}^{(2)}(\epsilon_t)\hat{\eta}_t]\right]$$

$$K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix}1\\ \frac{X_t-x}{h_2}\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix}. \tag{B.55}$$

Considering $\sum_{t=1}^{n}\phi_{h_1}^{(2)}(\epsilon_t)\hat{\eta}_t K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix}1\\ \frac{X_t-x}{h_2}\end{pmatrix}$, by calculating and the results from Lemma B.2.6, we obtain

$$\sum_{t=1}^{n}\phi_{h_1}^{(2)}(\epsilon_t)R(X_t)K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix}1\\ \frac{X_t-x}{h_2}\end{pmatrix}-\sum_{t=1}^{n}\phi_{h_1}^{(2)}(\epsilon_t)K\left(\frac{X_t-x}{h_2}\right)$$
$$\begin{pmatrix}(\hat{\sigma}^2(x)-\sigma^2(x))+(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))(X_t-x)\\ \frac{X_t-x}{h_2}[(\hat{\sigma}^2(x)-\sigma^2(x))+(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))(X_t-x)]\end{pmatrix}$$

$$=\sum_{t=1}^{n}\phi_{h_1}^{(2)}(\epsilon_t)R(X_t)K\left(\frac{X_t-x}{h_2}\right)\begin{pmatrix}1\\ \frac{X_t-x}{h_2}\end{pmatrix}-\sum_{t=1}^{n}\phi_{h_1}^{(2)}(\epsilon_t)K\left(\frac{X_t-x}{h_2}\right)$$
$$\begin{pmatrix}1 & \frac{X_t-x}{h_2}\\ \frac{X_t-x}{h_2} & \frac{(X_t-x)^2}{h_2^2}\end{pmatrix}\begin{pmatrix}\hat{\sigma}^2(x)-\sigma^2(x)\\ h_2(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))\end{pmatrix}$$
$$=\frac{nh_2^3}{2}E(\phi_{h_1}^{(2)}(\epsilon_t)\mid X_t=x)\ddot{\sigma}^2(x)f_X(x)\begin{pmatrix}\mu_2\\ \mu_3\end{pmatrix}(1+o_p(1))-nh_2E(\phi_{h_1}^{(2)}(\epsilon_t)\mid X_t=x)$$
$$f_X(x)\begin{pmatrix}\mu_0 & \mu_1\\ \mu_1 & \mu_2\end{pmatrix}(1+o_p(1))\begin{pmatrix}\hat{\sigma}^2(x)-\sigma^2(x)\\ h_2(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x))\end{pmatrix}. \tag{B.56}$$

Meanwhile, with consistency property (it can be proved by following the arguments in Proof of Theorem 3.2.11), we know that

$$\sup_t|\hat{\eta}_t|=\sup_t\mid R(X_t)-\left(\hat{\sigma}^2(x)-\sigma^2(x)\right)-\left(\hat{\dot{\sigma}}^2(x)-\dot{\sigma}^2(x)\right)(X_t-x)\mid$$

$$\leq \sup_t |R(X_t)| + |\hat{\sigma}^2(x) - \sigma^2(x)| + h_2|\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x)|$$

$$= O_p\left(h_2^2 + \left(\hat{\sigma}^2(x) - \sigma^2(x)\right) + h_2\left(\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x)\right)\right) = o_p(1) \tag{B.57}$$

with $|X_t - x| \leq h_2$. Thus, we do not need to consider the third part of the above Taylor expansion equation. Then, we have

$$\begin{pmatrix} \hat{\sigma}^2(x) - \sigma^2(x) \\ h_2(\hat{\dot{\sigma}}^2(x) - \dot{\sigma}^2(x)) \end{pmatrix} = \frac{1}{nh_2} E^{-1}(\phi_{h_1}^{(2)}(\epsilon_t) \mid X_t = x) f_X^{-1}(x) \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}^{-1} (1 + o_p(1)) W_n$$

$$+ \frac{h_2^2}{2} \ddot{\sigma}^2(x) \begin{pmatrix} \mu_0 & \mu_1 \\ \mu_1 & \mu_2 \end{pmatrix}^{-1} \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} (1 + o_p(1)). \tag{B.58}$$

With the results from Lemma B.2.7 and Slutsky's theorem, we complete the proof.

Proof of Theorem 3.3.15

We follow the lines of the proof in Cheng et al. (2007). Recall that

$$E\left(\hat{\sigma}^2(x)\right) = \sigma^2(x) + \left(\frac{h_2^2}{2}\mu_2 \ddot{\sigma}^2(x) - \frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0 \mid x)}{g_\epsilon^{(2)}(0 \mid x)}\right) + o(h_1^2 + h_2^2 + (nh_2h_1^3)^{-1/2}). \tag{B.59}$$

$$\tilde{\sigma}^2(x) = \frac{r(r-1)}{2}\hat{\sigma}^2(x-(r+1)\beta h_2) + (1-r^2)\hat{\sigma}^2(x-r\beta h_2) + \frac{r(r+1)}{2}\hat{\sigma}^2(x-(r-1)\beta h_2). \tag{B.60}$$

By applying Taylor expansion, we can have

$$\begin{aligned}
&E\left(\tilde{\sigma}^2(x)\right) \\
=&E\left(\frac{r(r-1)}{2}\hat{\sigma}^2(x-(r+1)\beta h_2) + (1-r^2)\hat{\sigma}^2(x-r\beta h_2) + \frac{r(r+1)}{2}\hat{\sigma}^2(x-(r-1)\beta h_2)\right) \\
=&\left(\frac{r(r-1)}{2} + (1-r^2) + \frac{r(r+1)}{2}\right)\left(\sigma^2(x) + \left(\frac{h_2^2}{2}\mu_2\ddot{\sigma}^2(x) - \frac{h_1^2}{2}\frac{g^{(3)}(0 \mid x)}{g^{(2)}(0 \mid x)}\right)\right) \\
&+ \dot{\sigma}^2(x)\frac{r(r-1)}{2}(-(r+1)\beta h_2) + \dot{\sigma}^2(x)(1-r^2)(-r\beta h_2) \\
&+ \dot{\sigma}^2(x)\frac{r(r+1)}{2}(-(r-1)\beta h_2) + \frac{1}{2}\ddot{\sigma}^2(x)\frac{r(r-1)}{2}(-(r+1)\beta h_2)^2
\end{aligned}$$

$$+\frac{1}{2}\ddot{\sigma}^2(x)(1-r^2)(-r\beta h_2)^2+\frac{1}{2}\ddot{\sigma}^2(x)\frac{r(r-1)}{2}(-(r-1)\beta h_2)^2+o(h_1^2+h_2^2+(nh_2h_1^3)^{-1/2}). \tag{B.61}$$

It indicates that

$$E\left(\tilde{\sigma}^2(x)\right)=\sigma^2(x)+\left(\frac{h_2^2}{2}\mu_2\ddot{\sigma}^2(x)-\frac{h_1^2}{2}\frac{g_\epsilon^{(3)}(0\mid x)}{g_\epsilon^{(2)}(0\mid x)}\right)+o(h_1^2+h_2^2+(nh_2h_1^3)^{-1/2}). \tag{B.62}$$

From the Proof of Theorem 3.3.14, we know that the main equation we need to deal with for calculating the variance of $\hat{\sigma}^2(x)$ at point x is

$$\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\right)\\
=&E\left(-\frac{1}{nh_1^2h_2}\sum_{t=1}^{n}\phi^{(1)}\left(\frac{\epsilon_t}{h_1}\right)K\left(\frac{X_t-x}{h_2}\right)\right)^2(1+o_p(1))\\
=&\frac{1}{nh_1^4h_2^2}\iint\phi^{(1)2}\left(\frac{\epsilon}{h_1}\right)g_\epsilon(\epsilon\mid X)K^2\left(\frac{X-x}{h_2}\right)f_X(X)d\epsilon dX(1+o_p(1))\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_1^3h_2}f_X(x)v_0g_\epsilon(0\mid x)(1+o_p(1)).
\end{aligned} \tag{B.63}$$

In terms of the variance of $\tilde{\sigma}^2(x)$, we have

$$\begin{aligned}
&\mathrm{Var}\left(\tilde{\sigma}^2(x)\right)\\
=&\mathrm{Var}\left(\frac{r(r-1)}{2}\hat{\sigma}^2(x-(r+1)\beta h_2)+(1-r^2)\hat{\sigma}^2(x-r\beta h_2)+\frac{r(r+1)}{2}\hat{\sigma}^2(x-(r-1)\beta h_2)\right)\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau f_X(x)}{nh_1^3h_2g_\epsilon^{(2)}(0\mid x)^2}g_\epsilon(0\mid x)\int\left(\frac{r(r-1)}{2}K(s)+(1-r^2)K(s+\beta h_2)\right.\\
&\left.+\frac{r(r+1)}{2}K(s+2\beta h_2)\right)ds\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau f_X(x)}{nh_1^3h_2g_\epsilon^{(2)}(0\mid x)^2}g_\epsilon(0\mid x)\Big\{v_0-\frac{3r^2(1-r^2)}{2}\int K^2(s)ds\\
&+2r^2(1-r^2)\int K(s)K(s+\beta h_2)ds-\frac{r^2(1-r^2)}{2}\int K(s)K(s+2\beta h_2)ds\Big\}\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_1^3h_2g_\epsilon^{(2)}(0\mid x)^2}f_X(x)g_\epsilon(0\mid x)(v_0-r^2(1-r^2)C(\beta)),
\end{aligned} \tag{B.64}$$

where $C(\beta)=1.5C(0,\beta)-2C(0.5,\beta)+0.5C(1,\beta)$ and $C(\beta,t)=\int K(w-t\beta)K(w+t\beta)dw$.

Proof of Theorem 3.5.17

Recall that

$$Q_n(\beta_1(x), \beta_2(x)) = \frac{1}{nh_3h_4} \sum_{t=1}^{n} \phi\left(\frac{\hat{r}_t - \Psi(\beta_1(x) + \beta_2(x)(X_t - x))}{h_3}\right) K\left(\frac{X_t - x}{h_4}\right). \quad \text{(B.65)}$$

Following the notations in Ziegelmann (2002), we define $L(X_t - x, \theta) = \Psi(\beta_1(x) + \beta_2(x)(X_t - x))$ and $L^{(i)}(X_t - x, \theta) = (\partial/\partial(X_t - x))^i L(X_t - x, \theta)$, where $\theta = (\beta_1(x), \beta_2(x))$. By applying Taylor expansion, we can obtain

$$Q_n(\theta) = \frac{1}{nh_3h_4} \sum_{t=1}^{n} \phi\left(\frac{\epsilon_t + R^*(X_t)}{h_3}\right) K\left(\frac{X_t - x}{h_4}\right), \quad \text{(B.66)}$$

where $R^*(X_t) = 2^{-1}\ddot{\sigma}^2(x)(X_t - x)^2 - 2^{-1}L^{(2)}(0, \theta)(X_t - x)^2 + \textit{terms of smaller order}$. Then, we can follow the same steps as Proof of Theorem 3.3.14 to obtain Theorem 3.5.17.

Appendix C

Appendix for Chapter 4

C.1 Identification with Monotonicity

If we impose monotonicity for the modal function $Mode(Y \mid X)$, we can identify CMTE without using the full continuity assumption. In particular, we impose the following conditions.

Assumption 3 *$f_{Y_1}\left(Y \mid X = \bar{X}\right)$ and $f_{Y_0}\left(Y \mid X = \bar{X}\right)$ are strictly positive and unimodal distributions, such that $\sup_{Y:|Y-Y_1^m|>\varepsilon} f_{Y_1}(Y \mid X = \bar{X}) < f_{Y_1}\left(Y_1^m \mid X = \bar{X}\right)$ and $\sup_{Y:|Y-Y_0^m|>\varepsilon} f_{Y_0}(Y \mid X = \bar{X}) < f_{Y_0}\left(Y_0^m \mid X = \bar{X}\right)$ for all $\varepsilon > 0$.*

Assumption 4 *(i) $X \mapsto Mode(Y_1 \mid X)$ and $X \mapsto Mode(Y_0 \mid X)$ are monotone in some neighborhood of $\bar{X}$; (ii) $Mode(Y_1 \mid X = \bar{X}) \geq Mode(Y_0 \mid X = \bar{X})$ in the non-decreasing case or $Mode(Y_1 \mid X = \bar{X}) \leq Mode(Y_0 \mid X = \bar{X})$ in the non-increasing case.*

Assumption 5 *Under Assumption 4, $X \mapsto Mode(Y_1 \mid X)$ is right-continuous at the cutoff $\bar{X}$ and $X \mapsto Mode(Y_0 \mid X)$ is left-continuous at the cutoff $\bar{X}$.*

Assumption 3 is nearly identical to Assumption 1 except for the continuity, which is imposed to ensure that both distributions of Y_1 and Y_0 are unimodal in the presence of X. It guarantees the existence of the modal estimator around the cutoff $\bar{X}$. Assumption 4 assures that all limits exist at the discontinuity point and necessitates local responsiveness to treatment at the cutoff. Assumption 5, which is imposed on the conditional modes of potential outcomes, is weaker than the assumption of full continuity at the cutoff. It demonstrates that continuity at the cutoff of both conditional modal functions is not required for identification and that under monotonicity restriction, the one-sided continuity is both necessary and sufficient for identification. We then have the following lemma.

Lemma C.1.8 *Under the model settings in the paper and Assumptions 3-5, by defining* $m_{Y_1}(\bar{X}) = \lim_{X \downarrow \bar{X}} m_{Y_1}(X)$ *and* $m_{Y_0}(\bar{X}) = \lim_{X \uparrow \bar{X}} m_{Y_0}(X)$*, the conditional mode effect of the treatment on the outcome at the cutoff can be identified as*

$$\begin{aligned}\tau_{RD} &= Mode\left(Y_1 \mid \bar{X}\right) - Mode\left(Y_0 \mid \bar{X}\right) = \lim_{X \downarrow \bar{X}} Mode\left(Y_1 \mid X\right) - \lim_{X \uparrow \bar{X}} Mode\left(Y_0 \mid X\right) \\ &= \lim_{X \downarrow \bar{X}} Mode(Y \mid X) - \lim_{X \uparrow \bar{X}} Mode(Y \mid X) = m_{Y_1}(\bar{X}) - m_{Y_0}(\bar{X}),\end{aligned}$$

where the second equation follows under Assumption 5, the third equation is a consequence of $Y = Y_1 D + Y_0(1 - D)$ *and* $D = \mathbf{1}\left(X \geq \bar{X}\right)$*, and the fourth one is from Assumptions 4 and 5.*

Proof. To prove the above lemma, we only need to show the fourth equation. Because the treatment of non-increasing and non-decreasing cases is similar, we only discuss the former. Under the first part of Assumption 4, we have $Mode(Y_1 \mid X = \bar{X}) \leq \lim_{X \downarrow \bar{X}} Mode\left(Y_1 \mid X\right) = \lim_{X \downarrow \bar{X}} Mode(Y \mid X)$ and $\lim_{X \uparrow \bar{X}} Mode(Y \mid X) = \lim_{X \uparrow \bar{X}} Mode$

$(Y_0 \mid X) \leq Mode(Y_0 \mid X = \bar{X})$. Then, with the second part of Assumption 4, we obtain $\lim_{X\uparrow\bar{X}} Mode(Y \mid X) \leq Mode(Y_0 \mid X = \bar{X}) \leq Mode(Y_1 \mid X = \bar{X}) \leq \lim_{X\downarrow\bar{X}} Mode(Y \mid X)$. Finally, τ_{RD} is defined as the difference between the right and left limits of the conditional modal regression functions evaluated at the cutoff $\bar{X}$ by virtue of continuity under Assumption 5. ■

C.2 Asymptotic Properties of $\hat{m}_{Y_0}(x)$ and $\hat{m}^{(1)}_{Y_0}(x)$

Theorem C.2.21 *Under the regularity conditions C1-C4, with probability approaching one, as* $n_- \to \infty, h_{1,-} \to 0, h_{2,-} \to 0$, $h^2_{2,-}/h_{1,-} \to 0$, *and* $n_- h_{2,-} h^5_{1,-} \to \infty$, *there exist consistent maximizers* $(\hat{m}_{Y_0}(x), h_{2,-}\hat{m}^{(1)}_{Y_0}(x))$ *of* (4.11) *such that*

i. $|\hat{m}_{Y_0}(x) - m_{Y_0}(x)| = O_p\left(\left(n_- h_{2,-} h^3_{1,-}\right)^{-1/2} + h^2_{1,-} + h^2_{2,-}\right)$,

ii. $|h_{2,-}(\hat{m}^{(1)}_{Y_0}(x) - m^{(1)}_{Y_0}(x))| = O_p\left(\left(n_- h_{2,-} h^3_{1,-}\right)^{-1/2} + h^2_{1,-} + h^2_{2,-}\right)$.

Theorem C.2.22 *With* $n_- h^5_{2,-} h^3_{1,-} = O(1)$ *and* $n_- h_{2,-} h^7_{1,-} = O(1)$, *under the same conditions as Theorem C.2.21, the parameters satisfying the consistency results in Theorem C.2.21 have the following asymptotic result*

$$\sqrt{n_- h_{2,-} h^3_{1,-}}\left(\begin{bmatrix} \hat{m}_{Y_0}(x) - m_{Y_0}(x) \\ h_{2,-}(\hat{m}^{(1)}_{Y_0}(x) - m^{(1)}_{Y_0}(x)) \end{bmatrix} - \hat{\Gamma}^{-1}\left(\frac{h^2_{2,-}}{2} m^{(2)}_{Y_0}(x)\hat{\Lambda}_2 - \frac{h^2_{1,-}}{2}\frac{g^{(3)}_{\epsilon_-}(0 \mid x)}{g^{(2)}_{\epsilon_-}(0 \mid x)}\hat{\Lambda}_1\right)\right)$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{g_{\epsilon_-}(0 \mid x)\int \tau^2\phi^2(\tau)d\tau}{\left(g^{(2)}_{\epsilon_-}(0 \mid x)\right)^2 f_{X_-}(x)}\hat{\Gamma}^{-1}\hat{\Sigma}\hat{\Gamma}^{-1}\right).$$

If we allow $n_- h^5_{2,-} h^3_{1,-} \to 0$ *and* $n_- h_{2,-} h^7_{1,-} \to 0$, *the asymptotic theorem becomes*

$$\sqrt{n_- h_{2,-} h^3_{1,-}}\begin{pmatrix} \hat{m}_{Y_0}(x) - m_{Y_0}(x) \\ h_{2,-}(\hat{m}^{(1)}_{Y_0}(x) - m^{(1)}_{Y_0}(x)) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \frac{g_{\epsilon_-}(0 \mid x)\int \tau^2\phi^2(\tau)d\tau}{\left(g^{(2)}_{\epsilon_-}(0 \mid x)\right)^2 f_{X_-}(x)}\hat{\Gamma}^{-1}\hat{\Sigma}\hat{\Gamma}^{-1}\right).$$

where $\hat{\Gamma} = \begin{bmatrix} \int_{-M}^{\bar{c}} K(w)dw & \int_{-M}^{\bar{c}} wK(w)dw \\ \int_{-M}^{\bar{c}} wK(w)dw & \int_{-M}^{\bar{c}} w^2K(w)dw \end{bmatrix}$, $\hat{\Lambda}_1 = \begin{bmatrix} \int_{-M}^{\bar{c}} K(w)dw \\ \int_{-M}^{\bar{c}} wK(w)dw \end{bmatrix}$,

$\hat{\Lambda}_2 = \begin{bmatrix} \int_{-M}^{\bar{c}} w^2K(w)dw \\ \int_{-M}^{\bar{c}} w^3K(w)dw \end{bmatrix}$, *and* $\hat{\Sigma} = \begin{bmatrix} \int_{-M}^{\bar{c}} K^2(w)dw & \int_{-M}^{\bar{c}} wK^2(w)dw \\ \int_{-M}^{\bar{c}} wK^2(w)dw & \int_{-M}^{\bar{c}} w^2K^2(w)dw \end{bmatrix}$.

The proofs of the aforementioned two theorems can be obtained directly by following the procedures for proving Theorems 4.2.18 and 4.2.19, which are omitted in this paper for space reasons.

C.3 Modal Inference for $\tau_{RD}^{(1)}$

Theorem C.3.23 *Under the regularity conditions C1-C4, with* $n_+h_{2,+}^5h_{1,+}^3 = O(1)$, $n_+h_{2,+}h_{1,+}^7 = O(1)$, $n_-h_{2,-}^5h_{1,-}^3 = O(1)$, *and* $n_-h_{2,-}h_{1,-}^7 = O(1)$, *as both* $n_+ \to \infty$ *and* $n_- \to \infty$, *we have*

$$\frac{\hat{\tau}_{RD}^{(1)} - \tau_{RD}^{(1)} - Bias(\hat{\tau}_{RD}^{(1)})}{\sqrt{Var(\hat{\tau}_{RD}^{(1)})}} \xrightarrow{d} \mathcal{N}(0,1).$$

If we allow $n_+h_{2,+}^5h_{1,+}^3 \to 0$, $n_+h_{2,+}h_{1,+}^7 \to 0$, $n_-h_{2,-}^5h_{1,-}^3 \to 0$, *and* $n_-h_{2,-}h_{1,-}^7 \to 0$, *as both* $n_+ \to \infty$ *and* $n_- \to \infty$, *we have*

$$(Var(\hat{\tau}_{RD}^{(1)}))^{-1/2}(\hat{\tau}_{RD}^{(1)} - \tau_{RD}) \xrightarrow{d} \mathcal{N}(0,1),$$

where $Bias(\hat{\tau}_{RD}^{(1)}) = \left\{ \frac{h_{2,+}^2}{2} m_{Y_1}^{(2)}(\bar{X}) \frac{\mu_{+,0}\mu_{+,3} - \mu_{+,1}\mu_{+,2}}{\mu_{+,0}\mu_{+,2} - \mu_{+,1}^2} - \frac{h_{2,-}^2}{2} m_{Y_0}^{(2)}(\bar{X}) \frac{\mu_{-,0}\mu_{-,3} - \mu_{-,1}\mu_{-,2}}{\mu_{-,0}\mu_{-,2} - \mu_{-,1}^2} \right\}$

$(1 + o_p(1))$ *and* $Var(\hat{\tau}_{RD}^{(1)}) = Var(\hat{m}_{Y_1}^{(1)}(\bar{X})) + Var(\hat{m}_{Y_0}^{(1)}(\bar{X}))$

$$= \left\{ \frac{\int \tau^2\phi^2(\tau)d\tau f_{X_+}^{-1}(\bar{X})}{n_+h_{2,+}h_{1,+}^3} \frac{g_{\epsilon_+}(0 \mid \bar{X})}{(g_{\epsilon_+}^{(2)}(0 \mid \bar{X}))^2} \frac{\mu_{+,1}^2v_{+,0} - 2\mu_{+,1}\mu_{+,0}v_{+,1} + \mu_{+,0}^2v_{+,2}}{(\mu_{+,0}\mu_{+,2} - \mu_{+,1}^2)^2} \right.$$

$$\left. + \frac{\int \tau^2\phi^2(\tau)d\tau f_{X_-}^{-1}(\bar{X})}{n_-h_{2,-}h_{1,-}^3} \frac{g_{\epsilon_-}(0 \mid \bar{X})}{(g_{\epsilon_-}^{(2)}(0 \mid \bar{X}))^2} \frac{\mu_{-,1}^2v_{-,0} - 2\mu_{-,1}\mu_{-,0}v_{-,1} + \mu_{-,0}^2v_{-,2}}{(\mu_{-,0}\mu_{-,2} - \mu_{-,1}^2)^2} \right\}$$

$$(1 + o_p(1)).$$

C.4 Monte Carlo Experiment

It is observed that when the data are symmetrically distributed, the modal and mean regression lines are identical to one another. However, little is known about the behavior of the modal-based robust regression at the boundary point, which is a building block of the modal-based robust RD estimator. We in this section numerically show that we can utilize the proposed modal RD regression to estimate mean treatment effects, and have some efficiency gain against mean regression when the data are non-normal or have outliers.

To illustrate the applicability of the proposed modal regression on the symmetric case where the modal RD design is identical to the mean RD design (mainly focus on fuzzy RD designs), we generate the random samples from the following DGP

$$Y_i = m(X_i) + D_i\tau + X_i\epsilon_i,\ i = 1, \cdots, n,$$

where $m(X_i) = X_i + X_i^2$, $D_i = \mathbf{1}\left(X_i \geq \bar{X}_i\right)$, $\tau = 1$, and X_i follows the uniform distribution on $[-2, 2]$. The sample size considered for this experiment is $n \in \{200, 400, 600, 1000\}$. To show the superiority of the modal-based estimation, we set (1) $\epsilon_i \sim$ Laplace with $\mu = 0$ and $\sigma = 1$; (2) $\epsilon_i \sim t$ distribution with 3 degrees of freedom; and (3) $\epsilon_i \sim$ mixture normal $0.9N(0, 1) + 0.1N(0, 9)$. The model has a jump at $\bar{X}_i = 0.5$ which is assumed to be known in advance. Thus, we have the conditional modal/mean function

$$Mode(Y_i \mid X_i) = E(Y_i \mid X_i) = X_i + X_i^2 + \mathbf{1}\left(X_i \geq 0.5\right)\tau,$$

where $\lim_{X\uparrow 0.5} Mode(Y_0 \mid X_i = \bar{X}_i) = \lim_{X\uparrow 0.5} E(Y_0 \mid X_i = \bar{X}_i) = 0.75$ and $\lim_{X\downarrow 0.5} Mode(Y_1 \mid X_i = \lim_{X\downarrow 0.5} E(Y_1 \mid X_i = \bar{X}_i) = \bar{X}_i) = 1.75$. We can then have the direct causal effect of interest $\tau_{RD} = \tau_{mean} = 1$. For easy illustration, we set all bandwidths associated with h_2

in mean and modal regression equal, which are calculated by R package *rdrobust*. For the bandwidths associated with h_1 in modal regression, we use the rule of thumb to set them as $1.05\sigma_X n^{-1/5}$ in which σ_X is the standard deviation of variable X. For each data set, a total of 200 simulation relocations are conducted.

Table C.1 presents the simulation results of the studied estimators, where the bold number indicates the smaller value of the results obtained from the mean and modal-based regressions, and the values in the brackets represent standard errors. The results are in qualitative agreement with the theoretical intuition and show that the modal-based estimation produces highly accurate/efficient estimates and outperforms the local linear mean regression in all three error distributions under consideration.

Table C.1: Results of Simulations

Sample Size	Mean Treatment Effect (SE)	MSE	Modal-Based (SE)	MSE
		$\underline{L_p(0,1)}$		
n=200	1.0643 (0.3461)	0.1233	1.0836 (**0.2920**)	**0.0918**
n=400	1.0536 (0.2480)	0.0641	1.0448 (**0.2161**)	**0.0485**
n=600	1.0364 (0.1918)	0.0379	1.0216 (**0.1587**)	**0.0255**
n=1000	1.0222 (0.1620)	0.0266	1.0223 (**0.1290**)	**0.0171**
		$\underline{t(3)}$		
n=200	1.0331 (0.4311)	0.1860	1.0974 (**0.3902**)	**0.1610**
n=400	1.0232 (0.3153)	0.0995	1.0119 (**0.2757**)	**0.0758**
n=600	1.0342 (0.2526)	0.0647	1.0260 (**0.2332**)	**0.0548**
n=1000	1.0079 (0.1780)	0.0316	1.0188 (**0.1594**)	**0.0256**
		$\underline{0.9\text{N}(0,1)+0.1\text{N}(0,9)}$		
n=200	1.1121 (0.8105)	0.6662	1.0770 (**0.4021**)	**0.1668**
n=400	1.0905 (0.5607)	0.3210	1.0469 (**0.3012**)	**0.0925**
n=600	0.9507 (0.4522)	0.2059	1.0097 (**0.2494**)	**0.0620**
n=1000	1.0482 (0.3647)	0.1347	1.0474 (**0.1975**)	**0.0410**

The visual results for one set of simulated observations according to different values of sample size are presented in Figure C.1, which shows that both mean and modal-based

regression lines can capture the true RD regression lines. When we have symmetric data with outliers or a heavy-tailed distribution, the modal regression line is the same as the mean regression line, but modal-based estimation can give superior or more efficient estimators.

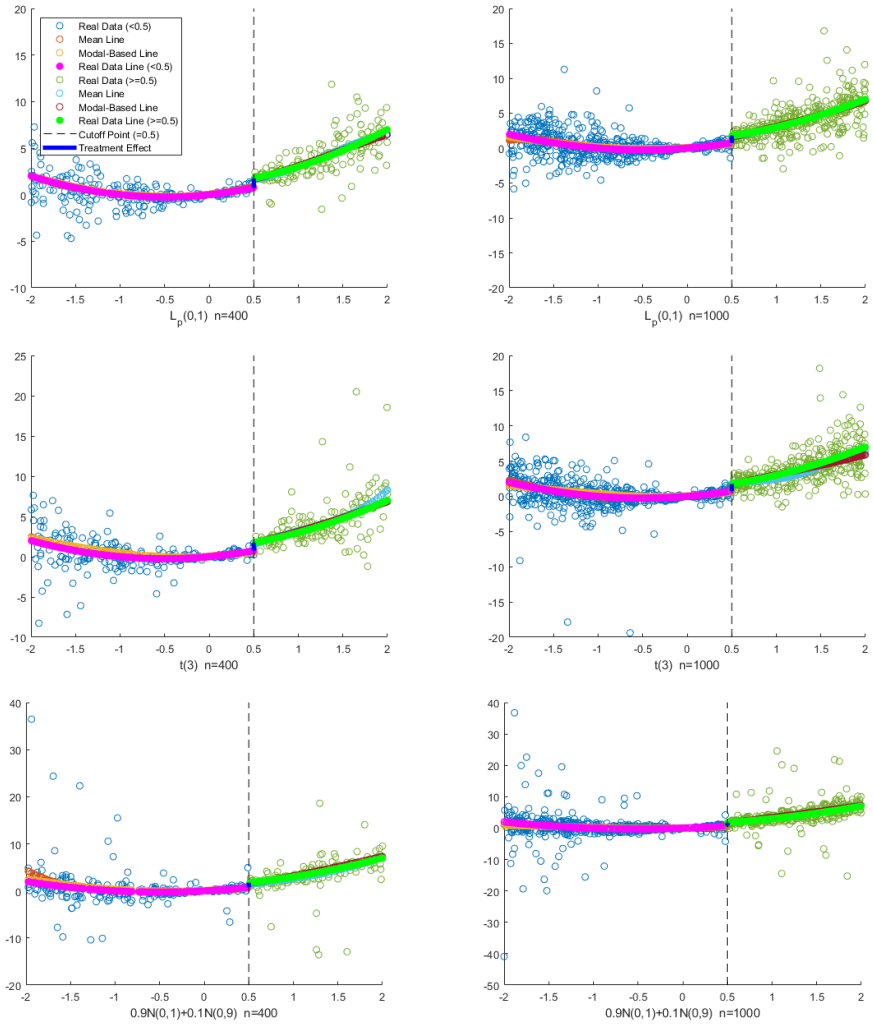

Figure C.1: Visual Results for One Set of Simulated Observations

C.5 Technical Proofs

Proof of Lemma 4.2.2

We can easily prove the lemma on the basis of the definition of derivative. Notice that we already know

$$\begin{aligned} Mode(Y_0 \mid \bar{X}) &= \lim_{\varepsilon \to 0} Mode\left(Y_0 \mid \bar{X} - \varepsilon\right) = \lim_{\varepsilon \to 0} Mode\left(Y_0 \mid \bar{X} - \varepsilon, D = 0\right) \\ &= \lim_{\varepsilon \to 0} Mode\left(Y \mid \bar{X} - \varepsilon\right) = m_{Y_0}(\bar{X}) \end{aligned} \tag{C.1}$$

when ε is sufficiently small. Based on Assumption 2, since $Mode(Y_0 \mid X)$ is continuously differentiable in a neighborhood of the cutoff $\bar{X}$, $Mode^{(1)}(Y_0 \mid \bar{X})$ will equal its own one-sided derivative. We therefore have

$$\begin{aligned} Mode^{(1)}(Y_0 \mid \bar{X}) &= \lim_{\varepsilon \to 0}(Mode(Y_0 \mid \bar{X} - \varepsilon) - Mode(Y_0 \mid \bar{X}))/\varepsilon \\ &= \lim_{\varepsilon \to 0}(Mode(Y_0 \mid \bar{X} - \varepsilon, D = 0) - Mode(Y \mid \bar{X}))/\varepsilon \\ &= \lim_{\varepsilon \to 0}(Mode(Y \mid \bar{X} - \varepsilon) - Mode(Y \mid \bar{X}))/\varepsilon = m_{Y_0}^{(1)}(\bar{X}). \end{aligned} \tag{C.2}$$

Similar results are obtained for $m_{Y_1}^{(1)}(\bar{X})$. We then prove the lemma.

Proof of Theorem 4.2.18

Define $\theta = (a_+, h_{2,+}b_+)^T$, $\theta_0 = (m_{Y_1}(x), h_{2,+}m_{Y_1}^{(1)}(x))^T$, $X_{+,i}^* = (1, (X_{+,i} - x)/h_{2,+})^T$, where θ_0 is the true value of the parameter, we have

$$Q_{n_+}(\theta) = \frac{1}{n_+ h_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi\left(\frac{Y_{1,i} - X_{+,i}^{*T}\theta}{h_{1,+}}\right) K\left(\frac{X_{+,i} - x}{h_{2,+}}\right). \tag{C.3}$$

Define $\delta_{n_+} = h_{1,+}^2 + h_{2,+}^2 + \sqrt{(n_+ h_{1,+}^3 h_{2,+})^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_{n_+}\left(\boldsymbol{\theta}_0+\delta_{n_+}\boldsymbol{\mu}\right)<Q_{n_+}\left(\boldsymbol{\theta}_0\right)\right\}\geq 1-\eta, \tag{C.4}$$

where $\|\cdot\|$ represents the Euclidean distance. (C.4) implies that with a probability tending to one, there is a local maximum in the ball $\{\boldsymbol{\theta}_0+\delta_{n_+}\boldsymbol{\mu}:\|\mu\|\leq c\}$. Using Taylor expansion, it follows that

$$\begin{aligned}
&Q_{n_+}\left(\boldsymbol{\theta}_0+\delta_{n_+}\boldsymbol{\mu}\right)-Q_{n_+}\left(\boldsymbol{\theta}_0\right)\\
=&\frac{1}{n_+h_{1,+}h_{2,+}}\sum_{i=1}^{n_+}\left[\phi\left(\frac{R(X_{+,i})+\epsilon_{+,i}-\delta_{n_+}\mu^TX^*_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right.\\
&\left.-\phi\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right]\\
=&\frac{1}{n_+h_{1,+}h_{2,+}}\sum_{i=1}^{n_+}\left[-\phi^{(1)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^TX^*_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right.\\
&+\frac{1}{2}\phi^{(2)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^TX^*_{+,i}}{h_{1,+}}\right)^2K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\\
&\left.-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon^*_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^TX^*_{+,i}}{h_{1,+}}\right)^3K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right]\\
=&I_1+I_2+I_3,
\end{aligned}\tag{C.5}$$

where $\epsilon^*_{+,i}$ is between $R(X_{+,i})+\epsilon_{+,i}$ and $R(X_{+,i})+\epsilon_{+,i}-\delta_{n_+}\mu^TX^*_{+,i}$, and $R(X_{+,i})=\sum_{j=2}(m^{(j)}_{Y_1}(x)\ /j!)(X_{+,i}-x)^j$. Based on the result $T_{n_+}=E\left(T_{n_+}\right)+O_p(\sqrt{\mathrm{Var}\left(T_{n_+}\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, $I_1=\frac{1}{n_+h_{1,+}h_{2,+}}\sum_{i=1}^{n_+}\Big(-\phi^{(1)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^TX^*_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\Big)$, by Taylor expansion, we can rewrite it as

$$\begin{aligned}
E(I_1)&=\frac{-\delta_{n_+}}{h_{1,+}h_{2,+}}E\left(\phi^{(1)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\mu^TX^*_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)\\
&=\frac{-\delta_{n_+}}{h_{1,+}h_{2,+}}E\left(\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\mu^TX^*_{+,i}}{h_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&+\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{R(X_{+,i})\boldsymbol{\mu}^T X^*_{+,i}}{h^2_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\\
&+\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon^{**}_{+,i}}{h_{1,+}}\right)\frac{R^2(X_{+,i})\boldsymbol{\mu}^T X^*_{+,i}}{h^3_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\Bigg)\\
&=I_{11}+I_{12}+I_{13},
\end{aligned}
\tag{C.6}
$$

where $\epsilon^{**}_{+,i}$ is between $\epsilon_{+,i}$ and $\epsilon_{+,i}+R(X_{+,i})$. Notice that as the order of $\epsilon^{**}_{+,i}$ is the same as that of $\epsilon_{+,i}$, when we do the calculations associated with I_{13}, we instead use $\epsilon_{+,i}$ directly. By some direct calculations for each part, we can get

$$
I_{11}=\frac{-\delta_{n_+}}{h_{1,+}h_{2,+}}E\left(\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\boldsymbol{\mu}^T X^*_{+,i}}{h_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)=O_p(\delta_{n_+}ch^2_{1,+}).
\tag{C.7}
$$

$$
\begin{aligned}
I_{12}=&\frac{-\delta_{n_+}}{h_{1,+}h_{2,+}}E\left(\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\boldsymbol{\mu}^T X^*_{+,i}}{h_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\frac{R(X_{+,i})}{h_{1,+}}\right)\\
=&\frac{-\delta_{n_+}}{h_{1,+}}\iint\phi(\tau)(\tau^2-1)\boldsymbol{\mu}^T X^*_+ g_{\epsilon_+}(\tau h_{1,+}|x)K(w)\frac{R(X_+)}{h_{1,+}}f_{X_+}(wh_{2,+}+x)dwd\tau\\
=&O_p(\delta_{n_+}ch^2_{2,+}).
\end{aligned}
\tag{C.8}
$$

$$
\begin{aligned}
I_{13}\approx&\frac{-\delta_{n_+}}{h_{1,+}h_{2,+}}E\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{R^2(X_{+,i})\boldsymbol{\mu}^T X^*_{+,i}}{h^3_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)\\
\le&\frac{-\delta_{n_+}h^4_{2,+}}{2}\iint\phi(\tau)(3\tau-\tau^3)\frac{(m^{(2)}_{Y_1}(x))^2\boldsymbol{\mu}^T X^*_+}{4h^3_{1,+}}g_{\epsilon_+}(\tau h_{1,+}|x)K(w)w^4 f_{X_+}(wh_{2,+}+x)\\
&dwd\tau\{1+o_p(1)\}\\
=&o_p(\delta_{n_+}ch^2_{2,+}).
\end{aligned}
\tag{C.9}
$$

Meanwhile, with the condition $h^2_{2,+}/h_{1,+}\to 0$ held, we obtain

$$
\frac{\delta^2_{n_+}}{h^2_{1,+}h^2_{2,+}}E\left(\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\boldsymbol{\mu}^T X^*_{+,i}}{h_{1,+}}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)^2=O_p(\delta^2_{n_+}c^2(h^3_{1,+}h_{2,+})^{-1}).
\tag{C.10}
$$

$$\frac{\delta_{n_+}^2}{h_{1,+}^2 h_{2,+}^2} E\left(\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\boldsymbol{\mu}^T X_{+,i}^*}{h_{1,+}} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\frac{R(X_{+,i})}{h_{1,+}}\right)^2$$

$$\leq \frac{\delta_{n_+}^2 h_{2,+}^3}{h_{1,+}^5}\iint \phi^2(\tau)(\tau^2-1)^2(\boldsymbol{\mu}^T X_+^*)^2 g_{\epsilon_+}(\tau h_{1,+}|x) w^4 K^2(w)\frac{(m_{Y_1}^{(2)}(x))^2}{4} f_{X_+}(wh_{2,+}+x)$$

$$dwd\tau\{1+o_p(1)\} = o_p(\delta_{n_+}^2 c^2 (h_{1,+}^3 h_{2,+})^{-1}). \tag{C.11}$$

The above equations show that $I_1 = O_p(\delta_{n_+} c(h_{1,+}^2 + h_{2,+}^2)) + O_p(\sqrt{\delta_{n_+}^2 c^2 (n_+ h_{1,+}^3 h_{2,+})^{-1}}) = O_p(\delta_{n_+}^2 c)$.

(ii) For the second part, $I_2 = \frac{1}{n_+ h_{1,+} h_{2,+}} \sum_{i=1}^{n_+}\left(\frac{1}{2}\phi^{(2)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^T X_{+,i}^*}{h_{1,+}}\right)^2 K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)$, we can rewrite it as

$$\begin{aligned}
E(I_2) =& \frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}} E\left(\phi^{(2)}\left(\frac{R(X_{+,i})+\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^T X_{+,i}^*}{h_{1,+}}\right)^2 K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)\\
=& \frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}} E\Bigg(\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^2} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\\
&+\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{R(X_{+,i})(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^3} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\\
&+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_{+,i}^{**}}{h_{1,+}}\right)\frac{R^2(X_{+,i})(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^4} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\Bigg)\\
=& I_{21}+I_{22}+I_{23},
\end{aligned} \tag{C.12}$$

where $\epsilon_{+,i}^{**}$ lies between $\epsilon_{+,i}$ and $\epsilon_t + R(X_{+,i})$. Notice that as the order of $\epsilon_{+,i}^{**}$ is the same as that of $\epsilon_{+,i}$, when we do the calculations associated with I_{23}, we instead use $\epsilon_{+,i}$ directly. By some calculations for each part, we can get

$$\begin{aligned}
I_{21} =& \frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}} E\left(\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^2} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)\\
=& \frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}}\iint \phi^{(2)}\left(\frac{\epsilon_+}{h_{1,+}}\right)\frac{(\boldsymbol{\mu}^T X_+^*)^2}{h_{1,+}^2} g_{\epsilon_+}(\epsilon_+|X_+) K\left(\frac{X_+-x}{h_{2,+}}\right) f_{X_+}(X_+) d\epsilon_+ dX_+
\end{aligned}$$

$$
\begin{aligned}
=&\frac{\delta_{n_+}^2}{2h_{1,+}^2}\iint \phi(\tau)(\tau^2-1)(\boldsymbol{\mu}^T X_+^*)^2 g_{\epsilon_+}(\tau h_{1,+}|x)K(w)f_{X_+}(wh_{2,+}+x)dwd\tau \\
=&O_p((\delta_{n_+}c)^2). \qquad \text{(C.13)}
\end{aligned}
$$

$$
\begin{aligned}
I_{22} =&\frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}}E\left(\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{R(X_{+,i})(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^3}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right) \\
=&\frac{\delta_{n_+}^2}{2h_{2,+}h_{1,+}}\iint \phi^{(3)}\left(\frac{\epsilon_+}{h_{1,+}}\right)\frac{R(X_+)(\boldsymbol{\mu}^T X_+^*)^2}{h_{1,+}^3}g_{\epsilon_+}(\epsilon_+|X_+)K\left(\frac{X_+-x}{h_{2,+}}\right)f_{X_+}(X_+)d\epsilon_+ dX_+ \\
\leq&\frac{\delta_{n_+}^2 h_{2,+}^2}{2h_{1,+}^3}\iint \phi(\tau)(3\tau-\tau^3)\frac{m_{Y_1}^{(2)}(x)}{2}(\boldsymbol{\mu}^T X_+^*)^2 g_{\epsilon_+}(\tau h_{1,+}|x)w^2K(w)f_{X_+}(wh_{2,+}+x)dwd\tau \\
&\{1+o_p(1)\} \\
=&o_p((\delta_{n_+}c)^2). \qquad \text{(C.14)}
\end{aligned}
$$

Meanwhile, we can prove that $I_{23} = o_p((\delta_{n_+}c)^2)$ as well. Following the same steps in (i), we obtain the following result

$$
\begin{aligned}
&\frac{\delta_{n_+}^4}{4h_{2,+}^2h_{1,+}^2}E\left(\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{(\boldsymbol{\mu}^T X_{+,i}^*)^2}{h_{1,+}^2}K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)^2 \\
=&\frac{\delta_{n_+}^4}{4h_{2,+}^2h_{1,+}^2}\iint \phi^{(2)2}\left(\frac{\epsilon_+}{h_{1,+}}\right)\frac{(\boldsymbol{\mu}^T X_+^*)^4}{h_{1,+}^4}g_{\epsilon_+}(\epsilon_+|X_+)K^2\left(\frac{X_+-x}{h_{2,+}}\right)f_{X_+}(X_+)d\epsilon_+ dX_+ \\
=&\frac{\delta_{n_+}^4}{4h_{2,+}h_{1,+}^2}\iint \phi^2(\tau)(\tau^2-1)^2\frac{(\boldsymbol{\mu}^T X_+^*)^4}{h_{1,+}^4}g_{\epsilon_+}(\tau h_{1,+}|x)K^2(w)f_{X_+}(wh_{2,+}+x)dwd\tau \\
=&O_p((\delta_{n_+}c)^4(h_{2,+}h_{1,+}^5)^{-1}). \qquad \text{(C.15)}
\end{aligned}
$$

With the condition $n_+h_{1,+}^5h_{2,+}\to\infty$ held, the above equations indicate that the second part will dominate the first part when we choose c big enough.

(iii) The same way is used to calculate the third part. As the order of $\epsilon_{+,i}^*$ is the same as the order of $\epsilon_{+,i}$, we can obtain $I_3 \approx \frac{1}{n_+h_{1,+}h_{2,+}}\sum_{i=1}^{n_+}\left(-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\left(\frac{\delta_{n_+}\mu^T X_{+,i}^*}{h_{1,+}}\right)^3 K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)$. By direct calculations, we can get

$$
\begin{aligned}
&\frac{\delta_{n_+}^3}{6h_{2,+}h_{1,+}} E\left(\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\left(\boldsymbol{\mu}^T X_{+,i}^*\right)^3}{h_{1,+}^3} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)\\
=&\frac{\delta_{n_+}^3}{6h_{2,+}h_{1,+}}\iint \phi^{(3)}\left(\frac{\epsilon_{+}}{h_{1,+}}\right)\frac{\left(\boldsymbol{\mu}^T X_{+}^*\right)^3}{h_{1,+}^3} g_{\epsilon_+}(\epsilon_+|X_+) K\left(\frac{X_+-x}{h_{2,+}}\right) f_{X_+}(X_+) d\epsilon_+ dX_+\\
=&\frac{\delta_{n_+}^3}{6}\iint \phi(\tau)(3\tau-\tau^3)\frac{\left(\boldsymbol{\mu}^T X_{+}^*\right)^3}{h_{1,+}^3} g_{\epsilon_+}(\tau h_{1,+}|x) K(w) f_{X_+}(wh_{2,+}+x) dw d\tau\\
=&O_p(\delta_{n_+}^3).
\end{aligned}
\tag{C.16}
$$

$$
\begin{aligned}
&\frac{\delta_{n_+}^6}{36h_{2,+}^2h_{1,+}^2} E\left(\phi^{(3)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)\frac{\left(\boldsymbol{\mu}^T X_{+,i}^*\right)^3}{h_{1,+}^3} K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\right)^2\\
=&\frac{\delta_{n_+}^6}{36h_{2,+}^2h_{1,+}^2}\iint \phi^{(3)2}\left(\frac{\epsilon_{+}}{h_{1,+}}\right)\frac{\left(\boldsymbol{\mu}^T X_{+}^*\right)^6}{h_{1,+}^6} g_{\epsilon_+}(\epsilon_+|X_+) K^2\left(\frac{X_+-x}{h_{2,+}}\right) f_{X_+}(X_+) d\epsilon_+ dX_+\\
=&\frac{\delta_{n_+}^6}{36h_{2,+}h_{1,+}}\iint \phi^2(\tau)(3\tau-\tau^3)^2\frac{\left(\boldsymbol{\mu}^T X_{+}^*\right)^6}{h_{1,+}^6} g_{\epsilon_+}(\tau h_{1,+}|x) K^2(w) f_{X_+}(wh_{2,+}+x) dw d\tau\\
=&O_p(\delta_{n_+}^6(h_{2,+}h_{1,+}^7)^{-1}).
\end{aligned}
\tag{C.17}
$$

These indicate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that I_2 dominates both I_1 and I_3 with probability $1-\eta$. Because the second term is negative, $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_{n_+}\left(\boldsymbol{\theta}_0+\delta_{n_+}\boldsymbol{\mu}\right) < Q_{n_+}\left(\boldsymbol{\theta}_0\right)\} \geq 1-\eta$ holds.

Proof of Theorem 4.2.19

Following the same steps as proving Theorem 4.2.18, recall that

$$
Q_{n_+}(\theta) = \frac{1}{nh_{1_+}h_{2_+}}\sum_{i=1}^{n_+}\phi\left(\frac{Y_1 - X_{+,i}^{*T}\theta}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right). \tag{C.18}
$$

Define $\hat{\theta} = (\hat{m}_{Y_1}(x), h_{2,+}\hat{m}_{Y_1}^{(1)}(x))$, it must satisfy the following equation

$$
-\frac{1}{n_+h_{1,+}^2h_{2,+}}\sum_{i=1}^{n_+}\phi^{(1)}\left(\frac{\epsilon_{+,i}+\tilde{R}(X_{+,i})}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^* = 0, \tag{C.19}
$$

where $\tilde{R}(X_{+,i}) = \sum_{j=2}(m_{Y_1}^{(j)}(x)/j!)(X_{+,i}-x)^j - X_{+,i}^{*T}(\hat{\theta}-\theta_0) = R(X_{+,i}) - X_{+,i}^{*T}(\hat{\theta}-\theta_0)$. We can then rewrite (C.19) as

$$-\frac{1}{n_+h_{1,+}^2h_{2,+}}\sum_{i=1}^{n_+}\phi^{(1)}\left(\frac{\epsilon_{+,i}+R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}=0. \quad \text{(C.20)}$$

By taking Taylor expansion, we can obtain

$$\begin{aligned}
&-\frac{1}{n_+h_{1,+}^2h_{2,+}}\sum_{i=1}^{n_+}\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}\\
&+\frac{1}{n_+h_{1,+}^3h_{2,+}}\sum_{i=1}^{n_+}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}(R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0))\\
&-\frac{1}{n_+h_{1,+}^4h_{2,+}}\sum_{i=1}^{n_+}\phi^{(3)}\left(\frac{\tilde{\epsilon}_{+,i}^{*}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}\left(R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^2=0,
\end{aligned} \quad \text{(C.21)}$$

where $\tilde{\epsilon}_{+,i}^{*}$ is between $\epsilon_{+,i}$ and $\epsilon_{+,i}+R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)$. From Theorem 4.2.18, we know $\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(\delta_{n_+})$, which indicates that

$$\begin{aligned}
\sup_{i:|X_{+,i}-x|/h_{2,+}\le 1}|R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)| &\le \sup_{i:|X_{+,i}-x|/h_{2,+}\le 1}\{|R(X_{+,i})|+|X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)|\}\\
&= O_p(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|) = O_p(\delta_{n_+}).
\end{aligned} \quad \text{(C.22)}$$

Combining this with the results in the Proof of Theorem 4.2.18, we can see that the third part which is associated with $\boldsymbol{X}_{+,i}^{*}\left(R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)^2$ is dominated by the second part which is associated with $\boldsymbol{X}_{+,i}^{*}\left(R(X_{+,i})-X_{+,i}^{*T}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right)$. We then mainly focus on the first two parts of the left side of (C.19).

Considering $-\frac{1}{n_+h_{1,+}^2h_{2,+}}\sum_{i=1}^{n_+}\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}+\frac{1}{n_+h_{1,+}^3h_{2,+}}\sum_{i=1}^{n_+}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}R(X_{+,i})$, by some direct calculations, we can obtain

$$E\Bigg\{-\frac{1}{n_+h_{1,+}^2h_{2,+}}\sum_{i=1}^{n_+}\phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^{*}$$

$$
\begin{aligned}
&+\frac{1}{n_+h_{1,+}^3h_{2,+}}\sum_{i=1}^{n_+}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^*R(X_{+,i})\Bigg\}\\
=&-\frac{1}{h_{1,+}^2h_{2,+}}\iint\phi^{(1)}\left(\frac{\epsilon_+}{h_{1,+}}\right)\boldsymbol{X}_+^*g_{\epsilon_+}(\epsilon_+|X_+)K\left(\frac{X_+-x}{h_{2,+}}\right)f_{X_+}(X_+)d\epsilon_+dX_+\\
&+\frac{1}{h_{1,+}^3h_{2,+}}\iint\phi^{(2)}\left(\frac{\epsilon_+}{h_{1,+}}\right)\boldsymbol{X}_+^*g_{\epsilon_+}(\epsilon_+|X_+)K\left(\frac{X_+-x}{h_{2,+}}\right)R(X_+)f_{X_+}(X_+)d\epsilon_+dX_+\\
=&\frac{1}{h_{1,+}}\iint\phi(\tau)\tau\boldsymbol{X}_+^*g_{\epsilon_+}(\tau h_{1,+}|x)K(w)f_{X_+}(wh_{2,+}+x)dwd\tau\\
&-\frac{1}{h_{1,+}^2}\iint\phi(\tau)(\tau^2-1)\boldsymbol{X}_+^*g_{\epsilon_+}(\tau h_{1,+}|x)K(w)R(X_+)f_{X_+}(wh_{2,+}+x)dwd\tau\\
=&\left\{\frac{h_{1,+}^2}{6}f_{X_+}(x)\begin{bmatrix}\mu_{+,0}g_{\epsilon_+}^{(3)}(0|x)\\ \mu_{+,1}g_{\epsilon_+}^{(3)}(0|x)\end{bmatrix}-\left(\frac{h_{2,+}^2m_{Y_1}^{(2)}(x)}{2}f_{X_+}(x)\begin{bmatrix}\mu_{+,2}g_{\epsilon_+}^{(2)}(0|x)\\ \mu_{+,3}g_{\epsilon_+}^{(2)}(0|x)\end{bmatrix}\right)\right\}\{1+o_p(1)\},
\end{aligned}
\tag{C.23}
$$

where $\int\tau^4\phi(\tau)d\tau=3$, $\int\tau^2\phi(\tau)d\tau=1$, and $\int_{-\bar{x}}^{M}w^jK(w)dw=\mu_{+,j}$ for $j=0,1,2,3$.

Considering $\frac{1}{n_+h_{1,+}^3h_{2,+}}\sum_{i=1}^{n_+}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^*X_{+,i}^{*T}$, by direct calculations, we have

$$
\begin{aligned}
&E\left(\frac{1}{n_+h_{1,+}^3h_{2,+}}\sum_{i=1}^{n_+}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^*X_{+,i}^{*T}\right)\\
=&E\left(\frac{1}{h_{1,+}^3h_{2,+}}\phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right)K\left(\frac{X_{+,i}-x}{h_{2,+}}\right)\boldsymbol{X}_{+,i}^*X_{+,i}^{*T}\right)\\
=&\frac{1}{h_{1,+}^3h_{2,+}}\iint\phi^{(2)}\left(\frac{\epsilon_+}{h_{1,+}}\right)\boldsymbol{X}_+^*\boldsymbol{X}_+^{*T}g_{\epsilon_+}(\epsilon_+|X_+)K\left(\frac{X_+-x}{h_{2,+}}\right)f_{X_+}(X_+)d\epsilon_+dX_+\\
=&\frac{1}{h_{1,+}^2}\iint\phi(\tau)(\tau^2-1)\boldsymbol{X}_+^*\boldsymbol{X}_+^{*T}g_{\epsilon_+}(\epsilon_+|x)K(w)f_{X_+}(wh_{2,+}+x)dwd\tau(1+o_p(1))\\
=&f_{X_+}(x)\begin{bmatrix}\mu_{+,0}g_{\epsilon_+}^{(2)}(0|x) & \mu_{+,1}g_{\epsilon_+}^{(2)}(0|x)\\ \mu_{+,1}g_{\epsilon_+}^{(2)}(0|x) & \mu_{+,2}g_{\epsilon_+}^{(2)}(0|x)\end{bmatrix}.
\end{aligned}
\tag{C.24}
$$

Based on the above two equations (C.23) and (C.24), we can achieve

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \begin{bmatrix} \mu_{+,0} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) \\ \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,2} g^{(2)}_{\epsilon_+}(0|x) \end{bmatrix}^{-1}$$
$$\left(\frac{h^2_{1,+}}{6} f_{X_+}(x) \begin{bmatrix} \mu_{+,0} g^{(3)}_{\epsilon_+}(0|x) \\ \mu_{+,1} g^{(3)}_{\epsilon_+}(0|x) \end{bmatrix} - \left(\frac{h^2_{2,+} m^{(2)}_{Y_1}(x)}{2} f_{X_+}(x) \begin{bmatrix} \mu_{+,2} g^{(2)}_{\epsilon_+}(0|x) \\ \mu_{3,+} g^{(2)}_{\epsilon_+}(0|x) \end{bmatrix} \right) \right)$$
$$\{1 + o_p(1)\}. \tag{C.25}$$

Meanwhile, with the condition $h^2_{2,+}/h_{1,+} \to 0$ held, we can obtain

$$\begin{aligned}
&\mathrm{Var}\Big\{ -\frac{1}{n_+ h^2_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right) \boldsymbol{X}^*_{+,i} \\
&+ \frac{1}{n_+ h^3_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi^{(2)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right) \boldsymbol{X}^*_{+,i} R(X_{+,i}) \Big\} \\
=&E\left(-\frac{1}{n_+ h^2_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right) \boldsymbol{X}^*_{+,i} \right) \\
&\left(-\frac{1}{n_+ h^2_{1,+} h_{2,+}} \sum_{i=1}^{n_+} \phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right) \boldsymbol{X}^*_{+,i} \right)^T (1 + o_p(1)) \\
=&\frac{1}{n_+ h^4_{1,+} h^2_{2,+}} \iint \phi^{(1)2}\left(\frac{\epsilon_+}{h_{1,+}}\right) \boldsymbol{X}^*_+ \boldsymbol{X}^{*T}_+ g_{\epsilon_+}(\epsilon_+|X_+) K^2\left(\frac{X_+ - x}{h_{2,+}}\right) f_{X_+}(X_+) d\epsilon_+ dX_+ \\
&(1 + o_p(1)) \\
=&\frac{\int \tau^2 \phi^2(\tau) d\tau}{n_+ h^3_{1,+} h_{2,+}} f_{X_+}(x) \begin{bmatrix} v_{0,+} g_{\epsilon_+}(0|x) & v_{1,+} g_{\epsilon_+}(0|x) \\ v_{1,+} g_{\epsilon_+}(0|x) & v_{2,+} g_{\epsilon_+}(0|x) \end{bmatrix} (1 + o_p(1)),
\end{aligned} \tag{C.26}$$

where $v_{+,j} = \int_{-\bar{x}}^{M} w^j K^2(w) dw$ for $j = 0, 1, 2$. Then, we can get

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) = &\frac{\int \tau^2 \phi^2(\tau) d\tau}{n_+ h^3_{1,+} h_{2,+} f_{X_+}(x)} \begin{bmatrix} \mu_{+,0} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) \\ \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,2} g^{(2)}_{\epsilon_+}(0|x) \end{bmatrix}^{-1} \\
&\begin{bmatrix} v_{0,+} g_{\epsilon_+}(0|x) & v_{1,+} g_{\epsilon_+}(0|x) \\ v_{1,+} g_{\epsilon_+}(0|x) & v_{2,+} g_{\epsilon_+}(0|x) \end{bmatrix} \begin{bmatrix} \mu_{+,0} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) \\ \mu_{+,1} g^{(2)}_{\epsilon_+}(0|x) & \mu_{+,2} g^{(2)}_{\epsilon_+}(0|x) \end{bmatrix}^{-1} (1 + o_p(1)).
\end{aligned} \tag{C.27}$$

Define $W_{n_+} = \frac{1}{n_+ h_{1,+}^2 h_{2,+}} \sum_{i=1}^{n_+} \phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) K\left(\frac{X_{+,i}-x}{h_{2,+}}\right) \boldsymbol{X}_{+,i}^*$. By Slutsky's theorem, to show Theorem 4.2.19, it is sufficient to show that

$$T_{n_+} = \sqrt{n_+ h_{2,+} h_{1,+}^3} W_{n_+} \xrightarrow{d} \mathcal{N}(0, T), \tag{C.28}$$

where $T = \int \tau^2 \phi^2(\tau) d\tau f_{X_+}(x) \begin{bmatrix} v_{0,+} g_{\epsilon_+}(0|x) & v_{1,+} g_{\epsilon_+}(0|x) \\ v_{1,+} g_{\epsilon_+}(0|x) & v_{2,+} g_{\epsilon_+}(0|x) \end{bmatrix}$. We then prove that for any unit vector $\boldsymbol{d} \in R^2$,

$$\left\{\boldsymbol{d}^T \operatorname{Cov}\left(T_{n_+}\right) \boldsymbol{d}\right\}^{-1/2} \left\{\boldsymbol{d}^T T_{n_+} - \boldsymbol{d}^T E\left(T_{n_+}\right)\right\} \xrightarrow{d} N(0,1). \tag{C.29}$$

We therefore check Lyapunov's condition. Let $\xi_i = \sqrt{h_{2,+} h_{1,+}^3 / n_+} K\left(\frac{X_+ - x}{h_{2,+}}\right) \frac{1}{h_{1,+} h_{2,+}} \phi^{(1)}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) \boldsymbol{d}^T \boldsymbol{X}_{+,i}^*$, we need to prove $n_+ E|\xi_1|^3 \to 0$. As $\left(\boldsymbol{d}^T \boldsymbol{X}_{+,i}^*\right)^2 \le \|\boldsymbol{d}\|^2 \left\|\boldsymbol{X}_{+,i}^*\right\|^2$, $\phi^{(1)}(\cdot)$ is bounded, and $K(\cdot)$ has a compact support, we have

$$n_+ E|\xi|^3 \le O\left(n_+ n_+^{-3/2} h_{2,+}^{-3/2} h_{1,+}^{3/2}\right) E\left|K^3\left(\frac{X_+ - x}{h_{2,+}}\right) \phi^{(1)3}\left(\frac{\epsilon_{+,i}}{h_{1,+}}\right) \boldsymbol{d}^T \boldsymbol{X}_{+,i}^*\right| \to 0. \tag{C.30}$$

Thus, the asymptotic normality for T_{n_+} holds.