Semiparametric Partially Linear Varying Coefficient Modal Regression*

Aman Ullah† Tao Wang‡ Weixin Yao§

First Version: September 23, 2020

This Version: June 15, 2022

Abstract

We in this paper propose a *semiparametric partially linear varying coefficient (SPLVC) modal regression*, in which the conditional mode function of the response variable given covariates admits a partially linear varying coefficient structure. In comparison to existing regressions, the newly developed SPLVC modal regression captures the "most likely" effect and provides superior prediction performance when the data distribution is skewed. The consistency and asymptotic properties of the resultant estimators for both parametric and nonparametric parts are rigorously established. We employ a kernel-based objective function to simplify the computation and a modified modal-expectation-maximization (MEM) algorithm to estimate the model numerically. Furthermore, taking the residual sums of modes as the loss function, we construct a goodness-of-fit testing statistic for hypotheses on the coefficient functions, whose limiting null distribution is shown to follow an asymptotically χ^2-distribution with a scale dependent on density functions. To achieve sparsity in the high-dimensional SPLVC modal regression, we develop a regularized estimation procedure by imposing a penalty on the coefficients in the parametric part to eliminate the irrelevant variables. Monte Carlo simulations and two real-data applications are conducted to examine the performance of the suggested estimation methods and hypothesis test. We also briefly explore the extension of the SPLVC modal regression to the case where some varying coefficient functions admit higher-order smoothness.

Keywords: Goodness-of-fit test, MEM algorithm, Modal regression, Oracle property, Partially linear varying coefficient.

JEL Classification: C01, C12, C14, C50.

*We are grateful to the Co-Editor Xiaohong Chen, an anonymous Associate Editor, and two anonymous referees for their constructive comments, which have greatly improved the previous version of the paper. We also thank participants at the 55th Annual Conference of the CEA and the 96th Annual Conference of the WEAI for valuable comments and suggestions.

†Department of Economics at University of California, Riverside, CA 92521. E-mail: aman.ullah@ucr.edu.

‡Department of Economics at University of Victoria, Victoria BC, V8P 5C2. E-mail: taow@uvic.ca.

§Department of Statistics at University of California, Riverside, CA 92521. E-mail: weixin.yao@ucr.edu.

1 Introduction

Semiparametric models have become the latest state-of-the-art in recent years due to the flexible specification that allows traditional linear models to be combined with nonparametric models. They can reduce the possibility of model misspecification and ameliorate some of the drawbacks of a fully nonparametric model, such as the "curse of dimensionality" and lack of extrapolation capability. One popular semiparametric specification is a *semiparametric partially linear varying coefficient (SPLVC)* regression, which models the key covariates linearly and the rest of the variables nonparametrically. In particular, the response variable $Y \in \mathbb{R}$ depends on the associated covariates $\boldsymbol{X}_i \in \mathbb{R}^p$, $\boldsymbol{Z}_i \in \mathbb{R}^k$, and $\boldsymbol{U}$ in the structure of (Ahmad et al., 2005; Fan and Huang, 2005; Zhou and Liang, 2009)

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(\boldsymbol{U}_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \epsilon_i, \ i = 1, \cdots, n, \tag{1.1}$$

where $\boldsymbol{U}_i = (U_{i1}, \cdots, U_{iq})$ is a $1 \times q$ vector of index variables, $\boldsymbol{\alpha}(\boldsymbol{U}_i) = (\alpha_1(\boldsymbol{U}_i), \cdots, \alpha_p(\boldsymbol{U}_i))^T$ is a $p \times 1$ vector of unknown smooth nonparametric functions of $\boldsymbol{U}_i$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_k)^T$ is a k-dimensional vector of unknown parameters, the superscript T denotes the transpose of a vector or matrix, and ϵ_i is the unobservable random error that satisfies certain additional properties, such as conditional zero mode in this paper. To avoid the "curse of dimensionality", we let $U_i \in \mathbb{R}$ be a scalar throughout the rest of this paper, which ranges over a nondegenerate compact interval assumed to be the unit interval [0, 1] with little loss of generality. The extension to multivariate $\boldsymbol{U}_i$ involves no fundamentally new ideas but more complicated notations. According to Fan and Huang (2005), (1.1) allows for nonlinear interaction between the covariates U_i and $\boldsymbol{X}_i$ to such an extent that the impact of $\boldsymbol{X}_i$ varies at different levels of the covariate U_i with different linear models, thus increasing the flexibility of the model. (1.1) is a prevalent model in sociology, economics, finance, and statistics because it provides a parsimonious approach for inference in a variety of contents. Of particular interest is that (1.1) is flexible enough to form a general family of numerous multidimensional models. For instance, it includes the varying coefficient model when $\boldsymbol{Z}_i = 0$, the semiparametric partially linear model when $\boldsymbol{X}_i = 1$ and $p = 1$, and the single index model when $\boldsymbol{Z}_i = 0$, $\boldsymbol{X}_i = 1$, and $p = 1$. This suggests that the technical results developed in what follows can be straightforwardly extended to other non- and semiparametric modal regression models (Remark 4).

A large number of estimation methods, such as the local linear method, the profile least squares method, the average derivatives method, and the smoothing spline method, are already well established for estimating (1.1) built upon the idea of mean or quantile; see Ahmad et al. (2005) for more details. Aside from these, we can alternatively estimate (1.1) by Robinson (1988) model utilizing a two-step estimation approach,[1] in which we concentrate out the unknown

[1]The well-known Robinson (1988) type estimation approach may not be appropriate for the SPLVC modal

functional coefficient $\boldsymbol{\alpha}(U_i)$ by using a generalization of residual regression. All of the proposals discussed thus far, however, are concerned with the conventional mean or quantile regression. When the data contain a number of outliers or have a skewed distribution (non-Gaussian errors), the traditional nonparametric regressions applied to the SPLVC model may struggle to extract the intrinsic trends, resulting in degraded performance. For example, the mean regression may break down in practice if the error distribution lacks a finite second moment (e.g., Cauchy distribution).[2] To gain new insights into the underlying structure of skewed data, we investigate (1.1) under the content of modal regression and introduce a so-called *SPLVC modal regression* to target the most probable value of a dependent variable given covariates. Besides presenting a comprehensive description of how the conditional mode of the response variable depends on covariates, SPLVC modal regression can provide a shorter prediction interval than mean or quantile regression because with the same interval length, the interval around the conditional mode covers more samples than the interval around the conditional mean or quantile (Figure 1). Meanwhile, SPLVC modal regression can capture the "most likely" effect of certain covariates that would otherwise be missed by mean or quantile regression. Therefore, it is of interest and desire to develop a statistical methodology to complement the current modal regression literature. To the extent of our knowledge, the present paper is the first work to systematically develop the theory and methodology for flexible SPLVC modal regression.

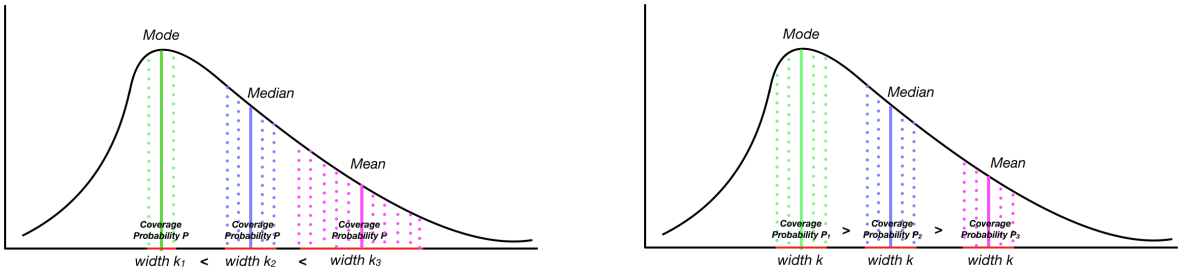

Figure 1: Comparison of Prediction Performance with Skewed Data

Note: Among the three location or centre measures (mean, median, and mode), with the same coverage probability, modal prediction provides the narrowest width (left plot); while it produces the highest coverage probability given the same prediction width (right plot).

There are two types of modal regression studied in the literature: unimodal regression and multimodal regression. Both of them can be obtained by optimizing a joint or conditional distribution function. Suppose that X is univariate with a compactly supported density. Modal

regression (as mode does not have the additive property in general), unless some restrictive conditions, such as unimodal symmetric distribution, are imposed. Nevertheless, the mean estimate is identical to the modal estimate in this case, and we would prefer to use mean estimation owing to the faster convergence rate.

[2]Although quantile regression can describe the entire situation of the conditional distribution of a dependent variable given covariates, it fails to reveal how the conditional mode depends on covariates directly to detect the "most likely" effect and may produce low density point predictions; see Figure 1.

regression can be defined as

$$m_g(x) = Mode(Y|X=x) = \arg\max_Y f(Y|X=x) = \arg\max_Y f(Y,x), \tag{1.2}$$

where $f(Y|X=x) = f(Y,x)/f(x)$, $f(Y|X=x)$ is the conditional density of Y given x, $f(Y,x)$ is the joint distribution of Y and x, and $f(x)$ is the marginal distribution of x. We can then utilize some indirect density-based estimation methods to capture modal regression lines (Chen et al., 2016). However, such an idea based on density estimation is not particularly feasible in the presence of high-dimensional covariates and typically has very poor convergence rates. To avoid estimating the density function, researchers explored the direct imposition of certain mode structures on $Mode(Y|X=x)$. The path-breaking papers of modal regression in econometrics are Lee (1989, 1993), in which they investigated modal regression by observing that the conditional mode from the truncated data provides a consistent estimate of the conditional mean for the original non-truncated data. To achieve consistency, Lee (1989, 1993) required constant tuning parameters and symmetric error density assumptions around zero on models, implying that their modal regression estimator is essentially a kind of robust regression estimator. Until only very recently, Kemp and Santos Silva (2012) and Yao and Li (2014) proposed a kernel-based objective function with the bandwidth approaching zero to achieve a consistent modal estimator even for skewed data, where they forced a linear regression structure on $Mode(Y|X)$. Such findings significantly simplify computations and widen the applicability of modal regression, making it a valuable addition to the regression tools for social, economic, financial, and statistical sciences. Since then, there has been an upsurge of interest and effort in developing modal regression; see for example, the work of Chen et al. (2016), Yao and Xiang (2016), Krief (2017), Chen (2018), Ota et al. (2019), Zhou and Huang (2019), Feng et al. (2020), Kemp et al. (2020), Zhang et al. (2021), and Ullah et al. (2021, 2022) and references therein.

Motivated by the aforementioned literature, we devote to investigating (1.1) in the context of modal regression, where we treat all bandwidths as going to zero and enable the density of error terms to be skewed and dependent on covariates. Differing from traditional mean regression under the error condition $\mathbb{E}\left(\epsilon_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i\right) = 0$, we have SPLVC modal regression

$$Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T\boldsymbol{\beta} \tag{1.3}$$

with the assumptions that ϵ_i is independent and identically distributed ($i.i.d.$) and admits a unique global zero mode such that $Mode\left(\epsilon_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i\right) = 0$. The primary goal of this paper is to develop theories and methods for estimating the unknown parameter $\boldsymbol{\beta}$ and the unknown functional coefficient $\boldsymbol{\alpha}(U_i)$, which can be naturally interpreted as the effects of covariates on the "most likely" data points of Y or the change in the mode of the response variable Y corresponding to a unit change in the covariate. Because $\boldsymbol{\alpha}(U_i)$ is modeled nonparametrically, it is reasonable to consider local linear estimation. Nonetheless, since the arguments of $\boldsymbol{\beta}$ (i.e., global

estimator) and $\boldsymbol{\alpha}(U_i)$ (i.e., local estimator) are different, they should be estimated with *modal* parametric and nonparametric rates of convergence, respectively. We thus develop a three-stage estimation procedure to estimate (1.3) by approximating $\boldsymbol{\alpha}(U_i)$ with a local linear function and updating estimates in different stages through a kernel-based objective function.

We obtain the convergence rates and establish the asymptotic distributions of the finite dimensional parameters and varying coefficients in different stages under regularity conditions. We show that the second- and third-stage estimators are oracles in the sense that the asymptotic properties are unaffected by the unknown components. The convergence rate of SPLVC modal regression is slower than that of mean regression, which is the cost we must pay to estimate the conditional mode (Parzen, 1962). Nevertheless, based on the simulation results in this paper, the SPLVC modal regression generally still provides estimates with smaller mean squared errors ($MSEs$) and narrower prediction intervals than the SPLVC mean regression for finite sample performance with skewed data. Since there are no closed-form solutions for the SPLVC modal regression, we introduce a modified MEM algorithm to efficiently achieve numerical estimates with the use of a normal kernel function. Note that the proposed estimation procedure implicitly assumes that all varying coefficient functions possess the same minimum degree of smoothness and hence can be approximated equally effectively. If some functional coefficients are known to have higher-order of differentiability, the bias rate of all estimated varying coefficient functions will be determined by the rate of the local polynomial with the lowest degree. In this case, the suggested estimation procedure based on local linear approximation may be ineffective (in the sense of optimal convergence rate). We in the supplementary note S1 generalize the proposed model to the case where some varying coefficient functions admit higher-order smoothness, and present a two-step estimation method that can attain the optimal convergence rate.

Furthermore, the most essential assumption in the developed estimation procedure is that the subset of variables with a constant or varying effect on the response is known in advance. However, it is impractical to accomplish this artificially in the application. Due to the difference in estimation rate, treating the parametric component of the SPLVC modal regression as a nonparametric function would incur a loss in efficiency. Therefore, it is of particular interest and importance to determine whether the varying coefficient functions truly vary with a certain variable or follow the linear form. Because of differences in function estimation, the classical profile likelihood ratio test for the mean estimate cannot be utilized directly for testing varying coefficient functions in modal regression. To develop an easily understandable and generally applicable method for the SPLVC regression regarding conditional mode processes, we extend the generalized likelihood technique of Fan et al. (2001) to propose a goodness-of-fit testing statistic for hypotheses on the coefficient functions by taking a kernel-based function as the loss function. The asymptotic behavior of the suggested test demonstrates that its limiting null distribution follows a χ^2-distribution with a scale depending on unknown density functions. Because the asymptotic distribution heavily relies on many unknown terms and is associated

with diverging degrees of freedom, obtaining an accurate distribution for the testing statistic under consideration is difficult. To avoid density estimation, we construct a residual-based modal bootstrap procedure to consistently approximate the unknown distribution of the test statistic and compute the analogous p-value. The simulation results show that the resulting testing procedure performs fairly effectively.

Many variables in practical applications might be irrelevant or insignificant, and their inclusion would cause a substantial loss in estimation accuracy. As a result, variable selection should be carried out prior to modeling. In recent decades, many researchers have developed variable selection procedures based on the concept of penalty functions to model the mean or quantile of a response variable Y as a function of a selected vector $\boldsymbol{X}$; see Su and Zhang (2013) for a detailed review. Although selecting relevant variables in the SPLVC model is not a new problem, there has been no formal work elaborating variable selection for the SPLVC modal regression to our knowledge. Generally, variable selection for semiparametric regression models consists of two components: identifying significant variables in the nonparametric component and selecting significant variables in the parametric component. Because the proposed test can be used to identify significant variables in the nonparametric component of the SPLVC modal regression,[3] we concentrate on the variable section utilizing the penalty function for the parametric component in this paper. Particularly, the form "Kernel-based objective function+Penalty" is adopted, which we call *penalized SPLVC modal regression*. With the proper regularization parameters and the assumption that the dimension k of the parameter $\boldsymbol{\beta}$ is fixed, the penalized modal estimator is shown to possess an oracle property. This implies that the estimator can correctly select the nonzero coefficients with a probability converging to one and has the same asymptotic distribution as if the subset of true zero coefficients and the varying coefficient functions were known. With the aid of a local quadratic approximation, the proposed variable selection method is computationally convenient by a modified MEM algorithm.

The rest of this paper is organized as follows. In Section 2, we focus on the construction of a three-stage estimation procedure relying on the local linear approximation to estimate the newly developed SPLVC modal regression, and present the large sample properties of the resulting estimators. We also explore bandwidth selection and suggest a varying coefficient test. Section 3 investigates SPLVC modal regression variable selection with penalty function, where the oracle property of the variable selection procedure is investigated. The results of applications to simulated and real datasets are reported in Section 4. Conclusions are given in Section 5. In the supplementary note, we address the extension of the proposed SPLVC modal regression to the case where some coefficient functions admit higher-order smoothness, and provide all technical proofs and additional simulation results.

[3]Model structure identification is critical since the validity of the inference is largely reliant on how well the model structure is specified. Although the developed test can be used to identify nonparametric components, it is challenging to implement in practice because of the dramatic increase in computational burden (i,e, for each submodel we need to estimate the varying coefficient functions).

2 SPLVC Modal Regression

We in this section propose a three-stage estimation procedure to achieve the optimal convergence rates for both global parameters and varying coefficients. Specifically, in the **first stage**, the local linear approximation is employed to get the initial estimators; in the **second stage**, we obtain the optimal convergence rate for the parametric estimator using all data points after plugging in the estimates of varying coefficients; and in the **third stage**, we re-estimate the varying coefficient functions through a local linear method after plugging in the estimates of the finite dimensional parameters. Following that, we discuss practical bandwidth selection and introduce a goodness-of-fit testing statistic to test whether the varying coefficients are constant.

2.1 Local Linear Modal Estimators

Suppose that $\{(Y_i, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n$ are *i.i.d.* samples and $\boldsymbol{\alpha}(U_i)$ is smooth enough that its first and second derivatives exist. We then estimate (1.3) by locally approximating the unknown nonparametric functional coefficient $\boldsymbol{\alpha}(U_i)$ with a linear function for a given $u \in \mathbb{R}$ in the neighborhood of U_i, i.e., $|U_i - u| = o(1)$,

$$\alpha_j(U_i) \approx \alpha_j(u) + \alpha_j^{(1)}(u)(U_i - u) \equiv \alpha_j(u) + b_j(u)(U_i - u), \quad j = 1, \cdots, p,$$

where "$a_n \approx b_n$" indicates that for sufficiently large n, $a_n/b_n \to 1$, i.e., $a_n = b_n + o_p(b_n)$, $o_p(b_n)$ represents the term with a probability order smaller than that of b_n, "$\equiv$" means "is defined as", and $\alpha_j^{(1)}(u) = b_j(u)$ is the first derivative of $\alpha_j(\cdot)$. We consider local linear approximation for ease of presentation, which has advantages in the ability of design adaptation and high asymptotic efficiency (Fan and Gijbels, 1996), but it is straightforwardly generalized to local polynomial estimation with the assumption of higher order derivatives.

Denoting $\boldsymbol{\alpha}(u) = (\alpha_1(u), \cdots, \alpha_p(u))^T \in \mathbb{R}^p$ and $\boldsymbol{b}(u) = (b_1(u), \cdots, b_p(u))^T \in \mathbb{R}^p$, we obtain the following local kernel-based objective function to recover the unknown functional coefficient $\boldsymbol{\alpha}(U_i)$ and parameter $\boldsymbol{\beta}$ at each data point

$$Q_n(\boldsymbol{\alpha}(u), \boldsymbol{b}(u), \boldsymbol{\beta}(u)) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} \phi\left(\frac{Y_i - \boldsymbol{X}_i^T\left(\boldsymbol{\alpha}(u) + \boldsymbol{b}(u)\left(U_i - u\right)\right) - \boldsymbol{Z}_i^T\boldsymbol{\beta}(u)}{h_1}\right) K\left(\frac{U_i - u}{h_2}\right), \tag{2.1}$$

where $\phi(\cdot)$ is a nonnegatively symmetric kernel function with bounded support and bandwidth $h_1 := h_1(n) \to 0$ as $n \to \infty$, and $K(\cdot)$ is a bounded and symmetric kernel function associated with the size of the local neighborhood bandwidth $h_2 := h_2(n) \to 0$ as n approaches infinity. To prevent notation confusion, we suppress the n for all bandwidths used in this paper. Note that the kernel function $K(\cdot)$ reflects the fact that (2.1) is only applied to data around u and gives a larger weight to data closer to the point u, which is consistent with the weight function

in classical nonparametric estimation. According to Yao and Li (2014) and Ullah et al. (2021, 2022), the choice of kernel function $\phi(\cdot)$ is not particularly crucial compared to the choice of bandwidths in modal regression. For computational simplicity, we choose a standard normal kernel for $\phi(\cdot)$ in this paper to develop a modified MEM algorithm. We henceforth use $\hat{\boldsymbol{\alpha}}(u)$, $\hat{\boldsymbol{b}}(u)$, and $\hat{\boldsymbol{\beta}}(u)$ to denote the naive modal estimators from (2.1).

Remark 1. (**Iterative Method**) *The objective of (2.1) is to provide initial estimators for the following two estimation stages and to make variable selection for the parametric component easier in Section 3. Nevertheless, we can employ an iterative procedure to obtain the final efficient estimators for the proposed SPLVC modal regression without conducting the first-stage estimation. Specifically, (2.3) is maximized for any given $\boldsymbol{\beta}$. Then, given the estimators of $\boldsymbol{\alpha}(u)$ and $\boldsymbol{b}(u)$, we can update $\boldsymbol{\beta}$ by solving the objective function (2.2). Iterate these two steps until a stopping rule is satisfied for the convergence of the estimator of $\boldsymbol{\beta}$. The advantage of this iterative method is that no undersmoothing is required to obtain the optimal modal convergence rate for the estimator of $\boldsymbol{\beta}$, and thus the common selection criteria for the optimal bandwidth in Remarks 6 can be directly employed. Such an iterative method, however, is computationally more expensive than the proposed three-stage estimation procedure.*

As $\boldsymbol{\beta}$ is a global parameter and only data in the local neighborhood of u are utilized, the naive estimator $\hat{\boldsymbol{\beta}}(u)$ converges in probability to its true value at a nonparametric rate, as demonstrated in Theorem 2.2. To achieve the optimal convergence rate for the modal estimator of $\boldsymbol{\beta}$, we substitute the varying coefficient $\boldsymbol{\alpha}(U_i)$ with its estimate in the first stage, transforming the SPLVC modal regression into a pseudo linear modal regression. We then apply the following global kernel-based objective function to re-estimate $\boldsymbol{\beta}$ using all data points

$$Q_n(\boldsymbol{\beta}) = \frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right), \tag{2.2}$$

where $h_3 := h_3(n) \to 0$ as $n \to \infty$ is an optimal bandwidth for estimating $\boldsymbol{\beta}$. Comparing the proposed global kernel-based objective function (2.2) to its local counterpart (2.1), we can observe that the naive modal estimator only uses local data, whereas the global modal estimator takes advantage of the full sample information. As expected, the global modal estimator has the optimal modal parametric convergence rate. Here and hereafter, we refer to $\tilde{\boldsymbol{\beta}}$ from (2.2) as the semi-modal estimator of $\boldsymbol{\beta}$.

Remark 2. (**Weighted Mean**) *Alternatively, we can calculate the globally weighted mean of all naive estimates $\hat{\boldsymbol{\beta}}(u)$ to estimate $\boldsymbol{\beta}$, i.e., $\tilde{\boldsymbol{\beta}} = Mean(\hat{\boldsymbol{\beta}}(U_i)) = \int \hat{\boldsymbol{\beta}}(U_i)dW(u)$, where $W(\cdot)$ is a deterministic weighting function with $\int dW(u) = 1$ that can be discrete or continuous, and the integral should be interpreted in the Stieltjes sense. This method can ensure that all data points are used and thus lead to a faster rate of convergence than the naive estimator. We conjecture that the global mean estimator should be $\sqrt{nh_1^3}$-consistent.*

After obtaining $\tilde{\boldsymbol{\beta}}$ from (2.2), it is natural to further re-estimate the nonparametric part by plugging in the semi-modal estimate, which converts the original model to a pseudo varying coefficient modal regression. The final estimator of $\boldsymbol{\alpha}(U_i)$ is then constructed by maximizing the following local kernel-based objective function

$$Q_n(\boldsymbol{\alpha}(u), \boldsymbol{b}(u)) = \frac{1}{nh_4h_5} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \left(\boldsymbol{\alpha}(u) + \boldsymbol{b}(u) \left(U_i - u \right) \right) - \boldsymbol{Z}_i^T \tilde{\boldsymbol{\beta}}}{h_4} \right) K \left(\frac{U_i - u}{h_5} \right), \quad (2.3)$$

where $h_4 := h_4(n) \to 0$ as $n \to \infty$ is the bandwidth that is optimal for the estimation of $\boldsymbol{\alpha}(U_i)$, and $h_5 := h_5(n) \to 0$ as $n \to \infty$ is the bandwidth that controls the size of a local neighborhood of U_i. We denote the estimators from (2.3) as $\tilde{\boldsymbol{\alpha}}(u)$ and $\tilde{\boldsymbol{b}}(u)$.[4]

Remark 3. *When the skewed error density is reduced to a symmetric one, the SPLVC modal regression accordingly degenerates in line with the SPLVC mean regression. The shrinking bandwidths associated with the error terms and the resulting slower convergence rates, however, make the suggested modal regression suboptimal for directly producing mean estimates in this instance. Researchers can resort to the model established in Zhang et al. (2013) to explore the robust estimators of the SPLVC mean regression on the basis of mode value to attain the mean convergence rates. Such a modal-based SPLVC regression can be regarded as a robust alternative against outliers in the variables or non-normal symmetric errors, with the resultant estimators achieving efficiency by utilizing constant bandwidths related to the error terms.*

Remark 4. (**Extended Models**) *The results in this paper encompass several alternative specifications of interest to econometricians, as stated in Section 1. The varying coefficient and the semiparametric partially linear modal regressions can be included straightforwardly. When* $\boldsymbol{Z}_i = 0$, $\boldsymbol{X}_i = 1$, *and* $p = 1$, *we can obtain the single-index modal regression. A more complicated estimation procedure is required for this new single index model. Assume that* $Y_i = g(\boldsymbol{\eta}_i^T \boldsymbol{\theta}) + \epsilon_i$, *in which* $\boldsymbol{\eta}_i = (\eta_{i1}, \cdots, \eta_{ip})^T$ *is the covariate,* $\boldsymbol{\theta} \in \mathbb{R}^p$ *is the unknown index parameter, and* $g(\cdot)$ *is an unspecified link function. For* $\boldsymbol{\eta}_i^T \boldsymbol{\theta}$ *in the neighborhood of* t, *the function* $g(\cdot)$ *can be locally linearly approximated as*

$$g(\boldsymbol{\eta}_i^T \boldsymbol{\theta}) \approx g(t) + g^{(1)}(t)(\boldsymbol{\eta}_i^T \boldsymbol{\theta} - t) \equiv d_0 + d_1(\boldsymbol{\eta}_i^T \boldsymbol{\theta} - t),$$

where $d_0 = g(t)$, $d_1 = g^{(1)}(t)$, *and* $g^{(1)}(\cdot)$ *is the first derivative of* $g(\cdot)$. *After that, we can maximize a local kernel-based objective function*

$$Q_n(d_0, d_1) = \frac{1}{nh_{s1}h_{s2}} \sum_{i=1}^{n} \phi \left(\frac{Y_1 - d_0 - d_1(\boldsymbol{\eta}_i^T \boldsymbol{\theta} - t)}{h_{s1}} \right) K \left(\frac{\boldsymbol{\eta}_i^T \boldsymbol{\theta} - t}{h_{s2}} \right)$$

with respect to the parameters d_0 *and* d_1 *given an initial estimator of* $\boldsymbol{\theta}$, *where* h_{s1} *and* h_{s2} *are two*

[4]Numerically, we can further iterate the last two estimation stages to update the estimates (until it converges). Nevertheless, such an iteration would add more computational workload while not significantly improving the estimations based on the simulation results.

bandwidths that approach zero as the sample size n processes infinity. Then, with all of the data available, we maximize a global kernel-based objective function similar to (2.2) to re-estimate $\boldsymbol{\theta}$. *Continue with the preceding procedures until convergence is reached. For identification, we also need the Euclidean norm* $\|\boldsymbol{\theta}\|_2 = 1$.

Algorithm 1 MEM Algorithm for SPLVC Modal Regression

First-Stage Equation (2.1)

Input: data $\{(Y_i, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n$, kernel bandwidths h_1 and h_2, and the initial guess $\boldsymbol{\alpha}(u)^{(0)} \in \mathbb{R}^p$, $\boldsymbol{b}(u)^{(0)} \in \mathbb{R}^p$, $\boldsymbol{\beta}(u)^{(0)} \in \mathbb{R}^k$.

Output: the estimated coefficients $\boldsymbol{\alpha}(u)^{(g+1)} \in \mathbb{R}^p$, $\boldsymbol{b}(u)^{(g+1)} \in \mathbb{R}^p$, $\boldsymbol{\beta}(u)^{(g+1)} \in \mathbb{R}^k$.

while the stopping criterion (e.g., Euclidean distance) is not satisfied **do**

- **E-Step** Update weight (posterior conditional probability) $\pi\left(i|\boldsymbol{\alpha}(u)^{(g)}, \boldsymbol{b}(u)^{(g)}, \boldsymbol{\beta}(u)^{(g)}\right)$

$$\pi\left(i|\boldsymbol{\alpha}(u)^{(g)}, \boldsymbol{b}(u)^{(g)}, \boldsymbol{\beta}(u)^{(g)}\right) = \frac{\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\left(\boldsymbol{\alpha}(u)^{(g)} + \boldsymbol{b}(u)^{(g)}(U_i - u)\right) - \boldsymbol{Z}_i^T\boldsymbol{\beta}(u)^{(g)}}{h_1}\right) K\left(\frac{U_i - u}{h_2}\right)}{\sum_{i=1}^n \phi\left(\frac{Y_i - \boldsymbol{X}_i^T\left(\boldsymbol{\alpha}(u)^{(g)} + \boldsymbol{b}(u)^{(g)}(U_i - u)\right) - \boldsymbol{Z}_i^T\boldsymbol{\beta}(u)^{(g)}}{h_1}\right) K\left(\frac{U_i - u}{h_2}\right)}.$$

- **M-Step** Update the values of coefficients with the weight calculated in E-Step by

$$\left(\boldsymbol{\alpha}(u)^{(g+1)}, \boldsymbol{b}(u)^{(g+1)}, \boldsymbol{\beta}(u)^{(g+1)}\right) = \arg\max_{\boldsymbol{\alpha}(u), \boldsymbol{b}(u), \boldsymbol{\beta}} \sum_{i=1}^n \left\{\pi\left(i|\boldsymbol{\alpha}(u)^{(g)}, \boldsymbol{b}(u)^{(g)}, \boldsymbol{\beta}(u)^{(g)}\right) \log\left(\frac{1}{h_1}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\left(\boldsymbol{\alpha}(u) + \boldsymbol{b}(u)\left(U_i - u\right)\right) - \boldsymbol{Z}_i^T\boldsymbol{\beta}(u)}{h_1}\right)\right)\right\}.$$

- Set $g := g + 1$.

end while

Second-Stage Equation (2.2)

Input: data $\{(Y_i, \boldsymbol{X}_i, \hat{\boldsymbol{\alpha}}(U_i), \boldsymbol{Z}_i)\}_{i=1}^n$, kernel bandwidth h_3, and the initial guess $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^k$.

Output: the estimated coefficient $\boldsymbol{\beta}^{(g+1)} \in \mathbb{R}^k$.

while the stopping criterion (e.g., Euclidean distance) is not satisfied **do**

- **E-Step** Update weight (posterior conditional probability) $\pi\left(i|\boldsymbol{\beta}^{(g)}\right)$ as

$$\pi\left(i|\boldsymbol{\beta}^{(g)}\right) = \frac{\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^{(g)}}{h_3}\right)}{\sum_{i=1}^n \phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^{(g)}}{h_3}\right)}.$$

- **M-Step** Update the value of $\boldsymbol{\beta}^{(g+1)}$ with the weight calculated in E-Step by

$$\boldsymbol{\beta}^{(g+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^n \left\{\pi\left(i|\boldsymbol{\beta}^{(g)}\right) \log\left(\frac{1}{h_3}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right)\right)\right\}.$$

- Set $g := g + 1$.

end while

Note: The algorithm for the third-stage equation (2.3) is similar to that for the first-stage equation (2.1), but associated with bandwidths h_4 and h_5.

The closed-form solutions for (2.1)-(2.3) are unavailable, suggesting that a mode-hunting procedure should be applied. Because maximizing the objective function $Q(\cdot)$ is equivalent to maximizing $\log(Q(\cdot))$, we suggest a modified MEM Algorithm 1 based on Li et al. (2007) and Yao et al. (2012) to estimate modal coefficients by iterating the E-step and M-step until the algorithm converges or a stopping criterion is satisfied. The monotone ascending property of the proposed MEM algorithm, i.e., $Q_n(\cdot^{(g+1)}) \geq Q_n(\cdot^{(g)})$, can be developed along the lines of Yao and Li (2014) by utilizing Jensen's inequality, which guarantees the stability and convergence of the algorithm (i.e., local optimum). The weight scheme in E-Step enables modal regression to reduce the effect of observations far away from the modal regression curve in order to achieve robustness, which is one of the benefits of modal regression over mean regression.

Because of the normal kernel function we use, a closed-form expression for the maximizers of the objective functions exists in M-Step, namely, $(\boldsymbol{X}^{*T}\boldsymbol{W}\boldsymbol{X}^{*T})^{-1}\boldsymbol{X}^{*T}\boldsymbol{W}\boldsymbol{Y}$, where $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with diagonal elements $\{\pi(i|\cdot)\}_{i=1}^{n}$ obtained in E-step, and $\boldsymbol{X}^* = (\boldsymbol{X}\ \boldsymbol{Z})$ is the corresponding variable matrix with $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)^T$ and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n)^T$. Such an explicit expression largely simplifies the computation. Although maximization is searched across the entire space, the converged value of the MEM algorithm is highly dependent on the initial point and consequently produces sub-optimal estimates (Yao and Li, 2014; Ullah et al., 2021, 2022). In practice, if the computation is feasible, it is advisable to begin with a variety of estimates obtained by different estimation techniques, such as mean, quantile, or other robust estimates, to verify the stability of the solution (too many local solutions indicate that the solution is not stable). By choosing the estimate with the largest value of the target function, we can avoid the ambush of potential local maxima.[5]

2.2 Asymptotic Properties

Before proceeding to the asymptotic theorems, it is convenient to introduce some notations that will be used throughout the remaining part of this paper. We define $\mu_2 = \int w^2 K(w)\mathrm{d}w < \infty$ and $v_j = \int w^j K^2(w)\mathrm{d}w < \infty$ for $j = 0, 1, 2$, $\hat{\boldsymbol{X}} = (\boldsymbol{X}, U, \boldsymbol{Z})$, $H_r = \mathrm{diag}(h_r, \cdots, h_r)_{p\times p}$, $r = 2$ or 5, and use "$\xrightarrow{d}$" and "$\xrightarrow{p}$" to denote the convergence in distribution and probability, respectively. Given random variables W_n for $n \geq 1$, we write $W_n = O_P(w_n)$ if $\lim_{b\to\infty} \limsup_n P(|W_n| \geq bw_n) = 0$ and $W_n = o_P(w_n)$ if $\lim_n P(|W_n| \geq bw_n) = 0$ for any constant $b \geq 0$. We let a function $f(n) = O(1)$ if there exist some nonzero constants c and N such that $f(n)/c \to 1$ for $n \geq N$, and $f_n \asymp g_n$ means $0 < \liminf_{n\to\infty}|f_n/g_n| \leq \limsup_{n\to\infty}|f_n/g_n| < \infty$. Let $\|\cdot\|$ represent the Euclidean norm, i.e., $\|A\| = [tr(AA^T)]^{1/2}$, in which $tr(A)$ is the trace of the matrix A, and $\boldsymbol{\alpha}^{(c)}(u) \in \mathbb{R}^p$ indicates the cth derivative of $\boldsymbol{\alpha}(u)$ with respect to u. The following technical

[5]It is well-known that starting values impact the quality of the EM algorithm's solution, and a particular set of starting values will always converge to the same solution after repeated initializations. In such a case, we can stop after a specified number of iterations have been reached and keep iterating only from the estimates with the largest value of the objective function.

conditions are then listed to establish the consistency and asymptotic properties of the resultant modal estimators.

C1 The true value of parameter $\boldsymbol{\beta}_0$ is in the interior of the known compact parameter space, which is a subset of $\mathbb{R}^k$.

C2 $\{(Y_i, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n$ is an *i.i.d.* random sequence drawn from the joint probability distribution $F(Y, \boldsymbol{X}, U, \boldsymbol{Z})$ on $\mathbb{R} \times \mathbb{R}^p \times \mathbb{R} \times \mathbb{R}^k$. The error term admits a unique global zero mode such that $Mode\,(\epsilon_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) \overset{a.s.}{=} 0 \Rightarrow Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) \overset{a.s.}{=} \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T\boldsymbol{\beta}$.

C3 The index variable U has a bounded support Ω. Without loss of generality, Ω is the unit interval $[0,\, 1]$. The marginal density function $f_U(u)$ is continuous in some neighborhoods of u and has a value of $f_U(u) > 0$ on $\{u : 0 < F_U(u) < 1\}$, where $F_U(u)$ is the cumulative distribution function and u is an interior point on its support Ω. $f_U(u)$ also has a continuous first derivative and is bounded away from infinity.

C4 $\alpha_j(u) \in \mathcal{V}$ is rth continuously differentiable on Ω for $j = 1, \cdots, p$, where $\mathcal{V}$ denotes the class of varying coefficient functions and $r \geq 2$.

C5 The kernel function $K(\cdot) : \mathbb{R} \to \mathbb{R}$ is a nonnegatively symmetric and bounded kernel with compact support and integrates to one.

C6 Let $f_\epsilon(\epsilon|\hat{\boldsymbol{X}})$ denote the conditional probability density function of ϵ given $\hat{\boldsymbol{X}}$. For any $\hat{\boldsymbol{X}}$ in the corresponding support set, $f_\epsilon(\epsilon|\hat{\boldsymbol{X}})$ is bounded away from zero and infinity and has the fourth continuous derivative with respect to ϵ in a neighbour of zero. Furthermore, $f_\epsilon(\epsilon|\hat{\boldsymbol{X}}) < f_\epsilon(0|\hat{\boldsymbol{X}})$ for any $\epsilon \neq 0$ and $f_\epsilon^{(1)}(0|\hat{\boldsymbol{X}}) = 0$ for any $\hat{\boldsymbol{X}}$, where $f_\epsilon^{(d)}(\cdot)$ represents the dth derivative of $f_\epsilon(\cdot)$ with respect to ϵ.

C7 There is a constant $s > 2$ such that $\mathbb{E}(\|\boldsymbol{X}\|^{2s}) < \infty$ and $\mathbb{E}(\|\boldsymbol{Z}\|^{2s}) < \infty$ with probability one. The matrices $\Gamma(u)$, $\tilde{\Gamma}(u)$, and $\tilde{\Gamma}^*(u)$ defined in the following theorems are negative definite matrices in a neighborhood of u, and the eigenvalues of $\Gamma(u)$, $\tilde{\Gamma}(u)$, and $\tilde{\Gamma}^*(u)$ are bounded away from zero and infinity for all $u \in \Omega$. Also, J is a negative definite matrix.

While the above conditions appear to be a little verbose at first glance, they are actually quite modest and simple to satisfy in the literature of the SPLVC model and modal regression; see Fan and Huang (2005), Kai et al. (2011), Kemp and Santos Silva (2012), Yao and Li (2014), and Ullah et al. (2021, 2022). C1 is an ordinary regularity condition that is usually easy to verify. C2 is standard in describing the sample generating process for modal regression. The *i.i.d.* assumption can be relaxed to cover the strictly α-mixing and stationary case but at the expense of more tedious proofs. C3 is a reasonable condition related to the localized behavior around $u \in \Omega$, which is required for establishing the consistency and asymptotic normality of the resulting estimators. If U does not have a compact support, a transformation from $(-\infty, +\infty)$

to [0, 1] may be employed. C4 is a commonly used condition on the smoothness of nonparametric functions in local linear fitting. It controls the precision in the approximation of the varying coefficient functions as the second derivative of $\boldsymbol{\alpha}(u)$ impacts the bias. The definition of $\mathcal{V}$ is stated in Ahmad et al. (2005). We emphasize that this paper implicitly assumes that all varying coefficient functions admit the same minimum degree (twice) of smoothness to perform local linear approximation. The situation where $\alpha_j^{(r)}(\cdot)$ is continuous in a neighborhood of u for $j = 1, \cdots, p-1$ while the functional coefficient $\alpha_p(\cdot)$ has a continuous $(r+2)$th derivative in a neighborhood of u is investigated in the supplementary note S1. C5 is a mild condition on the kernel function that is satisfied by many commonly used kernels to derive the asymptotic variance of estimators. The compact support condition for the kernel function is not essential and is imposed merely to simplify the proof. It can be eased as long as we put certain integrability restrictions on the kernel function's tail. Especially, the standard normal density function is permitted (Ullah et al., 2021, 2022), which is the default kernel used in numerical analysis in this paper. C6 enforces a certain smoothness on $f_\epsilon(\epsilon|\cdot)$ in the neighborhood of zero, which is necessary for identification. It implies that the conditional density of ϵ has a well-defined unique global mode at zero. Note that such a unique global mode assumption is used for simple illustration. The proposed SPLVC modal regression can be applied to the multimode setting when the population is not homogeneous, where the suggested estimation method will find multiple modal solutions if starting from multiple initial values, with each solution corresponding to one local modal estimator; see supplementary note S2. In contrast to mean regression, we do not need to impose any moment conditions on error terms and can allow $Var(\epsilon) = \infty$. C7 is the standard rank condition placing restrictions on the moments of covariates to ensure the existence of the asymptotic mean and variance for modal estimators. The nonsingular restriction on matrices guarantees that J is invertible, as are $\Gamma(u)$, $\tilde{\Gamma}(u)$, and $\tilde{\Gamma}^*(u)$ for all $u \in \Omega$. All of the bandwidth requirements that need to be satisfied are listed in the following theorems to guarantee the consistency of the modal estimators and the biases from previous stages are negligible at the later stages. As is typical in the semiparametric literature, undersmoothing is required for the developed SPLVC modal regression to asymptotically ignore biases.

We provide the asymptotic theorems for the modal estimators in the corresponding stages. Special care is needed to develop asymptotic theories for semiparametric modal estimators. For instance, the approximation error caused by local linear estimation must be taken into account. We first present the following theorem providing the convergence rates of the naive estimators $\hat{\boldsymbol{\alpha}}(u)$, $\hat{\boldsymbol{b}}(u)$, and $\hat{\boldsymbol{\beta}}(u)$, where $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{b}_0(\cdot)$ are the true parameter vectors.

Theorem 2.1. *Suppose that the regularity conditions C1-C7 are fulfilled. With probability appro -aching one, as $n \to \infty$, $h_1 \to 0$, $h_2 \to 0$, $h_2^2/h_1 \to 0$, and $nh_2h_1^5 \to \infty$, there exist consistent maximizers $(\hat{\boldsymbol{\alpha}}(u), \hat{\boldsymbol{b}}(u), \hat{\boldsymbol{\beta}}(u))$ of (2.1) such that*

i. $\|\hat{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\| = O_p\left((nh_2h_1^3)^{-1/2} + h_1^2 + h_2^2\right)$,

ii. $\|H_2\hat{\boldsymbol{b}}(u) - H_2\boldsymbol{b}_0(u)\| = O_p\left((nh_2h_1^3)^{-1/2} + h_1^2 + h_2^2\right)$,

iii. $\|\hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0\| = O_p\left((nh_2h_1^3)^{-1/2} + h_1^2 + h_2^2\right)$.

Theorem 2.1 implies that the magnitude of the bias term is bounded in probability by the best approximation rates obtained by local linear estimation, i.e, the bias of the naive estimator is of order $O_p(h_1^2 + h_2^2)$, while the variance is of order $O_p\left(nh_2h_1^3\right)^{-1}$. The first component of the bias results from the modal estimating process, and the second term is attributed to the local linear approximation of $\alpha_l(U_i)$. Such results demonstrate that treating bandwidth h_1 as a constant to estimate modal parameters will misbehave and result in the inconsistent estimation of the parameters if the density of error is skewed. The use of local data points substantially degrades the estimation efficiency of $\hat{\boldsymbol{\beta}}$; essentially the rate of convergence is $O(n^{-1/4})$ with the optimal bandwidth choice $h_1 \asymp h_2 \asymp n^{-1/8}$ achieved by minimizing the asymptotic MSE of the naive estimator. Although this rate is slower than that of the SPLVC mean or quantile regression estimator, it is faster than that of the modal estimator derived from nonparametric kernel density estimation; see Chen et al. (2016). The following theorem provides the asymptotic normality results for naive estimators when u is in the interior of Ω.

Theorem 2.2. *With $nh_2^5h_1^3 = O(1)$ and $nh_2h_1^7 = O(1)$, under the same conditions as Theorem 2.1, the estimators satisfying the consistency results in Theorem 2.1 have the following asymptotic result*

$$\sqrt{nh_2h_1^3}\left[\begin{pmatrix}\hat{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\\ h_2(\hat{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u))\\ \hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0\end{pmatrix} - \Gamma(u)^{-1}\left(\frac{h_2^2}{2}\Lambda_2(u)\boldsymbol{\alpha}_0^{(2)}(u) - \frac{h_1^2}{2}\Lambda_1(u)\right)\right] \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\int t^2\phi^2(t)dt}{f_U(u)}\Gamma(u)^{-1}\Sigma(u)\Gamma(u)^{-1}\right).$$

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

$$\sqrt{nh_2h_1^3}\begin{pmatrix}\hat{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\\ h_2(\hat{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u))\\ \hat{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\int t^2\phi^2(t)dt}{f_U(u)}\Gamma(u)^{-1}\Sigma(u)\Gamma(u)^{-1}\right),$$

where $\Gamma(u) = \mathbb{E}\left[\begin{pmatrix}\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & \mathbf{0} & \boldsymbol{X}\boldsymbol{Z}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ \mathbf{0} & \mu_2\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & \mathbf{0}\\ \boldsymbol{Z}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & \mathbf{0} & \boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\end{pmatrix}\middle| U = u\right]$,

$$\Lambda_1(u) = \mathbb{E}\left[\begin{pmatrix}\boldsymbol{X} f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})\\ \mathbf{0}\\ \boldsymbol{Z} f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})\end{pmatrix}\middle| U = u\right],\ \Lambda_2(u) = \mathbb{E}\left[\begin{pmatrix}\mu_2\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ \mathbf{0}\\ \mu_2\boldsymbol{Z}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\end{pmatrix}\middle| U = u\right],$$

$$\text{and } \Sigma(u) = \mathbb{E}\left[\begin{pmatrix} v_0 \boldsymbol{X}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}}) & \mathbf{0} & v_0 \boldsymbol{X}\boldsymbol{Z}^T f_\epsilon(0|\hat{\boldsymbol{X}}) \\ \mathbf{0} & v_2 \boldsymbol{X}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}}) & \mathbf{0} \\ v_0 \boldsymbol{Z}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}}) & \mathbf{0} & v_0 \boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(0|\hat{\boldsymbol{X}}) \end{pmatrix} \middle| U = u\right].$$

Because we only need data in a local neighborhood of u to obtain the naive estimator, Theorem 2.2 indicates that the estimator is $\sqrt{nh_2h_1^3}$-consistent. It also shows that the asymptotic bias term can be successfully removed under certain conditions, and that the naive estimators are asymptotically normally distributed, centered at the true values of the parameters of interest. However, the MSE-optimal bandwidths of h_1 and h_2 in Theorem 2.2 have the rate $n^{-1/8}$, which does not satisfy the condition that $\lim_{n\to\infty} nh_2^5h_1^3 = 0$ and $\lim_{n\to\infty} nh_2h_1^7 = 0$. As a result, undersmoothing is necessary to eliminate the asymptotic bias at the expense of a relatively slower convergence rate, which is a common requirement in semiparametric models. This will be incorporated later when we choose the bandwidths for practical purposes. Owing to the use of a symmetric kernel function for $K(\cdot)$, $\hat{\boldsymbol{b}}(u)$ is asymptomatically independent of $\hat{\boldsymbol{\alpha}}(u)$ and $\hat{\boldsymbol{\beta}}(u)$. Nevertheless, $\hat{\boldsymbol{\alpha}}(u)$ and $\hat{\boldsymbol{\beta}}(u)$ are dependent on each other regardless of the kernel function used, necessitating the third-stage estimation procedure to re-estimate $\boldsymbol{\alpha}(u)$ to improve efficiency. Despite concentrating on the interior point u, the above asymptotic result holds true when we investigate the boundary behavior.

Remark 5. *The bias term has a magnitude of $O_p(h_1^2 + h_2^2)$, which is only of theoretical importance. Generally, omitting it does not substantially impact the accuracy of the estimates. However, we still apply undersmoothing technology to make the bias ignorable in practice. Furthermore, the practical inferential use of asymptotic distribution on estimators is made difficult by the complex form of the asymptotic covariance matrix due to the existence of several unknown quantities, such as $f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})$ and $f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})$. We can instead utilize the bootstrap method for related inference (and bias adjustment for modal estimates). More specifically, we follow the procedures* **S1**-**S3** *in Algorithm 2 to obtain B bootstrap pointwise estimators $\hat{\boldsymbol{\alpha}}_l^*(u)$, $l = 1, \cdots, B$, such that the bias of $\hat{\boldsymbol{\alpha}}(u)$ is $\hat{\mathbf{b}}^{boot}(u) = \frac{1}{B}\sum_{l=1}^B \hat{\boldsymbol{\alpha}}_l^*(u) - \hat{\boldsymbol{\alpha}}(u)$ and the covariance of $\hat{\boldsymbol{\alpha}}(u)$ is $\hat{\mathbf{V}}^{boot}(u) = \frac{1}{B-1}\sum_{l=1}^B (\hat{\boldsymbol{\alpha}}_l^*(u) - \bar{\hat{\boldsymbol{\alpha}}}^*(u))(\hat{\boldsymbol{\alpha}}_l^*(u) - \bar{\hat{\boldsymbol{\alpha}}}^*(u))^T$, in which $\bar{\hat{\boldsymbol{\alpha}}}^*(u) = (1/B)\sum_{l=1}^B \hat{\boldsymbol{\alpha}}_l^*(u)$. Subsequently, we can compute the confidence intervals. The following Theorems 2.4 and 2.6 are both subject to the same comments.*

The convergence rates of the naive modal estimators are provided by the two theorems above, which is crucial in deriving the asymptotic distribution for the final modal estimators. Although (2.2) is a straightforward linear modal regression objective function, the extra bias factor from the previous stage needs to be taken care of. Following that, we characterize the consistency and asymptotic behavior of the semi-modal estimator $\tilde{\boldsymbol{\beta}}$ by plugging in the estimates of varying coefficients.

Theorem 2.3. *Under the regularity conditions C1-C7 and the additional bandwidth conditions $h_1/h_3 \to 0$ 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 $\tilde{\boldsymbol{\beta}}$ *of* (2.2) *such that*

$$\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(\left(nh_3^3\right)^{-1/2} + h_3^2\right).$$

Theorem 2.4. *With* $nh_3^7 = O(1)$*, under the same conditions as Theorem 2.3, the estimator satisfying the consistency result in Theorem 2.3 has the following asymptotic result*

$$\sqrt{nh_3^3}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 - \frac{h_3^2}{2}J^{-1}M\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dtJ^{-1}LJ^{-1}\right).$$

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

$$\sqrt{nh_3^3}\left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dtJ^{-1}LJ^{-1}\right),$$

where $J = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}))$*,* $L = \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(0|\hat{\boldsymbol{X}}))$*, and* $M = \mathbb{E}(\boldsymbol{Z} f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}}))$*.*

The first term in the convergence rate characterizes the magnitude of the estimation variance, whereas the second term h_3^2 captures the magnitude of the estimation bias. Interestingly, the rate of convergence ($\sqrt{nh_3^3}$) is slower than that of the SPLVC mean regression ($\sqrt{n}$). 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 $n \to \infty$, which is the case we must consider to reduce the influence of the bias from the first stage that may be brought to the second stage. The asymptotic result is identical to that for the feasible situation where $\boldsymbol{\alpha}(U_i)$ is known, implying that the asymptotic bias and variance of the semi-modal estimator $\tilde{\boldsymbol{\beta}}$ are independent of the naive modal estimators under suitable conditions. Similar phenomenon is observed in Ullah et al. (2021), where they presented a pseudo-demodeing method for estimating fixed effects modal coefficients for panel data.

Theorem 2.4 demonstrates that the semi-modal estimator $\tilde{\boldsymbol{\beta}}$ improves the convergence rate of the naive estimator $\hat{\boldsymbol{\beta}}$ to the linear modal regression one, $\sqrt{nh_3^3} = O(n^{-2/7})$, with the MSE-optimal bandwidth choice $h_3 = O(n^{-1/7})$. This finding is also compatible with the standard SPLVC mean regression result. The optimal bandwidth rate in Theorem 2.4 is larger than $n^{-1/8}$. Intuitively, a large bandwidth is required because the parametric coefficients are global parameters. Undersmoothing is also required to remove asymptotic bias, indicating that the estimator $\tilde{\boldsymbol{\beta}}$ can be asymptotically normal, centered at the true value under sufficient conditions. All of the aforementioned bandwidth considerations will be taken into account when selecting bandwidths in practice.

In what follows, with the available semi-modal estimator $\tilde{\boldsymbol{\beta}}$, we provide the consistency and asymptotic theorem for the final modal estimators $\tilde{\boldsymbol{\alpha}}(u)$ and $\tilde{\boldsymbol{b}}(u)$ at a fixed data point u in the interior of Ω. Similar to Theorem 2.4, we need to impose mild bandwidth conditions to ensure that the bias from the previous stage can be asymptotically disregarded and does not affect the

convergence rate of the final modal estimators.

Theorem 2.5. *Under the regularity conditions C1-C7 and the additional bandwidth condition* $h_3/h_5 \to 0$, *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* $(\tilde{\boldsymbol{\alpha}}(u), \tilde{\boldsymbol{b}}(u))$ *of* (2.3) *such that*

i. $\|\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\| = O_p\left((nh_5h_4^3)^{-1/2} + h_4^2 + h_5^2\right)$,

ii. $\|H_5\tilde{\boldsymbol{b}}(u) - H_5\boldsymbol{b}_0(u)\| = O_p\left((nh_5h_4^3)^{-1/2} + h_4^2 + h_5^2\right)$.

Theorem 2.6. *With* $nh_5^5h_4^3 = O(1)$ *and* $nh_5h_4^7 = O(1)$, *under the same conditions as Theorem 2.5, the estimators satisfying the consistency results in Theorem 2.5 have the following asymptotic result*

$$\sqrt{nh_5h_4^3}\left[\begin{pmatrix}\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\\ h_5(\tilde{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u))\end{pmatrix} - \tilde{\Gamma}(u)^{-1}\left(\frac{h_5^2}{2}\tilde{\Lambda}_2(u)\boldsymbol{\alpha}_0^{(2)}(u) - \frac{h_4^2}{2}\tilde{\Lambda}_1(u)\right)\right] \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\int t^2\phi^2(t)dt}{f_U(u)}\tilde{\Gamma}(u)^{-1}\tilde{\Sigma}(u)\tilde{\Gamma}(u)^{-1}\right).$$

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

$$\sqrt{nh_5h_4^3}\begin{pmatrix}\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\\ h_5(\tilde{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u))\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{\int t^2\phi^2(t)dt}{f_U(u)}\tilde{\Gamma}(u)^{-1}\tilde{\Sigma}(u)\tilde{\Gamma}(u)^{-1}\right),$$

where $\tilde{\Gamma}(u) = \mathbb{E}\left[\begin{pmatrix}\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & \mathbf{0}\\ \mathbf{0} & \mu_2\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\end{pmatrix}\Big|U = u\right]$, $\tilde{\Lambda}_1(u) = \mathbb{E}\left[\begin{pmatrix}\boldsymbol{X} f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})\\ \mathbf{0}\end{pmatrix}\Big|U = u\right]$, $\tilde{\Sigma}(u) = \mathbb{E}\left[\begin{pmatrix}v_0\boldsymbol{X}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}}) & \mathbf{0}\\ \mathbf{0} & v_2\boldsymbol{X}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}})\end{pmatrix}\Big|U = u\right]$, *and* $\tilde{\Lambda}_2(u) = \mathbb{E}\left[\begin{pmatrix}\mu_2\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ \mathbf{0}\end{pmatrix}\Big|U = u\right]$.

The same comments made with respect to naive estimators are applicable here as well. The condition $h_3/h_5 \to 0$ suggests that $h_3 \to 0$ is faster than $h_5 \to 0$ as $n \to \infty$, indicating that the bias from the previous stage is negligible when compared to the final stage bias at the order of h_5^2. The intuition behind this condition is that h_3 must fall to zero fast enough to guarantee the effect of the earlier stage biases does not carry over asymptotically to final stage estimation. The asymptotic properties of the final modal estimators are in line with those of the estimators in the varying coefficient modal regression. This observation reveals that the estimators are optimal for the final stage estimation in the sense that they reach the same rates as those whose parametric coefficient $\boldsymbol{\beta}$ is known, which is referred to as the oracle estimators.

Since the MSE-optimal bandwidths of h_4 and h_5 have a rate of $n^{-1/8}$, the estimator's bias can be negligible relative to the variance term by applying the undersmoothing technique, which enables modal estimators to achieve asymptotic normality centered at the true value. As a result, this does not produce a faster rate of convergence in probability. Although the convergence rate and asymptotic bias of $\tilde{\boldsymbol{\alpha}}(u)$ are theoretically the same as those of $\hat{\boldsymbol{\alpha}}(u)$, the third-stage estimator $\tilde{\boldsymbol{\alpha}}(u)$ should be more efficient. Because, unlike $\hat{\boldsymbol{\alpha}}(u)$, which estimates the parametric component locally, the third-stage estimator $\tilde{\boldsymbol{\alpha}}(u)$ does not need to account for the uncertainty of estimating the parametric component that has a faster convergence rate. Similar result is achieved in the SPLVC quantile regression investigated by Kai et al. (2011). Also, the behavior near the boundary, a well-known appealing property of local linear smoothers, can be demonstrated to carry over to the final stage estimation.

2.3 Bandwidth Selection

One practical issue concerning the implementation of the three-stage estimation procedure is the selection of bandwidths. Kernel functions have little effect on estimating in modal regression, whereas bandwidth strongly influences estimation accuracy since it can control the balance between mean and modal estimates (i.e., with a large bandwidth, we can achieve mean estimates). Furthermore, bandwidth is the tuning parameter that regulates the trade-off between bias and variance of the resultant estimator. Some literature has addressed the problem of bandwidth selection under the content of modal regression. For example, Chen et al. (2016) proposed choosing bandwidths by minimizing a loss function defined as the volume of the prediction band; Zhou and Huang (2019) developed two different cross-validation methods for obtaining bandwidths in multimodal regression; and Yao and Li (2014) and Ullah et al. (2021) applied the plug-in method for choosing bandwidths based on expressions of asymptotically optimal bandwidths. However, to the best of our knowledge, there appear to be no results available about selecting the bandwidth in the context of the SPLVC modal regression.

To strike a balance between the computation burden and efficiency of the estimators while minimizing model bias, we suggest a simple rule-of-thumb to select bandwidths in this paper based on the asymptotically optimal rates of bandwidths. The MSE-optimal rate for h_1, h_2, h_4, and h_5 is $n^{-1/8}$, while the MSE-optimal rate for h_3 is $n^{-1/7}$. Despite sharing the same rate, the roles of bandwidths h_1 (or h_4) and h_2 (or h_5) are quite different. Especially, the bandwidths h_2 and h_5 associated with variable U govern the smoothing of the regression function, whereas the bandwidths h_1, h_3, and h_4 associated with the response variable Y affect the number of estimated modes. When h_2 and h_5 are large, the undersmoothed estimations of the regression functions are obtained, and with small values of h_2 and h_5, the oversmoothed estimators are achieved. Based on these, we follow classical nonparametric estimation to fix the bandwidths $h_2 = Cn^{-0.15}$ and $h_5 = Cn^{-0.13}$ in numerical studies because we need to undersmooth the estim-

ators to avoid bias, where C is a constant set as the appropriate rule-of-thumb value $1.06\hat{\sigma}_U$ and $\hat{\sigma}_U$ is the standard deviation of variable U.

We work with the undersmoothing assumption on the bandwidths following Kemp and Santos Silva (2012) and Ullah et al. (2022) to apply the grid search method to select a number of potential bandwidths for h_1, h_3, and h_4, which control the number of modes. To be more specific, we first calculate the mean regression residual $\hat{\epsilon}_{mean}$, and then select 50 bandwidth values between $50MAD$ and $0.5MADn^{-\gamma_h}$ (the values of γ_h for bandwidths h_1, h_3, and h_4 are $\gamma_{h_1} = 0.15$, $\gamma_{h_3} = 0.143$, and $\gamma_{h_4} = 0.13$, respectively), in which MAD is the median value of the absolute deviation of the mean regression residual from the corresponding median value,

$$MAD = med_j\{|(Y_j - \hat{m}(\cdot)) - med_i(Y_i - \hat{m}(\cdot))|\},$$

$\hat{m}(\cdot)$ represents the associated mean estimate, and med denotes the median value. In the empirical analysis, the default bandwidth is $1.6MADn^{-\gamma_h}$. It is crucial to note that the aforementioned approach for selecting bandwidth may not yield best estimates, but it does provide a simple procedure to achieve optimal convergence rates for all estimators, and its satisfactory performance has been demonstrated in practice.

Remark 6. (**Modal Cross-Validation**) *We may adopt some other data-driven methods, such as leave-one-out cross-validation, to choose bandwidths in the proposed SPLVC modal regression. Nevertheless, MSE criterion-based cross-validation is inapplicable for modal regression. We recommend utilizing the cross-validation method based on the fact that the interval around the conditional mode should cover more samples with the same interval length. Particularly, combining with the grid search method, we maximize the following kernel-based equation*

$$CV(bandwidths) = \frac{1}{n\kappa}\sum_{i=1}^{n}\phi\left(\frac{Y_{-i} - \hat{Mode}_{-i}(\cdot)}{\kappa}\right),$$

where $\hat{Mode}_{-i}(\cdot)$ represents the estimated modal function value without the observation indexed by i, and κ is a constant that can be chosen as $MAD\max|Y_i - Y_j|$, $i, j = 1, \cdots, n$. Such a cross-validation method reveals mode characteristic and can make the prediction interval of the same width contain more sample points. The investigation of its asymptotic property, however, is beyond the scope of this paper.

2.4 Varying Coefficient Test

All of the preceding discussions, which provide a solid foundation for developing the SPLVC modal estimates, rely on the correctly specified semiparametric regression model. If the parametric modal regression is a valid specification, but the SPLVC modal regression under consideration is used, the estimation results based on the overspecified semiparametric model would

not only increase the complexity of the model but also decrease estimation accuracy. Most importantly, because of the difference in convergence rates, treating a parametric component as a nonparametric one can result in overfitting the data and efficiency loss. Therefore, it is critical for the proposed SPLVC modal regression to test whether the varying coefficient functions are constant, specifically testing the null hypothesis

$$\mathcal{H}_0 : \alpha_j(u) = \alpha_j \quad \text{for all} \quad j = 1, \cdots, p$$

with the alternative hypothesis $\mathcal{H}_1 : \alpha_j(u)$ varying with u for at least one of $j = 1, \cdots, p$, where α_j is assumed to be an unknown constant.

Inspired by Fan et al. (2001) and Fan and Jiang (2007), which offered a general technique for testing hypotheses regarding nonparametric functions, we construct a goodness-of-fit testing statistic by taking a kernel-based function as the loss function rather than the sum of squared errors. We compare the residual sums of modes under both the null and alternative hypotheses

$$T_0 \overset{\text{def}}{=} L(\mathcal{H}_1) - L(\mathcal{H}_0), \tag{2.4}$$

where the residual sum of modes under $\mathcal{H}_0$ is

$$L(\mathcal{H}_0) = \frac{1}{h_4} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\alpha}^* - \boldsymbol{Z}_i^T \boldsymbol{\beta}^*}{h_4} \right)$$

in which $\alpha_1^*, \cdots, \alpha_p^*$ and $\boldsymbol{\beta}^*$ are the parametric modal estimates, and the residual sum of modes under $\mathcal{H}_1$ is

$$L(\mathcal{H}_1) = \frac{1}{h_4} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T \tilde{\boldsymbol{\beta}}}{h_4} \right).$$

Remark 7. *We construct $L(\mathcal{H}_1)$ using the final estimators $\tilde{\boldsymbol{\alpha}}(U_i)$ and $\tilde{\boldsymbol{\beta}}$ with the same bandwidth h_4 in estimation. Under the null hypothesis, we treat varying coefficient functions as constants and utilize the linear modal regression in Yao and Li (2014) to estimate $\{\alpha_j\}_{j=1}^p$ and β directly by applying the plug-in method for bandwidth selection.*[6] *The key to the success of the developed test, as demonstrated in the following theorems, is to use the same bandwidth for constructing $\mathcal{L}(\cdot)$ functions under $\mathcal{H}_0$ and $\mathcal{H}_1$. Therefore, the bandwidth h_4 is utilized in the construction of $L(\mathcal{H}_0)$ as well to ensure that bandwidths do not affect kernel-based functions when comparing.*

Since the role of the inference function based on the kernel and the least square is comparable, $L(\mathcal{H}_0)$ and $L(\mathcal{H}_1)$ can be regarded as the degree to which the model fits the data under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Intuitively, the values of $L(\mathcal{H}_0)$ and $L(\mathcal{H}_1)$ will be extremely close under

[6] According to Remark 2, we can also use the suggested steps to obtain the estimators $\hat{\alpha}_1(u), \cdots, \hat{\alpha}_p(u)$ by disregarding the fact that $\alpha_1, \cdots, \alpha_p$ are constants and treating them as unknown varying coefficients. After that, we take the average over $\{U_i\}_{i=1}^n$ to have $\alpha_1^*, \cdots, \alpha_p^*$. Then, numerically, $\boldsymbol{\beta}^*$ will be the same as $\tilde{\boldsymbol{\beta}}$.

$\mathcal{H}_0$, while the value of $L(\mathcal{H}_1)$ will be sufficiently greater than $L(\mathcal{H}_0)$ if the alternative hypothesis $\mathcal{H}_1$ is true. Therefore, if T_0 is larger than an approximate critical value, the null hypothesis $\mathcal{H}_0$ is rejected. Defining t_0 as the observed value of T_0, the p-value of the test is followed as

$$p_0 = P_{\mathcal{H}_0}(T_0 > t_0), \tag{2.5}$$

where $P_{\mathcal{H}_0}(\cdot)$ refers to the probability computed under the null hypothesis $\mathcal{H}_0$. For a given significance level α, the null hypothesis $\mathcal{H}_0$ would be rejected if $p_0 < \alpha$; otherwise it would fail to reject $\mathcal{H}_0$. The test is supported by the following theorem.

Theorem 2.7. *Assume that the regularity conditions in Theorem 2.6 are attained. Under* $\mathcal{H}_0$,

$$P_{\mathcal{H}_0}(T_0 \to 0) \to 1 \quad as \quad n \to \infty.$$

Otherwise, if $\inf_{\alpha_l \in R} \|\alpha_l(\cdot) - \alpha_l\| > 0$, *there exists a constant* $t_0 > 0$ *such that*

$$P_{\mathcal{H}_0}(T_0 > t_0) \to 1 \quad as \quad n \to \infty.$$

Technically, we could extend the methodology introduced in Fan et al. (2001) to derive the asymptotic null distribution of the proposed test. Theorem 2.8 shows that the developed statistic T_0 converges in distribution as $n \to \infty$ under reasonable regularity conditions, and the scale r_K and degree of freedom $r_K\mu_n$ of the asymptotic χ^2-distribution are dependent on the unknown density functions. The Wilks phenomenon, which states that the asymptotic distribution is independent of the nuisance parameters, thus does not apply to modal regression. Because we treat h_4 as a shrinking bandwidth, the components in Theorem 2.8 are reliant on h_4. Nonetheless, if h_4 is treated as a constant when conducting the test, the result would be consistent with the classical mean case, i.e., $d_n = O_p(nh_5^4) + O_p(\sqrt{n}h_5^2)$.

Theorem 2.8. *Suppose that all of the conditions in Theorem 2.6 are met. With the additional constraint* $nh_4^3h_5 \to \infty$, *under* $\mathcal{H}_0$,

$$\sigma_n^{-1}(T_0 - \mu_n + d_n) \xrightarrow{d} N(0,1).$$

Furthermore, with a scale $r_K = 2\mu_n/\sigma_n^2$, *under* $\mathcal{H}_0$, *the test statistic* $r_K T_0$ *approximately follows a* χ^2*-distribution*

$$r_K T_0 \sim \chi^2(r_K\mu_n),$$

where " $\sim$ " *denotes generalized approximation,* $d_n = O_p(nh_5^4h_4^{-2}) + O_p(\sqrt{n}h_5^2h_4^{-3/2})$, σ_n *is the asymptotic variance shown in the supplementary note S5, and* $\mu_n = -[h_4^3h_5 f_\epsilon^{(2)}(0|\boldsymbol{X}, U, \boldsymbol{Z}) f_U(U)]^{-1} f_\epsilon(0|\boldsymbol{X}, U, \boldsymbol{Z}) \int t^2\phi^2(t)dt \ \boldsymbol{X}_i^T\boldsymbol{X}_i K(0)$.

However, as many researchers have pointed out, unless the bandwidths are sufficiently small so that the degree of freedom $r_K\mu_n$ is large, the p-value generated from the asymptotic null dis-

tribution of the test statistic may be erroneous in the context of finite sample sizes; see Fan and Jiang (2007). This is especially true for SPLVC modal regression because the asymptotic null distribution of the proposed statistic is dependent on the nuisance parameters. To perform the test, we apply the residual-based bootstrap algorithm with appropriate modal centering adjustments to approximate the null distribution of T_0 and the p-value of the test. In general, with a moderate sample size, the bootstrap method outperforms the asymptotic distribution-based method, as only the main order of the degrees of freedom is given in Theorem 2.8. The performance, including the validity of the bootstrap procedure in approximating the null distribution of the test statistic and the power of the test, is further demonstrated by Monte Carlo simulations in Section 4.

Algorithm 2 Bootstrap Algorithm for Estimating p-value

S1 Based on the data $\{(Y_i, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n$, estimate the SPLVC modal regression and obtain the residual $\tilde{\epsilon}_i = Y_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}$.

S2 Compute the centered-in-mode residual $\tilde{\epsilon}_i^* = \tilde{\epsilon}_i - Mode(\tilde{\epsilon}_i)$ and generate the bootstrap residuals $\{\tilde{\epsilon}_i^*\}_{i=1}^n$ with replacement from the empirical distribution function of $\tilde{\epsilon}_i^*$.

S3 Define $Y_i^* = \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) + \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_i^*$ and calculate the bootstrap test statistic T_0^* based on the samples $\{(Y_i^*, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n$.

S4 Repeat **S2**-**S3** for B (e.g. $B = 200$) times to obtain a bootstrap sample of the test statistics T_0 as $\{T_{0b}^*\}_{b=1}^B$. The p-value is estimated by $\hat{p} = \sum_{b=1}^B I\left(T_{0b}^* \geq t_0\right)/B$. Reject the null hypothesis $\mathcal{H}_0$ when T_0 is greater than the upper-α quantile of $\{T_{0b}^*\}_{b=1}^B$ or $\hat{p} < \alpha$.

We intend to approximate the distribution of T_0 using the sampling distribution of T_0^*, which is justified if T_0^* converges to the same limiting distribution as T_0. The following theorem demonstrates the consistency of the above bootstrap testing procedure.

Theorem 2.9. *Suppose that all of the conditions in Theorem 2.6 are fulfilled. Under $\mathcal{H}_0$, $r_K T_0^* \sim \chi^2(r_K\mu_n)$ and*

$$\sup_{z\in\mathbb{R}}|P(Z_0^* \leq z|\{(\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)\}_{i=1}^n) - P(N(0,1) \leq z)| \xrightarrow{p} 0,$$

in which $Z_0^* = \sigma_n^{-1}(T_0^* - \mu_n + d_n)$.

Remark 8. *We bootstrap the centralized residuals from the SPLVC modal regression rather than the residuals from the parametric modal regression as in Cai et al. (2000), which considered the goodness-of-fit test for the varying coefficient nonlinear time series model. Regardless of whether the null or alternative hypothesis is correct, the SPLVC modal estimate of the residual can always be consistent. However, since we are concentrating on modal regression, we need to calculate the centered-in-mode instead of the centered-in-mean residual to ensure $Mode(\hat{\epsilon}_i^*) = 0$. We uti-*

lize the same bandwidths in calculating T_0^* (including estimators) as in T_0, which satisfies the conditions on the bandwidths used in the optimal asymptotic performance of the proposed test.

We consider the test for all varying coefficient functions for concreteness. The same steps can be applied to test $\alpha_j(u) = \alpha_j$ for a subset of the index $j = 1, \cdots, p$. In addition, since the varying coefficient functions may be known in certain applications, the proposed goodness-of-fit testing statistic can be used to determine whether the varying coefficient functions are of some specific functional forms, where

$$L(\mathcal{H}_0) = \frac{1}{h_4} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\alpha}_0(U_i) - \boldsymbol{Z}_i^T \tilde{\boldsymbol{\beta}}}{h_4} \right)$$

and $\boldsymbol{\alpha}_0(U_i)$ is the true function under the null hypothesis. Furthermore, we can develop a test for the situation where α_j is a known constant. In such a case, $L(\mathcal{H}_0)$ needs to be modified to

$$L(\mathcal{H}_0) = \frac{1}{h_4} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\alpha}_0 - \boldsymbol{Z}_i^T \tilde{\boldsymbol{\beta}}}{h_4} \right)$$

with $\boldsymbol{\alpha}_0$ representing a known constant vector. Nonetheless, this may not be attractive in practice because researchers are more concerned with whether the varying coefficient functions are indeed constant without knowing specific values.

Remark 9. (Wald-Type Test) *It is natural to investigate the Wald test by directly examining the variability of the estimated coefficient $\tilde{\boldsymbol{\alpha}}(u)$. Given the null restricted modal regression Mode $(Y_i|\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T) = \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{Z}_i^T \boldsymbol{\beta}$, based on the results from Theorem 2.4, it can be seen that under certain regularity conditions, $\|\boldsymbol{\alpha}^* - \boldsymbol{\alpha}\| = O_p(\tilde{h}^2 + (n\tilde{h}^3)^{-1})$, where bandwidth $\tilde{h}$ is used in the modal regression under $\mathcal{H}_0$. If we choose $nh_5^5 h_4^3 \to 0$ and $nh_5 h_4^7 \to 0$, under $\mathcal{H}_0$, it can be obtained from Theorem 2.6 that*

$$\begin{aligned} \sqrt{nh_5 h_4^3}(\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}^*) &= \sqrt{nh_5 h_4^3}(\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}) - \sqrt{nh_5 h_4^3}(\boldsymbol{\alpha}^* - \boldsymbol{\alpha}) \\ &= \sqrt{nh_5 h_4^3}(\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}) + o_p(1) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Xi}(u)), \end{aligned}$$

where $\boldsymbol{\Xi}(u) = (\int t^2 \phi^2(t) dt / f_U(u))[\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|U = u)]^{-1}[\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T f_\epsilon(0|\hat{\boldsymbol{X}})|U = u)][\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|U = u)]^{-1}$. Following Yao and Li (2014) and Ullah et al. (2021) to consistently estimate $f_\epsilon(0|\hat{\boldsymbol{X}})$ and the corresponding derivatives, we can get the consistent estimate for the asymptotic covariance matrix $\boldsymbol{\Xi}(u)$, which is defined as $\hat{\boldsymbol{\Xi}}(u)$. We then have

$$\sqrt{nh_5 h_4^3} \hat{\boldsymbol{\Xi}}(u)^{-1}(\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}^*) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p),$$

where $\boldsymbol{I}_p$ is an identity matrix with dimension $p \times p$, and $\|\sqrt{nh_5 h_4^3} \hat{\boldsymbol{\Xi}}(u)^{-1}(\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}^*)\|^2 \xrightarrow{d} \chi^2(p)$. We finally can construct an asymptotically valid test for $\boldsymbol{\alpha}(u)$ across a range of u such that

$$T_0^* = \max_{1\le i\le n} \|\sqrt{nh_5h_4^3}\hat{\boldsymbol{\Xi}}(U_i)^{-1}(\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{\alpha}^*)\|^2 \xrightarrow{d} \max_{1\le i\le n} \chi_i^2(p).$$

Since the limiting distribution of T_0^* is a functional of independent χ^2 random variables (with p degrees of freedom) that is free of nuisance parameters, the critical value of T_0^* can be calculated.

3 Penalized SPLVC Modal Regression

Variable selection is crucial in high-dimensional econometrics since sparse modeling is often preferred due to enhanced model predictability and interpretability. This section utilizes the penalty function to simultaneously estimate parametric coefficients in the SPLVC modal regression and shrink some coefficients to zero. The theoretical properties of the procedure, including the consistency in variable selection and the oracle property in estimation, are established with appropriate choice of the tuning parameter.

3.1 Penalized Modal Estimators

Under the assumption that the dimension k of the parameter $\boldsymbol{\beta}$ is fixed, we propose the following penalized kernel-based objective function to conduct variable selection

$$\mathcal{L}_P(\boldsymbol{\beta}) = \frac{1}{h_3}\sum_{i=1}^{n}\phi_3\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right) - n\sum_{j=1}^{k} p_{\lambda_j}(|\beta_j|), \tag{3.1}$$

where $p_{\lambda_j}(\cdot)$ is the penalty function with a tuning parameter λ_j that includes commonly used penalty functions, such as the least absolute shrinkage and selection operator (LASSO), adaptive LASSO, smoothed clipped absolute deviation (SCAD), among others. Fan and Li (2001) studied the choice of penalty functions in depth and advocated for the use of a nonconcave penalty. The regularization parameters $\{\lambda_j\}_{j=1}^k$ are not necessarily the same for all j, which offers the flexibility of producing different shrinkage for different modal coefficients to keep some important variables in the final model. Practically, the regularization parameters can be chosen using the data-driven criterion BIC shown in Algorithm 3.

There are numerous penalty functions available for conducting variable section; see Su and Zhang (2013). In order to perform variable selection for the SPLVC modal regression in a computationally efficient manner, this section shall use the SCAD penalty function for the calculation. According to Fan and Li (2001), the SCAD penalty is defined as

$$p_{\lambda_j}(|t|) = \lambda_j(|t|)\{I((|t|) < \lambda_j)\} + \frac{(a - |t|/2\lambda_j)}{a-1}I(\lambda_j < (|t|) < a) + \frac{a^2\lambda_j}{2(a-1)(|t|)}I\{(|t|) \ge a\lambda_j\}$$

for some constant $a > 2$, where the first derivative $p_{\lambda_j}^{(1)}(0) = 0$, λ_j is a penalized parameter con-

trolling the trade-off between data fit and estimate roughness, and $a = 3.7$ as suggested in Fan and Li (2001) from a Bayesian perspective. The SCAD penalty function is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$ but singular at zero. Its derivative vanishes outside of $[-a\lambda_j, a\lambda_j]$. Consequently, it can obtain sparse solutions and unbiased estimates for large datasets by shrinking small coefficients toward zeros.

Remark 10. *To obtain the oracle property, the penalty function $p_{\lambda_j}(\cdot)$ used for the SPLVC modal regression should satisfy the following properties: (a) for nonzero fixed t, $\lim(nh_3^3)^{1/2}$ $p_{\lambda_j}(|t|) = 0$ and $\lim p_{\lambda_j}^{(1)}(|t|) = 0$; (b) for any $C > 0$, $\lim(nh_3^3)^{1/2} \inf_{|t| \le C(nh_3^3)^{-1/2}} p_{\lambda_j}(|t|) \to \infty$. Evidently, the SCAD penalty meets these two properties, which is the penalty we concentrate on in the main part of the paper. Note that adaptive LASSO can also be utilized for conducting variable selection, as illustrated in Remark 11.*

Since $\boldsymbol{\alpha}(U_i)$ consists of unknown nonparametric functions, (3.1) is not ready for optimization. We substitute the resulting estimate from the first-stage estimation into (3.1) and obtain the penalized equation shown below.

$$\mathcal{L}_P(\boldsymbol{\beta}) = \frac{1}{h_3} \sum_{i=1}^{n} \phi_3 \left(\frac{Y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T \boldsymbol{\beta}}{h_3} \right) - n \sum_{j=1}^{k} p_{\lambda_j} \left(|\beta_j| \right). \tag{3.2}$$

Optimizing (3.2) is a simultaneous estimation and variable selection procedure. However, it is not easy for achieving solutions because the penalty function is irregular at the origin and does not have a second derivative at some points. To tackle the challenging estimation problem, given an initial value of $\beta^{(0)}$ close to the maximizer of (3.2), we follow Fan and Li (2001) to apply a locally quadratic approximation for the penalty function such that

$$p_{\lambda_j} \left(|\beta_j| \right) \approx p_{\lambda_j} \left(|\beta_j^{(0)}| \right) + \frac{1}{2} \left\{ \frac{p_{\lambda_j}^{(1)} \left(|\beta_j^{(0)}| \right)}{|\beta_j^{(0)}|} \right\} \left(\beta_j^2 - \beta_j^{(0)2} \right) \text{ for } \beta_j \approx \beta_j^{(0)}, \tag{3.3}$$

in which $\beta_j^{(0)}$ is not very close to 0. Replace $p_{\lambda_j} \left(|\beta_j| \right)$ in (3.2) with (3.3), we can obtain

$$\mathcal{L}_P(\boldsymbol{\beta}) = \frac{1}{h_3} \sum_{i=1}^{n} \phi \left(\frac{Y_i - \boldsymbol{X}_i^T \hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T \boldsymbol{\beta}}{h_3} \right) - \frac{n}{2} \sum_{j=1}^{k} \left\{ \frac{p_{\lambda_j}^{(1)} \left(|\beta_j^{(0)}| \right)}{|\beta_j^{(0)}|} \right\} \beta_j^2. \tag{3.4}$$

By maximizing the above objective function with a proper penalty parameter λ_j, we can perform an automatic variable selection with a sparse estimator of $\boldsymbol{\beta}$, defined as $\hat{\boldsymbol{\beta}}^P$. Notice that for identifying variables in the nonparametric components, we can apply the aforementioned varying coefficient test by replacing $\boldsymbol{\beta}$ with the penalized estimator.

The sparse estimator is expected to possess the selection invariance property under regu-

larity conditions. If we have additional information regarding which components are zero, say $\beta_j = 0,\ j = s+1, \cdots, k$, we could take into account this prior knowledge to maximize the following constrained kernel-based objective function with all irrelevant variables removed.

$$\max_{\beta} \mathcal{L}_P(\boldsymbol{\beta}) = \frac{1}{h_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right) - \frac{n}{2}\sum_{j=1}^{s}\left\{\frac{p_{\lambda_j}^{(1)}\left(|\beta_j^{(0)}|\right)}{|\beta_j^{(0)}|}\right\}\beta_j^2,$$
$$\text{s.t. } \beta_j = 0,\ j = s+1, \cdots, k.$$

With approximate Karush-Kuhn-Tucker (KKT) condition

$$-\frac{Z_{ji}}{h_3}\phi^{(1)}\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}^{(rs)}}{h_3}\right) = o\left(n\left\{\frac{p_{\lambda_j}^{(1)}\left(|\beta_j^{(0)}|\right)}{|\beta_j^{(0)}|}\right\}\right),$$

where $\hat{\boldsymbol{\beta}}^{(rs)}$ is the solution to this constraint problem, we can have $(\hat{\boldsymbol{\beta}}^{(rs)}, \beta_{s+1}, \cdots, \beta_k) = \hat{\boldsymbol{\beta}}^P$ with probability converging to one.

Algorithm 3 MEM Algorithm for Penalized SPLVC Modal Regression

Selection of λ_j. Set $\lambda_j = \lambda SE(\hat{\beta}_j)$, where $SE(\hat{\beta}_j)$ is the standard error from (2.2) (i.e., bootstrap method). Then, use BIC to select λ

$$\lambda_{opt} = \arg\min_{\lambda} BIC(\lambda) = -\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\hat{\boldsymbol{\beta}}^P}{h_3}\right) + \frac{\log(nh_3^3)}{nh_3^3}df_\lambda,$$

where df_λ is the number of nonzero coefficients of $\hat{\boldsymbol{\beta}}^P$ with tuning parameter λ.

E-Step. Update weight (posterior conditional probability) $\pi\left(i|\boldsymbol{\beta}^{P(g)}\right)$ as

$$\pi\left(i|\boldsymbol{\beta}^{P(g)}\right) = \frac{\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^{P(g)}}{h_3}\right)}{\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}^{P(g)}}{h_3}\right)}.$$

M-Step. Update the value of $\boldsymbol{\beta}^{P(g+1)}$ with the weight calculated in E-Step by

$$\boldsymbol{\beta}^{P(g+1)} = \arg\max_{\boldsymbol{\beta}}\sum_{i=1}^{n}\left\{\pi\left(i|\boldsymbol{\beta}^{P(g)}\right)\log\left(\frac{1}{h_3}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right)\right)\right.$$
$$\left. - \frac{n}{2}\sum_{j=1}^{k}\left\{\frac{p_{\lambda_j}^{(1)}\left(|\beta_j^{P(g)}|\right)}{|\beta_j^{P(g)}|}\right\}\beta_j^2\right\} = (\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + n\Sigma_\lambda(\boldsymbol{\beta}^{P(g)}))^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y},$$

where $\Sigma_\lambda(\boldsymbol{\beta}^{P(g)}) = \text{diag}\left\{\frac{p_{\lambda_1}^{(1)}\left(|\beta_1^{P(g)}|\right)}{|\beta_1^{P(g)}|}, \cdots, \frac{p_{\lambda_k}^{(1)}\left(|\beta_k^{P(g)}|\right)}{|\beta_k^{P(g)}|}\right\}$.

As there are no available closed-form solutions for (3.4), we extend the modified MEM algorithm shown in Algorithm 3 to estimate (3.4). The E-step computes the weight provided that the current estimate gives the true parameter of the model, and the M-step updates the estimate through computationally simpler conditional maximization, which can result in a closed-form expression due to the use of a normal kernel $\phi(\cdot)$ and the quadratic approximation. The estimation process is repeated iteratively until convergence or stopping criterion is reached; see the related comments for Algorithm 1.[7]

The first term in the above BIC selection equation can be regarded as an "artificial" likelihood since it exhibits certain essential properties of a parametric log-likelihood. Due to the target of modal estimation, the effective sample size would be nh_3^3 rather than n in the classical mean regression. This is because the classical mean estimation is $\sqrt{n}$-consistent, while the convergence rate of modal estimation is $\sqrt{nh_3^3}$. Suppose that S_T denotes the true model and S_λ represents the set of the indices of the covariates selection by penalized modal regression with tuning parameter λ. We can construct a sequence of reference tuning parameters $\lambda_n = \log(nh_3^3)/\sqrt{nh_3^3}$ (i.e., $\lambda_n \to 0$ and $\sqrt{nh_3^3}\lambda_n \to \infty$). Because the penalty estimator $\hat{\boldsymbol{\beta}}^P_{\lambda_n}$ is identical to the oracle estimator, given the consistent estimator $\hat{\boldsymbol{\alpha}}(U_i)$, it follows immediately that $P(BIC_{\lambda_n} = BIC_{S_T}) \to 1$. By adopting the results presented in Wang et al. (2007) and Subsection 2.4, we can verify that $P(\inf_{\lambda \in \Omega_- \cup \Omega_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$ under mild conditions, where Ω_- and Ω_+ denote the underfitting and overfitting cases, respectively. This means that those λ's which fail to identify the true model cannot be selected by BIC asymptotically. As a result, we have $P(S_{\lambda_{opt}} = S_T) \to 1$.

Remark 11. (Identification and Variable Selection) *Besides the proposed goodness-of-fit test for identifying nonparametric coefficient functions and penalized SPLVC modal regression for selecting parametric coefficients, we can utilize penalization methods to directly identify the true structure of the SPLVC modal regression. Assume that $\boldsymbol{\beta}$ in (1.3) is a vector consisting of unknown functions of U_i. We can re-write (1.3) as a varying coefficient modal regression Mode $(Y_i|\mathbf{W}_i) = \mathbf{W}_i^T\boldsymbol{\theta}(U_i)$, where $\mathbf{W}_i = (\mathbf{X}_i^T, \mathbf{Z}_i^T)^T$ and $\boldsymbol{\theta}(U_i) = (\boldsymbol{\alpha}^T(U_i), \boldsymbol{\beta}^T)^T$. We then integrate local linear approximation with adaptive LASSO*

$$\frac{1}{nh_1^*h_2^*}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{W}_i^T\boldsymbol{\theta}(u) - \boldsymbol{W}_i^T\boldsymbol{\theta}^{(1)}(u)(U_i-u)}{h_1^*}\right)K\left(\frac{U_i-u}{h_2^*}\right) - \lambda_n\sum_{j=1}^{p+k}\frac{|\theta_j(u)|}{w_j} - \gamma_n\sum_{j=1}^{p+k}\frac{|\theta_j^{(1)}(u)|}{v_j}$$

to simultaneously identify whether a coefficient is parametric and select significant covariates in both nonparametric and parametric portions, where h_1^* and h_2^* are two bandwidths, $\boldsymbol{\theta}^{(1)}(\cdot)$ is the first derivative of $\boldsymbol{\theta}(\cdot)$, λ_n and γ_n are two tuning parameters, and w_j and v_j are two determined positive random quantities. By shrinking the first derivative of the varying coefficient function

[7] Penalized variable selection procedures are usually too conservative in the sense that they may over-shrink component estimates. To solve this problem in the SPLVC modal regression, we can further use the selected variables to re-estimate the non-penalty coefficients to improve the final performance of the estimates.

to zero, a parametric component can be detected. Following the results in this paper, the oracle properties of the nonzero coefficient function estimators can be established. Further theoretical analysis is needed in the future to develop such a variable selection model.

3.2 Large Sample Properties

Large sample properties of shrinkage estimation with the SCAD penalty, i.e., consistent variable selection and oracle property for parameter estimation, have been well established in the literature (Fan and Li, 2001; Kai et al., 2011; Zhang et al., 2013). In this subsection, we show that these theoretical results can be extended to the SPLVC modal regression scenario.

Note that maximizing (3.4) will result in a penalized modal regression estimator $\hat{\boldsymbol{\beta}}^P$. To investigate the asymptotic properties of the shrinkage modal estimator, we decompose the modal regression coefficient vector $\boldsymbol{\beta}_0$ into $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0'}^T, \boldsymbol{\beta}_{0''}^T)^T \in \mathbb{R}^k$ without loss of generality, where $\boldsymbol{\beta}_{0'} = (\beta_{01}, ..., \beta_{0s})^T \in \mathbb{R}^s$ consists of all nonzero components of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{0''} = (\beta_{0s+1}, ..., \beta_{0k})^T \in \mathbb{R}^{k-s}$ is made up of all of the zero components of $\boldsymbol{\beta}_0$. Define

$$a_n = \max_{1\le j\le k}\left\{|p_{\lambda_j}^{(1)}(|\beta_{0j}|)| : \beta_{0j} \neq 0\right\},\ b_n = \max_{1\le j\le k}\left\{|p_{\lambda_j}^{(2)}(|\beta_{0j}|)| : \beta_{0j} \neq 0\right\},$$
$$\Psi_\lambda = \left(p_{\lambda_1}^{(1)}(|\beta_{01}|), \ldots, p_{\lambda_s}^{(1)}(|\beta_{0s}|)\right)^T, \text{ and } \Phi_\lambda = \operatorname{diag}\left\{p_{\lambda_1}^{(2)}(|\beta_{01}|), \ldots, p_{\lambda_s}^{(2)}(|\beta_{0s}|)\right\},$$

where $p_{\lambda_j}^{(2)}(\cdot)$ is the second derivative of penalty. We can then establish the following theoretical properties about the consistency and sparsity property of the penalized modal estimator of the parametric part.

Theorem 3.1. *Suppose that the regularity conditions in Theorem 2.4 are fulfilled. With probability approaching one, as $b_n \to 0$ with $n \to \infty$, there exists a consistent maximizer $\hat{\boldsymbol{\beta}}^P$ of (3.4) such that*

$$\|\hat{\boldsymbol{\beta}}^P - \boldsymbol{\beta}_0\| = O_p\left((nh_3^3)^{-1/2} + h_3^2 + a_n\right).$$

The rate of convergence of the proposed penalized modal estimator in Theorem 3.1 is dependent on λ_j and bandwidth h_3. As a result, we can further improve the convergence rate to $\|\hat{\boldsymbol{\beta}}^P - \boldsymbol{\beta}_0\| = O_p\big((nh_3^3)^{-1/2} + h_3^2\big)$ with a slightly stronger assumption $\lambda_{max} = max_j\{\lambda_j\} \to 0$ $(i.e., a_n = 0)$. This demonstrates that the consistent penalized modal estimator indeed exists with probability tending to one.

Theorem 3.2. *Under the same conditions as Theorem 3.1, let $\delta_n = h_3^2 + (nh_3^3)^{-1/2}$ and $\lambda_{min} = min_j\{\lambda_j\}$, if $\lambda_{max} \to 0$, $\delta_n^{-1}\lambda_{min} \to \infty$ when $n \to \infty$, and $\liminf_{n\to 0}\liminf_{\beta_j\to 0^+} p_{\lambda_j}^{(1)}(|\beta_j|)/\lambda_j > 0$ for all j, then the penalized modal estimator can correctly identify all zero elements; that is*

$$P\left(\hat{\boldsymbol{\beta}}_{0''}^P = 0\right) \to 1.$$

Theorem 3.1 demonstrates the existence of the penalized modal estimator $\hat{\boldsymbol{\beta}}^P$ that converges to the true parameter at the rate of $O_p((nh_3^3)^{-1/2} + h_3^2 + a_n)$, indicating the dependence on the penalty function and the regularization parameter λ_j. It shows that the difference between the modal estimate with SCAD penalty and the true parameter is asymptotically negligible when λ_j is small enough. Theorem 3.2 states that the proposed penalized modal regression possesses the sparsity property; that is, by choosing an appropriate regularization parameter λ_j, the penalized modal estimator estimates a zero coefficient exactly as zero with a probability tending to one. According to the preceding two theorems, it is apparent that penalized modal estimator can achieve the optimal convergence rate for nonzero coefficients in large samples as if the subset of true zero coefficients were known.

Theorem 3.3. *With $nh_3^7 = O(1)$ and $nh_3^3\Psi_\lambda^2 = O(1)$, under the same conditions as Theorem 3.2, the estimator satisfying the consistency result in Theorem 3.1 has the following asymptotic result*

$$\sqrt{nh_3^3}(J_{(1)} + \Phi_\lambda)\left(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'} + (J_{(1)} + \Phi_\lambda)^{-1}\left(\Psi_\lambda - \frac{h_3^2}{2}M_{(1)}\right)\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dtL_{(1)}\right).$$

In addition, if $\sqrt{nh_3^3}\Psi_\lambda = o_p(1)$ and $\Phi_\lambda = o_p(1)$, we can obtain

$$\sqrt{nh_3^3}J_{(1)}\left(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'} - \frac{h_3^2}{2}J_{(1)}^{-1}M_{(1)}\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dtL_{(1)}\right).$$

Furthermore, if $nh_3^7 \to 0$, we have

$$\sqrt{nh_3^3}J_{(1)}\left(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'}\right) \xrightarrow{d} \mathcal{N}\left(0, \int t^2\phi^2(t)dtL_{(1)}\right),$$

where $J_{(1)}, M_{(1)}$, and $L_{(1)}$ are the $s \times s$ submatrices of J, M, and L corresponding to the nonzero components $\boldsymbol{\beta}_{0'}$, respectively.

In Theorem 3.3, we establish the asymptotic distributions of the resultant estimators for nonzero coefficients under suitable conditions, demonstrating that $\hat{\boldsymbol{\beta}}_{0'}^P$ has the oracle property, i.e., performs as well as an oracle estimator in the asymptotic sense (the estimators for the non-zero coefficients in the true model have the same asymptotic distribution as if the subset of true zero coefficient $\boldsymbol{\beta}_{0''}$ were already known or as if the true underlying model were given in advance). Then, according to the oracle properties, most statistical inferences for $\hat{\boldsymbol{\beta}}^P$ can be constructed exactly the same as the oracle estimator. Theorem 3.3 also indicates that undersmoothing is necessary to remove the asymptotic bias. Because the penalized modal estimator shares the same convergence rate as the parametric estimator $\tilde{\boldsymbol{\beta}}$, the suggested bandwidth in Subsection 2.3 can be adopted here as well.

Remark 12. *The theoretical results in this section are limited to the finite-parameter setting,*

which means that the dimension k of the parameter $\boldsymbol{\beta}$ is fixed. In a general setup, when the dimension of the parametric components is large, it is more realistic to regard it growing with sample size, that is, $k = k(n) \to \infty$ as $n \to \infty$. We can then establish the oracle property by requiring the tuning parameters h_3 and λ_j to approach zero with the rate depending on n and k.

4 Numerical Examples

We in this section investigate the finite sample performance of the developed estimation methods with Monte Carlo simulation studies and two real data analyses. Throughout the section, $\phi(\cdot)$ and $K(\cdot)$ are fixed as the normal kernel $(1/h\sqrt{2\pi})exp\,(-(\cdot)^2/2h^2)$, in which h is the bandwidth, and the word "SPLVC" is suppressed for regression whenever no confusion is caused. Some additional numerical results are contained in the supplementary note.

4.1 Monte Carlo Experiments

(1) (SPLVC Modal Regression) We carry out simulation experiments to illustrate the finite sample performances of the proposed estimators in this part, where two Monte Carlo experiments with different skewed error distributions are conducted (one of which is provided in the supplementary note *S4*). We use DGP to represent the data generating process in what follows, and compare the modal estimates to those of mean regression, which serves naturally as a competitor here. The sample sizes we consider are $n \in \{100, 200, 400, 600, 1000\}$. A total of M =200 simulation replications are conducted in all simulations.[8] We use the square root of average squared errors ($RASE$) to assess the performance of the nonparametric estimator $\tilde{\boldsymbol{\alpha}}(u)$

$$RASE(\tilde{\boldsymbol{\alpha}}(U_i)) = \left(\frac{1}{Mn} \sum_{j=1}^{M} \sum_{i=1}^{n} \|\tilde{\boldsymbol{\alpha}}^{(j)}(U_i) - \boldsymbol{\alpha}_0(U_i)\|^2 \right)^{1/2},$$

where $\tilde{\boldsymbol{\alpha}}^{(j)}(U_i)$ is the estimate in the jth replication, and utilize the generalized mean squared errors ($GMSE$) to evaluate the parametric component $\boldsymbol{\beta}$

$$GMSE(\tilde{\boldsymbol{\beta}}) = (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

We also provide the standard error and MSE for each parameter estimate. In accordance with Ullah et al. (2021, 2022), we present the shape of the empirical density of the standardized modal estimate to check the asymptotic normality property, as well as the coverage probabilities to measure the prediction performance of the newly introduced model.

[8]Taking into consideration the time required to conduct simulations to estimate semiparametric modal and mean regressions and many investigated simulation settings, we choose 200 as simulation replications. We try some simulations with 500 replications and the observed pattens do not change much.

DGP 1 We first generate random samples from the following model to illustrate the application of SPLVC modal regression

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i)\epsilon_i,$$

and ϵ_i follows a mixture normal distribution $0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$, which is skewed left with $\mathbb{E}(\epsilon)$=0 and $Mode(\epsilon)$=1 (Yao and Li, 2014; Ullah et al., 2021, 2022). We set the parameters and varying coefficient functions to $\boldsymbol{\beta} = (1, 2)^T$ and $\boldsymbol{\alpha}(U_i) = (\alpha_1(U_i), \alpha_2(U_i))^T$, in which $\alpha_1(U_i)$ = $exp(2U_i - 1)$ and $\alpha_2(U_i) = sin(2\pi U_i)$. The index variable U_i is simulated from the uniform distribution $U[0, 1]$. The covariate vector $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$ is normally distributed with mean 0, variance $I_{4\times4}$, and correlation $0.2^{|k-j|}$, where $k, j = 1, 2, 3, 4$. We consider three cases, where in case 1 we let $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = X_{1i} + Z_{1i}$, in case 2 we define $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = X_{1i}$, and in case 3 we allow $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = Z_{1i}$. We then have the following equations.

$$\text{Case 1 :} \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + Z_{1i} + 2Z_{2i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} + 2Z_{2i}; \end{cases}$$

$$\text{Case 2 :} \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + Z_{1i} + 2Z_{2i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 1) + X_{2i}sin(2\pi U_i) + Z_{1i} + 2Z_{2i}; \end{cases}$$

$$\text{Case 3 :} \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + Z_{1i} + 2Z_{2i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} + 2Z_{2i}. \end{cases}$$

The simulation results are summarized in Table 1, demonstrating that the developed estimation method is capable of estimating the modal regression effectively with finite samples. For case 1, the modal estimators of β_1 are slightly biased for small n, but there are substantial improvements with an increase in n. For case 2, when the error term is independent of $\boldsymbol{Z}$ and X_2, the proposed method works well even with small n. For case 3 in which the error term is independent of $\boldsymbol{X}$ and Z_2, the modal estimator of β_1 is kind of biased with finite samples. In all three cases, the modal estimators of β_2 are centered around the true parameter values. In addition, the values of $GMSEs$ and $RASEs$ obtained from modal regression are clearly smaller than those achieved from mean regression, indicating that when dealing with skewed data, modal regression (at least) in this example can provide more accurate estimators compared to mean regression. Moreover, as expected, when sample size n increases, the values of $MSEs$, $GMSEs$, and $RASEs$ are essentially decreased, corroborating the asymptotic theories.

Figure 2 depicts a set of varying coefficient estimators, which noticeably indicates that the suggested estimation procedure can capture the true varying coefficient functions in modal regression with finite samples and that the approximation accuracy increases with sample size.[9]

Table 1: The Results of Simulations for SPLVC Regressions

Method	Case	n	β_1	$MSE(\beta_1)$	β_2	$MSE(\beta_2)$	$GMSE(\boldsymbol{\beta})$	$RASE(\boldsymbol{\alpha}(U_i))$
Mode	Case 1 $\beta_{1,mode}=2$ $\beta_{2,mode}=2$	100	1.8046 (0.3079)	0.1325	1.9831 (0.2372)	0.0563	0.1829	0.9573
		200	1.9030 (0.1815)	0.0422	2.0089 (0.1404)	0.0197	0.0546	0.6132
		400	1.9390 (0.1167)	0.0173	1.9943 (0.0890)	0.0079	0.0228	0.5371
		600	1.9188 (0.0972)	0.0160	1.9934 (0.0678)	0.0046	0.0187	0.5135
		1000	1.9465 (0.0674)	0.0074	2.0007 (0.0562)	0.0031	0.0091	0.5012
	Case 2 $\beta_{1,mode}=1$ $\beta_{2,mode}=2$	100	1.0183 (0.1476)	0.0220	1.9975 (0.1400)	0.0195	0.0313	0.6822
		200	0.9960 (0.0934)	0.0087	2.0053 (0.0898)	0.0081	0.0140	0.5701
		400	0.9986 (0.0661)	0.0044	2.0025 (0.0619)	0.0038	0.0063	0.4689
		600	0.9957 (0.0496)	0.0025	1.9971 (0.0543)	0.0029	0.0044	0.3907
		1000	1.0052 (0.0393)	0.0016	1.9967 (0.0375)	0.0014	0.0022	0.3292
	Case 3 $\beta_{1,mode}=2$ $\beta_{2,mode}=2$	100	1.8448 (0.2329)	0.0781	1.9832 (0.1440)	0.0209	0.0902	0.5876
		200	1.8726 (0.1352)	0.0344	2.0028 (0.1008)	0.0101	0.0385	0.5271
		400	1.8774 (0.1001)	0.0250	1.9971 (0.0695)	0.0048	0.0281	0.4128
		600	1.8717 (0.0794)	0.0227	1.9970 (0.0511)	0.0026	0.0250	0.3240
		1000	1.8892 (0.0597)	0.0158	2.0020 (0.0402)	0.0016	0.0162	0.2871
Mean	Case 1 $\beta_{1,mean}=1$ $\beta_{2,mean}=2$	100	0.9898 (0.6519)	0.4229	2.0428 (0.6117)	0.3741	0.6377	1.6321
		200	0.9808 (0.4686)	0.2188	2.0091 (0.4068)	0.1648	0.3105	1.1480
		400	0.9790 (0.3046)	0.0927	1.9969 (0.3029)	0.0913	0.1603	0.8065
		600	0.9897 (0.2497)	0.0621	1.9828 (0.2475)	0.0613	0.1018	0.6554
		1000	1.0329 (0.2078)	0.0440	2.0078 (0.1865)	0.0347	0.0733	0.5289
	Case 2 $\beta_{1,mean}=1$ $\beta_{2,mean}=2$	100	1.0026 (0.3382)	0.1138	2.0187 (0.3259)	0.1060	0.1685	0.9865
		200	0.9994 (0.2276)	0.0516	1.9969 (0.2231)	0.0495	0.0755	0.7062
		400	0.9839 (0.1447)	0.0211	2.0001 (0.1559)	0.0242	0.0376	0.5054
		600	0.9918 (0.1217)	0.0148	1.9848 (0.1347)	0.0183	0.0269	0.4224
		1000	1.0169 (0.1048)	0.0112	2.0007 (0.1058)	0.0111	0.0185	0.3472
	Case 3 $\beta_{1,mean}=1$ $\beta_{2,mean}=2$	100	0.9880 (0.4185)	0.1744	2.0247 (0.3383)	0.1145	0.2458	0.8679
		200	0.9806 (0.3108)	0.0965	2.0122 (0.2330)	0.0542	0.1279	0.6182
		400	0.9940 (0.2089)	0.0435	1.9963 (0.1705)	0.0289	0.0663	0.4455
		600	0.9988 (0.1680)	0.0281	1.9980 (0.1297)	0.0167	0.0383	0.3693
		1000	1.0156 (0.1326)	0.0178	2.0074 (0.0992)	0.0098	0.0268	0.3099

Note: For case 1 and case 2, $\alpha_{1,mode}(U_i) = exp(2U_i - 1) + 1$ and $\alpha_{2,mode}(U_i) = sin(2\pi U_i)$; for case 3, $\alpha_{1,mode}(U_i) = exp(2U_i - 1)$ and $\alpha_{2,mode}(U_i) = sin(2\pi U_i)$. For all cases, $\alpha_{1,mean}(U_i) = exp(2U_i - 1)$ and $\alpha_{2,mean}(U_i) = sin(2\pi U_i)$.

To evaluate the asymptotic normality of the modal estimator, we compare the shape of the empirical density of the standardized modal estimate to that of the standard normal density.

[9]For space consideration, we only report the results for sample sizes 200 and 400 (same for the DGP 2 in the supplementary note *S4*). The results are comparable across sample sizes.

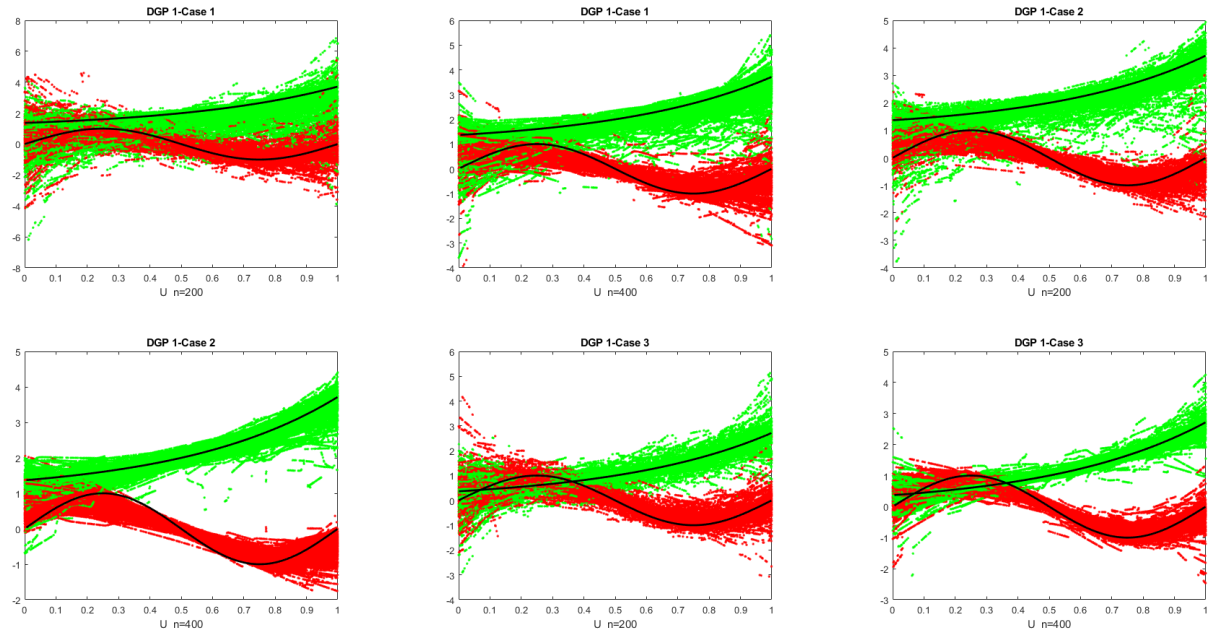

Figure 2: Fitted Varying Coefficient Functions with $n = 200$ or 400

Note: The black curves are the true (mean and modal) varying coefficient functions in each case. The red curves represent the corresponding estimates for $sin(\cdot)$ functions, while the green curves denote the analogous estimates for $exp(\cdot)$ functions with 200 replications.

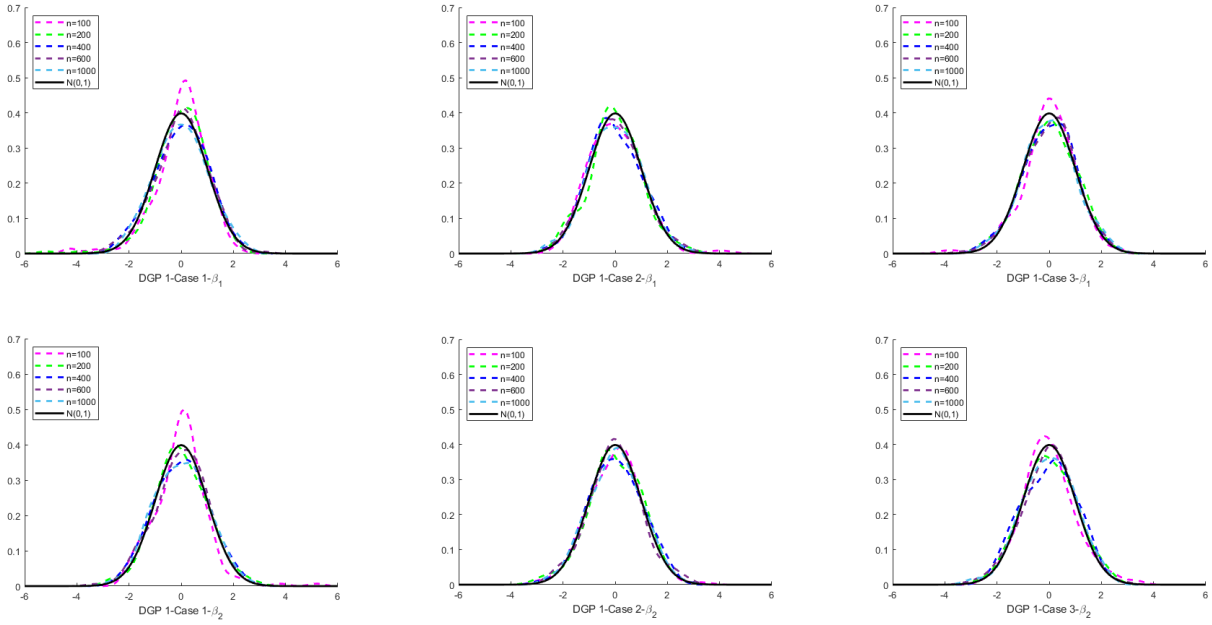

Figure 3: Empirical Density of the Standardized Estimate

Figure 3 shows that the sample distributions have a similar bell shape to the standard normal distribution. In accordance with the asymptotic property, the performance of the asymptotic normality approximation improves as the same size n increases. Note that there appears to pre-

sent some discrepancy between the sample distribution and the standard normal distribution when the sample size n is small, which may explain some of the facts that the convergence rate of modal regression is usually slow.

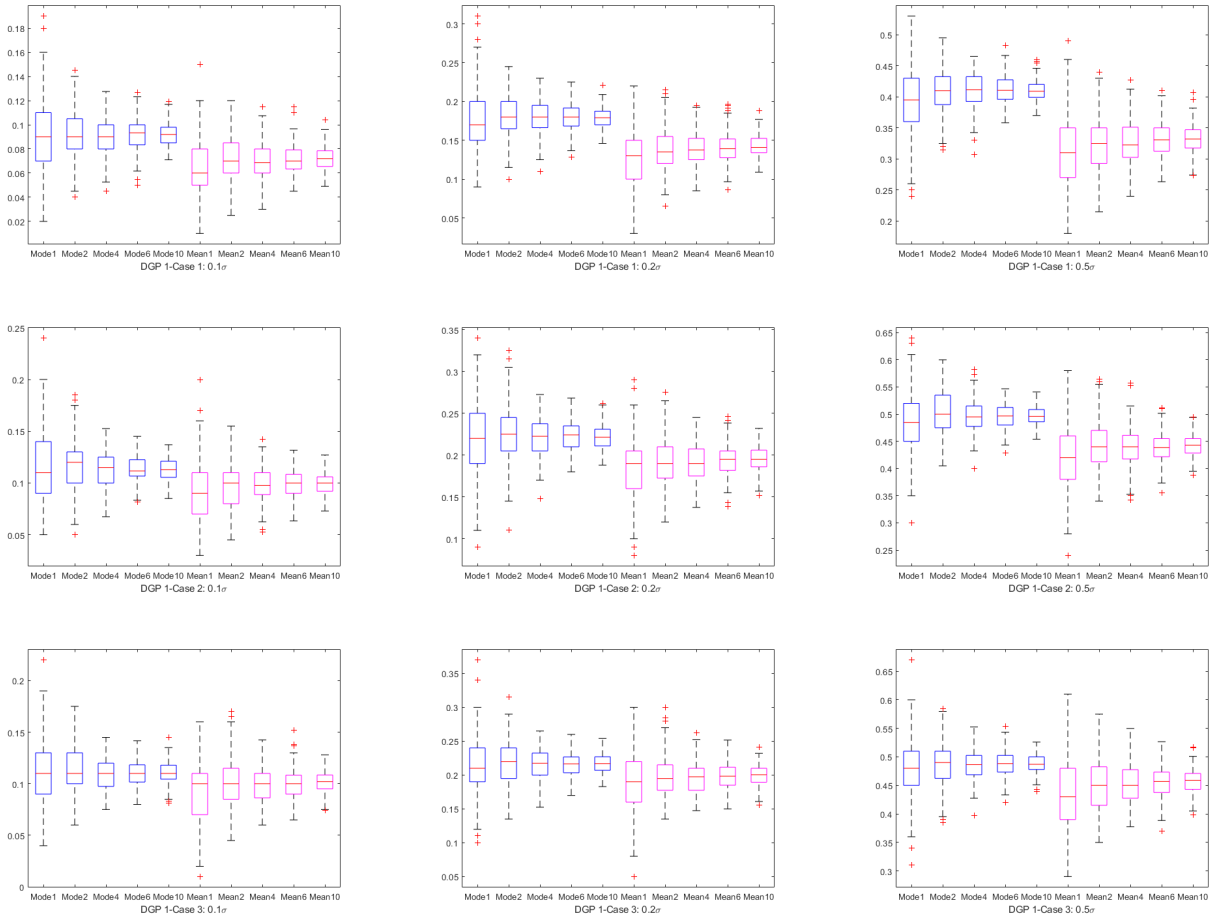

Figure 4: Boxplot of Average of Coverage Probability

Note: For each plot, the numbers 1, 2, 4, 6, and 10 represent the values of n=100, 200, 400, 600, and 1000, respectively.

The research focus in many empirical applications is more on prediction. As stated in Section 1, one of the main advantages of modal regression is having better prediction performance compared to existing regressions (Figure 1). Following Ullah et al. (2021, 2022), we report the average of the coverage probabilities when conducting predictions according to the same length of small intervals centered around each estimate. We consider 0.1σ, 0.2σ, and 0.5σ length of intervals, separately, where $\sigma \approx 2$ for $\epsilon_i \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$. We follow the same DGP process as the above three cases with the sample size $2n$, where we estimate the model with the first n data points and make out-of-sample predictions for the remaining n data points with 200 replications. The results are shown in Figure 4, which demonstrates that modal regression can obtain higher coverage probabilities than mean regression. With the increase of inter-

val length, the coverage probabilities for both modal and mean regressions are increasing and moving closer to each other as expected. The results are consistent with those reached in Yao and Li (2014) and Ullah et al. (2021, 2022).

(2) (Penalized SPLVC Modal Regression) We conduct a simulation experiment to illustrate the finite sample performance of the proposed estimator with variable selection in this part. We first generate random samples from the following model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i)\epsilon_i,$$

where $\boldsymbol{X}$ is composed of two covariates and $\boldsymbol{Z}$ is made up of ten covariates. The covariate vector $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i)^T$ is normally distributed with mean 0, variance $I_{12\times 12}$, and correlation $0.2^{|r-j|}$, where $r, j = 1, \cdots, 12$. For the purpose of selection, we set $\beta_1 = 2$, $\beta_2 = 1$, $\beta_3 = 1$, and $\beta_l = 0$ for $l = 4, \cdots, 10$, which indicates that only the first three variables are relevant and the rest are irrelevant. Other model settings are identical to those in DGP 1. Then, different modal and mean equations are attained as follows.

$$\text{Case 1}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} + Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 1) + X_{2i}sin(2\pi U_i) + 3Z_{1i} \\ \qquad\qquad +Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li} \end{cases}$$

$$\text{Case 2}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} + Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} \\ \qquad\qquad +Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li} \end{cases}$$

$$\text{Case 3}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + 2Z_{1i} + Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + 3Z_{1i} \\ \qquad\qquad +Z_{2i} + Z_{3i} + \sum_{l=4}^{10} 0Z_{li} \end{cases}$$

We report the average number of zero coefficients that are correctly estimated to be zero (denoted by C) and the average number of nonzero coefficients that are incorrectly estimated to be zero (indicated by IC). To present a more comprehensive picture, Table 2 also depicts other criteria for evaluating the performance of the developed model, including U-Fitted (underfitted) —the proportion of ignoring at least one of the nonzero coefficients in all replications, C-Fitted (correctly fitted)—the proportion of selecting all coefficients correctly in all replications, and O-

Table 2: The Results of Variable Selection for SPLVC Regressions

Method	Case	n	C	IC	C-Fitted	U-Fitted	O-Fitted	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	$GMSE(\boldsymbol{\beta})$-S	$RASE(\boldsymbol{\alpha}(U_i)$-S	$GMSE(\boldsymbol{\beta})$-O	$RASE(\boldsymbol{\alpha}(U_i)$-O
Mode	Case 1	100	6.9055	0.6617	0.4577	0.5174	0.0249	2.7728	0.7619	0.8193	0.0044	0.0057	0.0020	-0.0079	0.0022	0.0041	0.0000	0.9597	1.4951	0.4470	1.0516
		200	6.9552	0.4229	0.6467	0.3433	0.0100	2.8252	0.7950	0.9051	0.0000	0.0000	0.0000	0.0000	0.0000	0.0017	-0.0023	0.5190	0.4890	0.1728	0.3345
		400	6.9652	0.3184	0.7164	0.2836	0.0000	2.8510	0.8761	0.9179	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.3165	0.2295	0.0348	0.1267
		600	6.9652	0.2239	0.7960	0.2040	0.0000	2.8505	0.9346	0.9241	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2210	0.1876	0.0190	0.0863
		1000	6.9652	0.0746	0.9303	0.0697	0.0000	2.8634	0.9505	0.9921	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0791	0.1503	0.0106	0.0522
	Case 2	100	6.9502	0.2488	0.7463	0.2388	0.0149	2.0802	0.9305	0.9347	0.0000	-0.0018	0.0000	0.0000	0.0005	0.0000	0.0000	0.2334	0.7432	0.0377	0.7323
		200	6.9652	0.0945	0.9055	0.0945	0.0000	2.0615	0.9460	1.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0716	0.3025	0.0102	0.2279
		400	6.9652	0.0398	0.9602	0.0398	0.0000	2.0432	0.9607	1.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0289	0.1793	0.0044	0.0925
		600	6.9652	0.0249	0.9751	0.0249	0.0000	2.0158	0.9776	0.9987	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0142	0.0976	0.0028	0.0584
		1000	6.9652	0.0050	0.9950	0.0050	0.0000	1.9978	0.9884	0.9999	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0047	0.0715	0.0018	0.0413
	Case 3	100	6.9552	0.4478	0.6218	0.3682	0.0100	2.7093	0.8532	0.8596	0.0000	0.0000	0.0000	0.0000	0.0018	0.0000	0.0018	0.6783	0.4384	0.2321	0.3858
		200	6.9652	0.2637	0.7463	0.2537	0.0000	2.7880	0.9383	0.9042	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2971	0.1382	0.0996	0.1037
		400	6.9652	0.1294	0.8756	0.1244	0.0000	2.8425	0.9296	0.9877	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1322	0.0897	0.0313	0.0352
		600	6.9652	0.0697	0.9353	0.0647	0.0000	2.8446	0.9592	0.9814	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0849	0.0687	0.0124	0.0228
		1000	6.9652	0.0348	0.9652	0.0348	0.0000	2.8523	0.9703	1.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0457	0.0424	0.0087	0.0144
Mean	Case 1	100	4.9055	0.6617	0.4577	0.5174	0.0249	2.0069	0.9969	0.9986	-0.0158	0.0067	-0.0141	-0.0023	0.0031	0.0223	-0.0357	1.0592	1.5179	0.5981	1.1377
		200	4.9403	0.0149	0.4577	0.0149	0.5274	1.9969	0.9756	1.0304	0.0074	-0.0125	0.0252	0.0032	-0.0135	0.0010	0.0029	0.4717	0.7912	0.2061	0.7471
		400	5.8657	0.0050	0.4577	0.0050	0.5373	1.9729	0.9884	1.0166	0.0032	-0.0019	0.0027	-0.0005	-0.0041	-0.0085	-0.0012	0.1790	0.3935	0.1070	0.3755
		600	6.3333	0.0000	0.5522	0.0000	0.4478	1.9559	1.0076	0.9799	-0.0066	0.0008	-0.0033	-0.0029	-0.0015	-0.0011	0.0019	0.0903	0.2639	0.0763	0.2566
		1000	6.7463	0.0000	0.8060	0.0000	0.1940	1.9954	0.9940	0.9935	-0.0001	-0.0007	-0.0015	0.0005	-0.0014	0.0006	0.0000	0.0418	0.1657	0.0457	0.1653
	Case 2	100	5.9353	0.0597	0.4627	0.0597	0.4776	1.9684	0.9896	0.9819	0.0199	-0.0092	-0.0151	-0.0041	0.0105	-0.0154	0.0018	0.3134	0.7873	0.1430	0.7621
		200	6.6119	0.0050	0.7761	0.0050	0.2189	1.9901	0.9945	1.0102	0.0015	0.0001	-0.0043	-0.0047	-0.0024	0.0069	-0.0046	0.1041	0.3861	0.0772	0.3792
		400	6.9055	0.0000	0.9453	0.0000	0.0547	1.9812	0.9844	1.0013	-0.0019	-0.0026	-0.0005	-0.0005	0.0000	0.0012	-0.0002	0.0406	0.2112	0.0364	0.2111
		600	6.9453	0.0000	0.9801	0.0000	0.0199	1.9708	0.9981	0.9977	0.0000	0.0011	0.0000	0.0000	0.0000	-0.0018	0.0003	0.0246	0.1580	0.0234	0.1473
		1000	6.9652	0.0000	0.9950	0.0000	0.0050	1.9809	0.9915	1.0020	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0143	0.1028	0.0138	0.1002
	Case 3	100	5.8607	0.0299	0.4577	0.0299	0.5124	1.9784	0.9671	1.0143	0.0189	-0.0041	-0.0081	0.0043	0.0260	0.0094	-0.0020	0.4256	0.6197	0.2332	0.5966
		200	6.4776	0.0050	0.7313	0.0050	0.2637	2.0271	0.9852	0.9963	0.0039	0.0082	-0.0068	-0.0033	-0.0020	-0.0008	0.0007	0.1522	0.3084	0.1097	0.1931
		400	6.8905	0.0000	0.9303	0.0000	0.0697	1.9906	0.9902	0.9908	-0.0013	-0.0003	0.0012	-0.0004	-0.0006	0.0001	0.0000	0.0650	0.1609	0.0608	0.1475
		600	6.9502	0.0000	0.9851	0.0000	0.0149	1.9808	0.9900	0.9867	0.0000	0.0000	0.0000	0.0005	0.0000	0.0003	0.0000	0.0356	0.1127	0.0338	0.1016
		1000	6.9652	0.0000	0.9950	0.0000	0.0050	1.9915	0.9928	1.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0217	0.0743	0.0214	0.0712

Fitted (overfitted)—the proportion of correctly selecting all nonzero coefficients but including at least one zero coefficients in all replications. As seen from Table 2, the proposed modal regression variable selection procedure performs fairly well in terms of all evaluation criteria. It can select the true irrelevant variables with a high probability, and the percentage of incorrect selections steadily decreases as sample size increases. Compared to mean regression, modal regression could have better performance of oracle procedure in terms of accurate variable selections (higher rates of C and C-Fitted), and offers a more informative summary of the data. Especially, when n is not very large, mean regression cannot eliminate certain irrelevant variables.

To assess the accuracy of the resultant estimators, we use $GMSE$ and $RASE$ to compare the performance of different estimates. The results are shown in Table 2, where S represents the estimates with SCAD variable selection and O denotes the oracle estimates assuming the zero coefficients are known but the other coefficients are unknown. Note that O is only available in simulation studies and serves as a benchmark here for comparisons. The results show that the developed variable selection procedure can estimate all nonzero and zero coefficients more accurately compared to the mean selection method. There is an evident tendency that with the increase of sample size, the SCAD estimation procedure significantly improves the estimation accuracy, indicating the consistency of the suggested variable selection procedure. As expected, the S estimator performs comparably to the oracle procedure in terms of model error and model complexity as the sample size n increases. Furthermore, we find that modal regression performs better than mean regression in terms of $RASE$-Ss, but slightly worse in terms of $GMSE$-Ss. For large sample size, both mean and modal regression variable selection procedures perform reasonably well.

(3) (Varying Coefficient Test) To examine the finite sample performance of the test statistic, we generate random samples from the following model

$$Y_i = X_i\alpha(U_i) + Z_i\beta + \sigma(X_i, Z_i)\epsilon_i,$$

where $\beta = 3$, $U_i \sim U[0,1]$, $\epsilon_i \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$, and $(X_i, Z_i)^T$ is normally distributed with mean 0, variance $I_{2\times 2}$, and correlation 0.2. To test whether $\alpha(U_i)$ deviates from a constant, i.e., $\mathcal{H}_0 : \alpha(U_i) = \alpha \; vs. \; \mathcal{H}_1 : \alpha(U_i) \neq \alpha$, we set $\alpha(U_i) = \alpha + \gamma(cos(2U_i - 1) - \alpha)$, in which $\alpha = \int_0^1 cos(2u-1)du = 0.8415$. The parameter γ is chosen from the set $\{0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 4.0\}$, which can determine the extent that $\alpha(U_i)$ varies with U_i and evaluate the power of the proposed test. Accordingly, $\gamma = 0$ corresponds to the model under the null hypothesis, which examines the validity of the bootstrap procedure for approximating the null distribution of the test statistic T_0. When γ increases, the alternative moves farther away from the null hypothesis, where one would expect the rejection rates of the null hypothesis to get higher. To obtain the sizes and powers of the suggested test, the simulation replication is taken as 200 and the bootstrap replication B is set to 200. Due to the expensive computation, the sample

size is set to 100 or 200, separately. We define $\sigma(X_i, Z_i) = Z_i$ throughout this simulation.

To check whether the resultant statistic $r_K T_0$ asymptotically follows $\chi^2(d)$ with $d = r_K \mu_n = 2\mu_n^2/\sigma_n^2$, where μ_n and σ_n^2 are the simulated mean and variance of T_0, respectively, we plot the sampling distribution of 200 simulation statistics of T_0 against the $\chi^2(d)$ distribution in Figure 5. The two plots show that the empirical distribution of the developed statistic and the $\chi^2(d)$ are close to one another, demonstrating that the χ^2-distribution can satisfactorily approximate the null distribution of the proposed statistic.

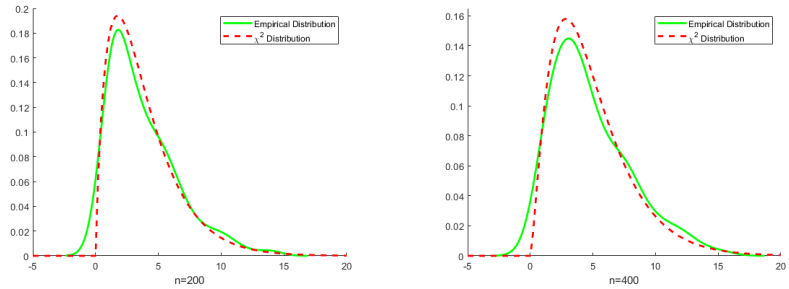

Figure 5: Empirical Distribution of the Proposed Statistic under $\mathcal{H}_0$

Figure 6 displays the relative frequencies of rejecting $\mathcal{H}_0$ at the significance levels $\alpha = 0.01$, 0.05, and 0.1, respectively. It shows that the proposed test statistic performs satisfactorily in terms of both size and power. The power, as expected, is a monotone increasing function of γ. Under the null hypothesis, i.e., $\gamma = 0$, the estimated sizes of the suggested test are closer to the nominal significance levels 1%, 5%, and 10%, indicating that the developed test can provide the appropriate levels of testing under these three different significance levels. Under the alternative hypothesis, i.e., $\gamma > 0$, when the sample size is 100, the powerful function does not increase rapidly to 1 as γ deviates from 0, and achieves 90% when $\gamma > 3$. However, when the sample size rises to 200, the power performance becomes better, achieving close to 1 with $\gamma > 1.5$. These findings suggest that the bootstrap estimate of the null distribution of the test statistic is approximately valid, and the developed test with the residual-based bootstrap is practically useful when we have a moderately large dataset.

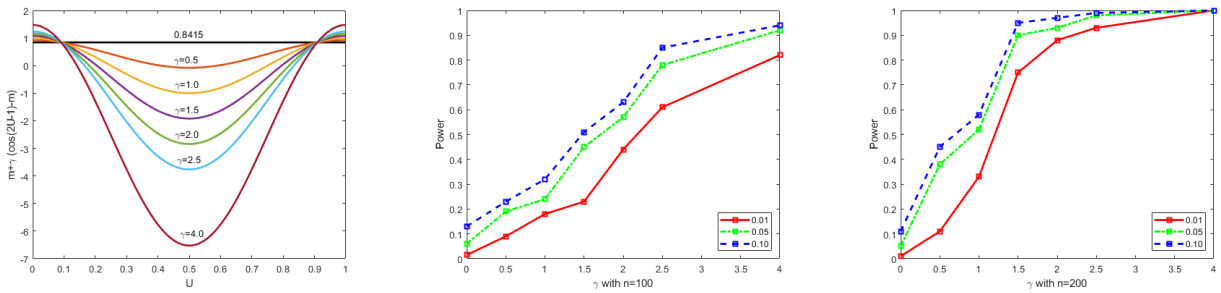

Figure 6: Type I Error and Power of Bootstrap Test

Note: The left plot represents the difference between the null ($\gamma = 0$) and the alternatives.

4.2 Empirical Analyses

Example 1: Application to Return to Education We demonstrate the effectiveness of the proposed estimating method and test procedure by an application to the return to education dataset. It is well-known that there is a nonlinear relationship between wage and educational level or work experience. A substantial amount of literature has been devoting effort to investigating the empirical relationship between earnings and education. For example, Su et al. (2013) introduced a local linear GMM estimation of varying coefficient instrumental variables model with an application to estimating the rate of return to schooling; Cai et al. (2006) utilized a two-step nonparametric procedure to estimate the return to education; among others. Note, however, that all of these studies are based on mean regression. To provide more empirical evidence to support the importance of education, we investigate the mode relationship between earnings and education using the developed SPLVC modal regression, where we utilize random samples from the 1985 wave of the Australian Longitudinal Survey (ALS). In the empirical setting, we choose $log(wage)$ to be the dependent variable Y and work experience as the index variable U. We use years of education as the variable X and the other four categorical variables as the control variables, namely, indicators for marital status (Z_1), union membership (Z_2), government employment (Z_3), and whether a person was born in Australia (Z_4). The model is defined as follows[10]

$$Y = \alpha_1(U) + \alpha_2(U)X + Z_1\beta_1 + Z_2\beta_2 + Z_3\beta_3 + Z_4\beta_4 + \epsilon, \tag{4.1}$$

where $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are unknown varying coefficient functions. The resulting sample contains 2041 observations with work experience of less than or equal to 8 years.[11] Table S3 in the supplementary note provides summary statistics for the sample.

Table 3: The Estimation Results of Equation (4.1)

Variable	SPLVC Modal	SPLVC Mean	Variable	SPLVC Modal	SPLVC Mean
Born in Australia	-0.0806	-0.0912	Government Employee	0.1422	0.1366
	(0.0017)	(0.0207)		(0.0016)	(0.0167)
Married	0.1867	0.1831	Union Member	0.0132	0.0098
	(0.0028)	(0.0190)		(0.0007)	(0.0151)
MAPE	14.55	15.19			

We utilize both the SPLVC mean and modal regressions to estimate (4.1). To evaluate the ability of reproducing data, we compare the in-sample prediction performance of these two

[10]Due to unobservable heterogeneity in schooling choices, education is an endogenous variable in the labour economics literature. We do not consider the endogenous issue in this paper, but it would be interesting to explore such a case in modal regression.

[11]The data from Su et al. (2013) have only eight observations with experience being more than or equal to 9 years. We delete these observations, yielding a total of 2041 observations. In addition, the distribution of $log(wage)$ is nearly symmetric, indicating that modal estimation should be similar to mean estimation.

models by reporting mean absolute percentage error ($MAPE$), defined as $MAPE = (100/n)$ $\sum_{i=1}^{n}|Y_i - \hat{Y}_i|/Y_i$, where $\hat{Y}_i$ is the estimated value and n is sample size. The standard error in parenthesis in Table 3 is obtained using the bootstrap technique with 200 replications (Ullah et al., 2021). As shown in Table 3, all coefficients $\boldsymbol{\beta}$ in modal and mean regressions are statistically significant at the 5% significance level. Modal regression has coefficient signs that are consistent with mean regression but have different magnitudes. In general, individuals who were not born in Australia, were married, worked for the government, and were members of a trade union have higher wages based on mode effect, which is aligned with the mean results obtained by Cai et al. (2006). There are some notable differences in the estimates of the coefficients of variables $\boldsymbol{Z}$ between mean and modal regressions. The modal coefficient of *Born in Australia* is larger than that of mean regression (negative values), indicating that although aliens' average salary is higher than those of natives, the effect based on mean is overestimated. With respect to the variable *Married*, it is generally known that married females are more mature and thus more attractive to employers. However, the effect is underestimated via mean regression. For variable *Government Employee*, modal regression provides a larger estimated effect compared to mean regression, which reveals the fact that people who work in government typically earn higher salaries. The coefficient of *Union member* of modal regression is larger than that of mean regression, indicating that in reality the benefit of female workers joining unions has been underestimated at some points. Moreover, compared to mean regression, modal regression produces a smaller $MAPE$, representing better performance in reproducing data.

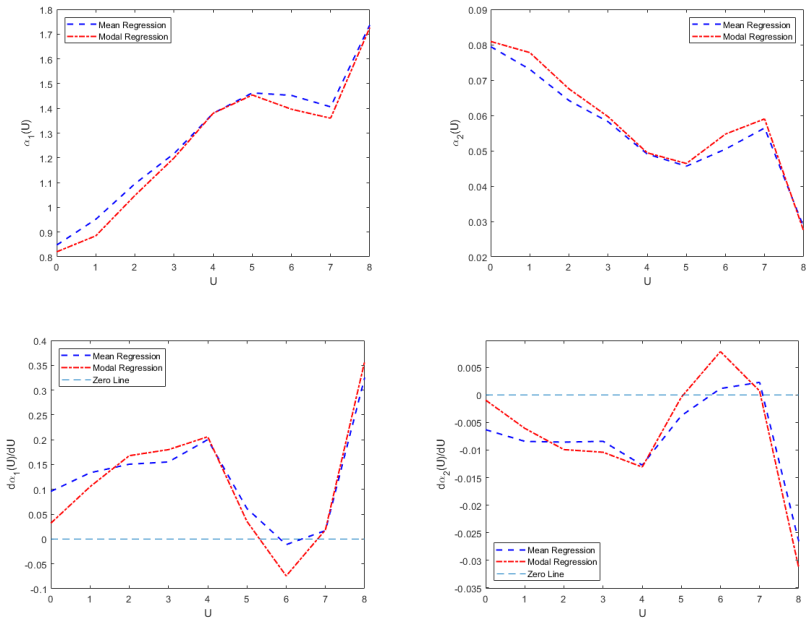

Figure 7: Estimated Curves for Wage Equation

Figure 7 reports the estimated curves for $\alpha_1(u)$, $\alpha_2(u)$, and the associated derivatives. In terms of $\alpha_1(u)$, both modal and mean regressions accurately reflect the nonlinear relationship

between experience and wage, implying that more experienced female workers tend to have higher wages in general. Mean regression, on the other hand, overestimates the effect of experience on wage. For the marginal effect of experience on wage ($\partial\alpha_1(u)/\partial u$), mean regression always indicates a positive relationship, whereas modal regression shows a negative relationship around six years. In regard to $\alpha_2(u)$, we can observe a positive relationship between education and wage based on modal and mean regressions, but the magnitude and shape are different, with mean regression underestimating the effect of education on wage. According to the plot of $\partial\alpha_2(u)/\partial u$, both modal and mean regressions suggest that the marginal effect of education decreases with the increase of experience for either low or high experienced workers. However, around the middle levels of experience (six years), modal regression reveals that there is a positive relationship between experience and the effect of education on wage.

Finally, we apply the proposed varying coefficient test to check whether $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ really vary over experience, i.e., $\mathcal{H}_0 : \alpha_1(U_i) = \alpha_1$ and $\alpha_2(U_i) = \alpha_2$ *vs.* $\mathcal{H}_1 : \alpha_1(U_i) \neq \alpha_1$ or $\alpha_2(U_i) \neq \alpha_2$, where α_1 and α_2 are two unknown constants. The obtained p-value from the bootstrap algorithm is 0.0150 for the considered null hypothesis, which suggests that we should reject the null hypothesis at the 5% significance level.

Example 2: Application to Boston Housing Dataset To illustrate variable selection for modal regression, we analyze the Boston Housing dataset (Fan and Huang, 2005), which contains 506 observations within the Boston Standard Metropolitan Statistical Area in 1970. We primarily employ a partially linear model to investigate the relationship between the median values of owner-occupied homes in the Boston area ($MEDV$) and the following covariates: $CRIM$ (per capita crime rate by town), RM (average number of rooms per dwelling), TAX (full-value property-tax rate per 10,000 USD), NOX (nitric oxides concentration parts per 10 million), $PTRATIO$ (pupil-teacher ratio by town), AGE (proportion of owner-occupied units built prior to 1940), B ($1000(Bk - 0.63)^2$ where Bk is the proportion of blacks in town), and $LSTAT$ (lower status of the population). See summary statistics for the sample in Table S4 in the supplementary note.

Table 4: The Selected Parametric Components

Variable	SPLVC Modal	SPLVC Mean	Variable	SPLVC Modal	SPLVC Mean
CRIM	-1.7616	-0.7459	RM	1.7102	3.6048
TAX	0	0	NOX	0	-2.1708
PTRATIO	0	0	AGE	-0.6307	-0.7808
B	0	0			
MAPE	86.53	129.90			

We choose scaled $U = \sqrt{LSTAT}$ on the interval $[0, 1]$ as the index variable and standardize all X-variables and the response variable to facilitate implementation. As argued by Fan and

Huang (2005), the influences of X-variables on $MEDV$ vary with the level of $LSTST$, thus it may be reasonable to fit a partially linear varying coefficient model. The main objective of this example, however, is to demonstrate the variable selection methodology proposed in this paper. As a result, we instead fit a partially linear model defined as

$$MEDV_i = \alpha(U_i) + \sum_{j=1}^{7} \beta_j X_{ij} + \epsilon_i, \ \ i = 1, \cdots, 506 \tag{4.2}$$

to reveal interesting data structures, where we utilize both mean and modal regressions to simultaneously conduct estimation and variable selection.

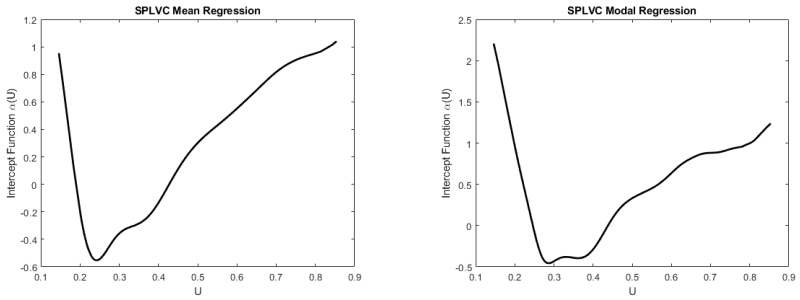

Figure 8: Estimated Curves ($\alpha(U_i)$) for Boston Housing Dataset

Table 4 shows the variable selection results and Figure 8 presents the estimated curves for $\alpha(U_i)$. According to Table 4, both mean and modal regressions can identify three nonzero coefficients, which indicates that the covariates $CRIM$, RM, and AGE have effects on the median value of owner-occupied home based on mean or mode. In contrast to mean regression, besides TAX, $PTRATIO$, and B, modal regression also shrinkages the coefficient of NOX to zero, implying that the covariate NOX has no effect on the median value of owner-occupied home when measured from the "most likely" value (mode). This further suggests that modal regression can provide a simpler model than mean regression and, on occasion, disclose new model characteristics that mean regression cannot reveal. In addition, modal regression has considerably better in-sample prediction performance than mean regression in terms of $MAPE$, demonstrating that modal regression with variable selection could be an attractive technique for simultaneously selecting variables and estimating coefficients.

5 Concluding Remarks

To broaden the scope of existing modal regression models, we in this paper propose a novel SPLVC modal regression and develop a computationally efficient three-stage estimation procedure to estimate the model. The asymptotic properties of the resultant estimators are studied, and the selection of bandwidths for the developed model is discussed. In contrast to condition-

al SPLVC mean or quantile regression, the introduced SPLVC modal regression provides additional information on how the "most likely" values of the dependent variable are affected by the regressors. It will be advantageous to consider the proposed modal regression as a complement to the existing regression tools and to employ it in situations where the distribution of the data is skewed. In addition, we investigate SPLVC modal regression with variable section to eliminate irrelevant variables while simultaneously estimate nonzero coefficients, and we develop a goodness-of-fit testing statistic for hypotheses on coefficient functions by taking a kernel-based function as the loss function instead of the traditional sum of squared errors. The modal variable selection procedure is shown to possess the oracle property subject to regularity assumptions. Monte Carlo simulations and empirical analyses reveal the reasonably good finite sample performance of the newly proposed model. In the supplementary note, we also discuss the extension of the SPLVC modal regression to the case where some varying coefficient functions admit higher-order smoothness.

As far as we are aware, this is the first paper that presents a systematic investigation of the SPLVC modal regression. Given the advantages and superior performance of this model over existing models with skewed datasets, further research into its applicability in other contexts would be worthwhile. For instance, measurement error models, also known as errors-in-variable models in the literature, are frequently encountered in practice when measurements on covariates contain errors. If the measurement error is ignored, the suggested three-stage estimation procedure will lead to biased estimators. As a result, it is worth extending the SPLVC modal regression to the case where covariates are measured with errors, which can be investigated using the deconvolution method. Furthermore, as mentioned in the empirical analysis, we in this paper do not address the endogeneity issue in the SPLVC modal regression. Endogeneity in modal regression will be important and meaningful to broaden the application of modal regression models. Such an endogeneity problem can be solved directly by applying the method of moments or instrumental variable estimation. Finally, due to the complexity of the objective function, the SPLVC modal regression lacks a convenient inference procedure with suitable bandwidth selection methods and reliable estimation algorithms when compared to mean or quantile regression. Although the suggested MEM algorithm can solve models efficiently, it is necessary to develop other algorithms that are less sensitive to initial values or bandwidths. All of these will be researched in more depth in the future.

References

Ahmad, I., Leelahanon, S., and Li, Q. (2005). Efficient Estimation of A Semiparametric Partially Linear Varying Coefficient Model. *The Annals of Statistics*, 33 (1), 258-283.

Cai, Z., Das, M., Xiong, H., and Wu, X. (2006). Functional Coefficient Instrumental Variables Models. *Journal of Econometrics*, 133, 207-241.

Cai, Z., Fan, J., and Yao, Q. (2000). Functional-Coefficient Regression Models for Nonlinear Time Series. *Journal of the American Statistical Association*, 95, 941-956.

Chen, Y. (2018). Modal Regression using Kernel Density Estimation: A Review. *Wiley Interdisciplinary Reviewers: Computational Statistics*, 10:e1431.

Chen, Y., Genovese, C. R., Tibshirani, R. J., and Wasserman, L. (2016). Nonparametric Modal Regression. *The Annals of Statistics*, 44 (2), 489-514.

De Jong, P. (1987). A Central Limit Theorem for Generalized Quadratic Forms. *Probab. Theor. Relat. Fields*, 75, 261-277.

Fan, J. and Gijbels, I. (1996). Local Polynomial Modeling and its Applications. *Chapman and Hall, London*.

Fan, J. and Huang, T. (2005). Profile Likelihood Inferences on Semiparametric Varying Coefficient Partially Linear Models. *Bernoulli*, 11, 1031-1057.

Fan, J. and Jiang, J. (2007). Nonparametric Inference with Generalized Likelihood Ratio Tests. *Test*, 16, 409-444.

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 Zhang, W. (1999). Statistical Estimation in Varying Coefficient Models. *The Annals of Statistics*, 27, 1491-1518.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, 29, 153-193.

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.

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.

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.

Krief, J. M. (2017). Semi-Linear Mode Regression. *Econometrics Journal*, 20, 149-167.

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.

Ota, H., Kato, K., and Hara, S. (2019). Quantile Regression Approach to Conditional Mode Estimation. *Electronic Journal of Statistics*, 13, 3120-3160.

Parzen, M. (1962). On Estimation of a Probability Density Function and Mode. *Philos. Trans. Roy. Soc. London Ser. A*, 186, 343-414.

Robinson, P. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56, 931-954.

Su, L., Murtazashvili, I., and Ullah, A. (2013). Local Linear GMM Estimation of Functional Coefficient IV Models with An Application to Estimating the Rate of Return to Schooling. *Journal of Business & Economic Statistics*, 31 (2), 184-207.

Su, L. and Zhang, Y. (2013). Variable Selection in Nonparametric and Semiparametric Regression Models. *Handbook in Applied Nonparametric and Semi-Nonparametric Econometrics and Statistics. Research Collection School Of Economics.*

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.

Wang, L., Li, R., and Tsai, C. L. (2007). Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, 94 (3), 553-568.

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*.

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.

Zhang, T., Kato, K., and Ruppert, D. (2021). Bootstrap Inference for Quantile-Based Modal Regression. *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2021.1918130.

Zhao, W., Zhang, R., Liu, J., and Lv, Y. (2014). Robust and Efficient Variable Selection for Semiparametric Partially Linear Varying Coefficient Model Based on Modal Regression. *Annals of the Institute of Statistical Mathematics*, 66 (1), 165-191.

Zhou, H. and Huang, X. (2019). Bandwidth Selection for Nonparametric Modal Regression. *Communications in Statistics-Simulation and Computation*, 48 (4), 968-984.

Zhou, Y. and Liang, H. (2009). Statistical Inference for Semiparametric Varying-Coefficient Partially Linear Models with Error-Prone Linear Covariates. *The Annals of Statistics*, 37 (1), 427-458.

Supplement to "Semiparametric Partially Linear Varying Coefficient Modal Regression"

Aman Ullah[a] Tao Wang[b] Weixin Yao[a]

a. University of California, Riverside *b*. University of Victoria

In this supplementary note, we extend the proposed SPLVC modal regression to the case where some coefficient functions admit higher-order smoothness, provide summary statistics for the samples in empirical Examples 1 and 2 (Section 4.2), present additional simulation results for SPLVC modal regression as well as the Monte Carlo experiment for SPLVC multimodal regression, and outline all the proofs for the theorems listed in the paper.

S1 Extension to Higher-Order Smoothness Case

The developed three-stage estimation procedure allows the varying coefficient functions to have different orders of derivatives as long as they are at least two, which is the underlying mechanism for the local linear approximation. It is nevertheless noticed that the suggested estimation procedure implicitly assumes that all varying coefficient functions possess the same minimum degree of smoothness and thus can be approximated equally effectively. When some components of $\boldsymbol{\alpha}(\cdot)$ are known to admit higher degrees of smoothness than others, the proposed three-stage estimation procedure may not be optimal for them (in the sense of optimal convergence rate). Intuitively, a smooth component requires a large bandwidth to decrease variation, whereas a rough component requires a small bandwidth to reduce bias. In this situation, the rate of the bias of all estimated varying coefficient functions will be determined by the rate of the local polynomial with the lowest degree. This implies that all components cannot be optimally evaluated with a single choice of bandwidth concurrently.

Such a problem has been raised explicitly by Fan and Zhang (1999) for investigating varying coefficient mean regression. To deal with this issue in SPLVC modal regression, we provide a two-step estimation procedure by extending the result of Fan and Zhang (1999) and derive the asymptotic properties. We emphasize that although in advance we cannot know the order of smoothness of the varying coefficient functions in practice, the extended estimation method is shown theoretically and numerically to have a significant gain when the considered varying coefficient function is smoother than the rest of the functions, and has the same performance as the introduced three-stage estimation procedure when they have the same minimum order of smoothness. Therefore, the extended two-step estimation procedure can be considered as an improved (and more reliable) version of the developed three-stage estimation method.

In the **first step**, we obtain the modal estimator $\tilde{\boldsymbol{\beta}}$ following the suggested method in Section 2, pretending that all of the components of $\boldsymbol{\alpha}(\cdot)$ possess about the same degrees of smoothness. We then define $\tilde{Y}_i = Y_i - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}$ to alter the original SPLVC modal regression to the following varying coefficient modal regression

$$Mode(\tilde{Y}_i|\boldsymbol{X}_i, U_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) + \underbrace{\boldsymbol{Z}_i^T\boldsymbol{\beta} - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}}_{o_p(1)}. \tag{S.1}$$

To illustrate the necessity of the second-step estimation, we first show that we cannot achieve the optimal estimator even if we utilize the higher order local approximation for some functions in the proposed three-stage estimation procedure. We assume that $\alpha_p(\cdot)$ is smoother than the rest of the varying coefficient functions without loss of generality, i.e.,

$$Mode(\tilde{Y}_i|\boldsymbol{X}_i, U_i) \approx \sum_{j=1}^{p-1}\alpha_j(U_i)X_{ij} + \alpha_p(U_i)X_{ip},\ j = 1, \cdots, p-1. \tag{S.2}$$

With the assumption that $\alpha_p(\cdot)$ has a bounded fourth derivative and others have a second derivative, we can locally approximate $\alpha_p(\cdot)$ by a cubic function

$$\alpha_p(U_i) \approx \alpha_p + b_p\left(U_i - u\right) + c_p\left(U_i - u\right)^2 + d_p\left(U_i - u\right)^3, \tag{S.3}$$

where U_i is in a neighborhood of u. Then, following the third-stage estimation procedure in Section 2, we should maximize the following local kernel-based objective function with respect to α_j, b_j, α_p, b_p, c_p, and d_p for given kernels $\phi(\cdot)$ and $K(\cdot)$

$$\frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\phi\left(\frac{\tilde{Y}_i - m(\{X_{ij}\}_{j=1}^{p-1}, X_{ip})}{\lambda_1}\right)K\left(\frac{U_i - u}{\lambda_2}\right), \tag{S.4}$$

where λ_1 and λ_2 are two bandwidths that depend on sample size n, and

$$m(\{X_{ij}\}_{j=1}^{p-1}, X_{ip}) = \sum_{j=1}^{p-1}\{\alpha_j + b_j\left(U_i - u\right)\}X_{ij} - \left(\alpha_p + b_p\left(U_i - u\right) + c_p\left(U_i - u\right)^2 + d_p\left(U_i - u\right)^3\right)X_{ip}.$$

Define

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{11}\left(U_1 - u\right) & \cdots & X_{1p} & X_{1p}\left(U_1 - u\right) & X_{1p}\left(U_1 - u\right)^2 & X_{1p}\left(U_1 - u\right)^3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n1}\left(U_n - u\right) & \cdots & X_{np} & X_{np}\left(U_n - u\right) & X_{np}\left(U_n - u\right)^2 & X_{np}\left(U_n - u\right)^3 \end{pmatrix}.$$

According to the Taylor expansion, we can re-write

$$m(\{X_{ij}\}_{j=1}^{p-1}, X_{ip}) \approx \mathbf{X}\big(\alpha_1(u), \alpha_1^{(1)}(u), \cdots, \alpha_{p-1}(u), \alpha_{p-1}^{(1)}(u), \alpha_p(u), \alpha_p^{(1)}(u), (1/2)\alpha_p^{(2)}(u),$$

$$(1/6)\alpha_p^{(3)}(u))^T + \tfrac{1}{2}\sum_{j=1}^{p-1}\begin{pmatrix} \alpha_j^{(2)}(u)(U_1-u)^2 X_{1j} \\ \vdots \\ a_j''(u)(U_n-u)^2 X_{nj} \end{pmatrix} + \tfrac{1}{4!}\begin{pmatrix} a_p^{(4)}(u)(U_1-u)^4 X_{1p} \\ \vdots \\ a_p^{(4)}(u)(U_n-u)^4 X_{np} \end{pmatrix},$$

where $\alpha^{(c)}(\cdot)$ denotes the cth derivative of $\alpha(\cdot)$. Following the same procedures for proving Theorem 2.6, we can show that the bias of the estimator of $\alpha_p(U_i)$ is $O_p(\lambda_1^2+\lambda_2^2)$ and the variance is $O_p((n\lambda_2\lambda_1^3)^{-1})$. Thus, the MSE of the estimator is only $O_p(\lambda_1^4+\lambda_2^4+(n\lambda_2\lambda_1^3)^{-1})$, which achieves the rate $O_p(n^{-1/2})$ when the bandwidth $\lambda_1=\lambda_2=O(n^{-1/8})$ is used. The above mathematical illustration demonstrates that the developed three-stage estimator for $\alpha_p(\cdot)$ inherits the non-negligible approximation error and is therefore not optimal.

To achieve the optimal estimator, we in the **second step** make use of the third-stage estimates of $\alpha_1(\cdot),\cdots,\alpha_{p-1}(\cdot)$. Following that, a local kernel-based objective function weighted by a kernel function $K(\cdot)$ is applied to estimate $\alpha_p(\cdot)$, i.e.,

$$\frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\phi\left(\frac{\tilde{Y}_i-\sum_{j=1}^{p-1}\tilde{\alpha}_j(U_i)X_{ij}-m(X_{ip})}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right), \tag{S.5}$$

where $m(X_{ip})=(\alpha_p+b_p(U_i-u)+c_p(U_i-u)^2+d_p(U_i-u)^3)X_{ip}$.

The solution of (S.5) gives the two-step modal estimators. Similar to the suggested SPLVC modal regression, we can utilize a modified MEM algorithm to numerically solve the preceding equation. Provided that the initial bandwidths in the first step are small enough (so that the bias of the first-step estimator is small), we have the following consistency and asymptotic normality results for the estimators, where we show that the two-step modal estimators can achieve the optimal convergence rate and share the same optimality as if $\{\alpha_j(\cdot)\}_{j=1}^{p-1}$ were known.

Theorem S1. *Suppose that the regularity conditions C1-C7 are met (instead of C4, we need $\alpha_j^{(2)}(u)$ to be continuous in a neighborhood of u for $j=1,\cdots,p-1$ and the functional coefficient $\alpha_p(u)$ has a continuous fourth derivative in a neighborhood of u). With probability approaching one, as $n\to\infty$, $\lambda_1\to 0$, $\lambda_2\to 0$, $h_4=o(\lambda_1^2)$, $h_5=o(\lambda_2^2)$, $\lambda_1^4/\lambda_2\to 0$, $h_3/h_5\to 0$, and $n\lambda_1\lambda_2^5\to\infty$, there exist consistent maximizers $(\hat{\alpha}_p(u),\hat{b}_p(u),\hat{c}_p(u),\hat{d}_p(u))$ of (S.5) such that*

i. $|\hat{\alpha}_p(u)-\alpha_{0p}(u)|=O_p\left((n\lambda_2\lambda_1^3)^{-1/2}+\lambda_1^2+\lambda_2^4\right)$,

ii. $|\lambda_2(\hat{b}_p(u)-b_{0p}(u))|=O_p\left((n\lambda_2\lambda_1^3)^{-1/2}+\lambda_1^2+\lambda_2^4\right)$,

iii. $|\lambda_2^2(\hat{c}_p(u)-c_{0p}(u))|=O_p\left((n\lambda_2\lambda_1^3)^{-1/2}+\lambda_1^2+\lambda_2^4\right)$,

iv. $|\lambda_2^3(\hat{d}_p(u)-d_{0p}(u))|=O_p\left((n\lambda_2\lambda_1^3)^{-1/2}+\lambda_1^2+\lambda_2^4\right)$,

where $\alpha_{0p}(u)$, $b_{0p}(u)$, $c_{0p}(u)$, and $d_{0p}(u)$ are the true parameters of (S.2).

Theorem S2. *With $n\lambda_2^5\lambda_1^3 = O(1)$ and $n\lambda_2\lambda_1^7 = O(1)$, under the same conditions as Theorem S1, the estimators satisfying the consistency results in Theorem S1 have the following asymptotic result*

$$\sqrt{n\lambda_2\lambda_1^3}\left[\begin{pmatrix}\hat{\alpha}_p(u)-\alpha_{0p}(u)\\ \lambda_2(\hat{b}_p(u)-b_{0p}(u))\\ \lambda_2^2(\hat{c}_p(u)-c_{0p}(u))\\ \lambda_2^3(\hat{d}_p(u)-d_{0p}(u))\end{pmatrix} - \tilde{\Gamma}(u)^{-1}\left(\frac{\lambda_2^4}{24}\tilde{\Lambda}_2(u)\alpha_{0p}^{(4)}(u) - \frac{\lambda_1^2}{2}\tilde{\Lambda}_1(u)\right)\right]$$
$$\xrightarrow{d} \mathcal{N}\left(0, \frac{\int\tau^2\phi^2(\tau)d\tau}{f_U(u)}\tilde{\Gamma}(u)^{-1}\tilde{\Sigma}(u)\tilde{\Gamma}(u)^{-1}\right).$$

If we allow $n\lambda_2^5\lambda_1^3 \to 0$ and $n\lambda_2\lambda_1^7 \to 0$, the asymptotic theorem becomes

$$\sqrt{n\lambda_2\lambda_1^3}\begin{pmatrix}\hat{\alpha}_p(u)-\alpha_{0p}(u)\\ \lambda_2(\hat{b}_p(u)-b_{0p}(u))\\ \lambda_2^2(\hat{c}_p(u)-c_{0p}(u))\\ \lambda_2^3(\hat{d}_p(u)-d_{0p}(u))\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \frac{\int\tau^2\phi^2(\tau)d\tau}{f_U(u)}\tilde{\Gamma}(u)^{-1}\tilde{\Sigma}(u)\tilde{\Gamma}(u)^{-1}\right),$$

$$\textit{where } \tilde{\Sigma} = \mathbb{E}\left[\begin{pmatrix} v_0XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0 & v_2XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0\\ 0 & v_2XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0 & v_4XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})\\ v_2XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0 & v_4XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0\\ 0 & v_4XX^Tf_\epsilon(0|\hat{\boldsymbol{X}}) & 0 & v_6XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})\end{pmatrix}\middle| U=u\right],$$

$$\tilde{\Gamma} = \mathbb{E}\left[\begin{pmatrix} XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0 & \mu_2XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0\\ 0 & \mu_2XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0 & \mu_4XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ \mu_2XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0 & \mu_4XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0\\ 0 & \mu_4XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}}) & 0 & \mu_6XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\end{pmatrix}\middle| U=u\right],$$

$$\tilde{\Lambda}_1 = \mathbb{E}\left[\begin{pmatrix} Xf_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})\\ 0\\ \mu_2Xf_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})\\ 0\end{pmatrix}\middle| U=u\right], \textit{ and } \tilde{\Lambda}_2 = \mathbb{E}\left[\begin{pmatrix} \mu_4XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ 0\\ \mu_6XX^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})\\ 0\end{pmatrix}\middle| U=u\right].$$

The bias of the two-step modal estimators is of $O_p(\lambda_1^2+\lambda_2^4)$, while the asymptotic variance has the same convergence rate as that of the three-stage modal estimators as long as the bandwidth conditions are satisfied. Also, the bias term is not dominated by the first-step estimation with the imposed bandwidth conditions. Therefore, when taking the optimal bandwidths λ_1 and λ_2 of order $n^{-1/15}$,[12] the MSE of the two-step estimator achieves the optimal rate of convergence $O_p(n^{-8/15})$. This indicates that when some varying coefficient functions

[12]The result demonstrates that the smoother the varying coefficient function is, the larger the optimal bandwidth for the estimator is. In precipice, we can follow the procedures described in Subsection 2.3 to select bandwidths λ_1 and λ_2 with MSE-optimal rates.

admit higher degrees of smoothness, the proposed three-stage estimation method in Section 2 will fail to achieve the optimal convergence rate and will transmit the approximation errors of $\alpha_1(\cdot), \cdots, \alpha_{p-1}(\cdot)$ to the bias of estimating $\alpha_p(\cdot)$. Furthermore, it is straightforward to show that the two-step estimators enjoy the same optimal rate of convergence as the ideal ones where $\alpha_1(\cdot), \cdots, \alpha_{p-1}(\cdot)$ are known, which is consistent with the property of the corresponding mean regression estimators. If $\alpha_p(\cdot)$ is assumed to have at least the same degree of smoothness as the rest of the functions, the two-step estimators will have the same performance as the three-stage modal estimators by using a local linear approximation in both steps, as it shares exactly the same asymptotic properties as $\tilde{\alpha}_p(\cdot)$ provided that some bandwidth conditions are met. This suggests that similar to the varying coefficient mean regression models, the extended two-step modal estimation procedure can be considered as an improved version of (and more efficient than) the developed three-stage estimation approach.

(Monte Carlo Experiment) We conduct a Monte Carlo experiment to provide some insight on the performance of the extended two-step estimation method, where the data are generated from the following models

$$\text{Model 1} : Y = sin(6\pi U)X_1 + sin(2\pi U)X_2 + 2Z + Z\epsilon,$$

$$\text{Model 2} : Y = cos(2\pi U)X_1 + sin(2\pi U)X_2 + 2Z + Z\epsilon,$$

in which U is simulated from a uniform distribution on $[0,\ 1]$ and $\epsilon \sim 0.5N(-1, 2.5^2) + 0.5N(1, 0.5^2)$. The covariate vector (X_1, X_2, Z) is normally distributed with mean 0, variance $I_{3\times3}$, and correlation $0.2^{|k-j|}$ in which $k, j = 1, 2, 3$. Thus, the varying coefficient functions in Model 1 admit different degrees of smoothness, whereas the varying coefficient functions in Model 2 possess the same degree of smoothness. We are primarily interested in $\alpha_2(U) = sin(2\pi U)$ for Model 1, which fluctuates less than $sin(6\pi U)$, and $\alpha_1(U) = cos(2\pi U)$ for Model 2. We conduct 200 simulations with sample sizes $n = 200, 400$, and 600, and simply set $\lambda_1 = \lambda_2 = \{0.1, 0.2, 0.3\}$ to study the influence of bandwidths. We calculate $RASE$ to assess the performance of the proposed three-stage estimation and the extended two-step estimation methods.

Table S1: The Performance of Different Estimation Methods

Bandwidth	Method	$n = 200$	$n = 400$	$n = 600$	Method	$n = 200$	$n = 400$	$n = 600$
				Model 1				
$\lambda_1 = \lambda_2 = 0.1$	Two-Step	0.2543	0.1910	0.1631	Three-Stage	0.3241	0.2685	0.2451
$\lambda_1 = \lambda_2 = 0.2$	Two-Step	0.3020	0.2367	0.2123	Three-Stage	0.3721	0.3243	0.3058
$\lambda_1 = \lambda_2 = 0.3$	Two-Step	0.3124	0.2567	0.2253	Three-Stage	0.4451	0.3893	0.3756
				Model 2				
$\lambda_1 = \lambda_2 = 0.1$	Two-Step	0.2526	0.1906	0.1697	Three-Stage	0.2591	0.2058	0.1919
$\lambda_1 = \lambda_2 = 0.2$	Two-Step	0.2999	0.2419	0.2058	Three-Stage	0.3182	0.2698	0.2448
$\lambda_1 = \lambda_2 = 0.3$	Two-Step	0.3261	0.2521	0.2327	Three-Stage	0.3502	0.3086	0.2807

The results in Table S1 show that the improvement of the two-step estimator is quite substantial for a wide range of bandwidths when some varying coefficient functions admit higher degrees of smoothness (results for Model 1), which is consistent with the asymptotic theory that the extended two-step method outperforms the proposed three-stage estimation procedure. The results for Model 2 indicate that the extended two-step estimation method performs nearly as well as the proposed three-stage estimation procedure when varying coefficient functions have the same minimum degrees of smoothness.

S2 SPLVC Multimodal Regression

As mentioned in the paper, the unique global mode assumption can be released without affecting the estimation procedure. The multimodal dataset is common in economics. For example, if we look carefully at the country's income distribution, we can see that there are two modes relating to developing and developed countries, which is consistent with a dichotomous world made up of countries with different incomes. SPLVC modal regression can then be used to capture these two different situations simultaneously. To demonstrate that the proposed estimation method can also be utilized to estimate SPLVC multimodal regression, we conduct a Monte Carlo simulation based on DGP 1. We generate random samples from the following model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i)\epsilon_i, \tag{S.6}$$

where we set the parameters and varying coefficients be $\boldsymbol{\beta} = (1, 2)^T$ and $\boldsymbol{\alpha}(U_i) = (\alpha_1(U_i), \alpha_2(U_i))^T$ in which $\alpha_1(U_i) = exp(2U_i - 1)$ and $\alpha_2(U_i) = sin(2\pi U_i)$. The index variable U_i is simulated from the uniform distribution $U[0, 1]$. The covariate vector $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$ is normally distributed with mean 0, variance $I_{4\times4}$, and correlation $0.2^{|k-j|}$, where $k, j = 1, 2, 3, 4$. For simplicity, we only consider the case where $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = X_{1i} + Z_{1i}$. To create SPLVC multimodal regression, we generate ϵ_i by mixing two normal distributions with equal weights, where one is centered at 0 and the other is centered at 4, and both variances equal 1 (Figure S1).

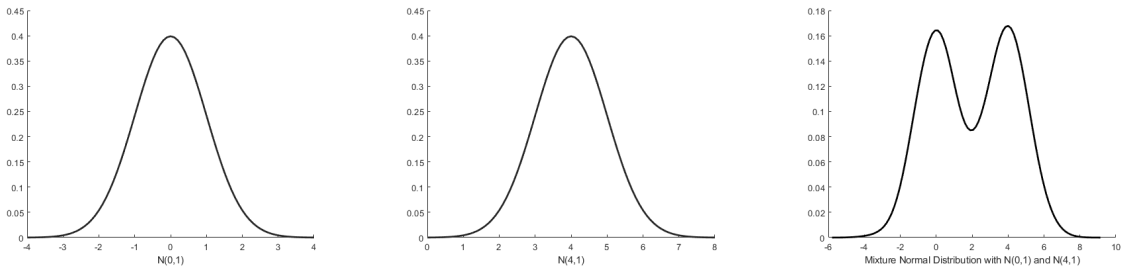

Figure S1: Mixture Normal Distribution with Two Modes

The generalized errors $\{\epsilon_i\}_{i=1}^n$ indicate that $\mathbb{E}(\epsilon_i) = 2$ and $Mode(\epsilon_i) = 0$ or 4. In this case, mean regression may produce misleading results by ignoring data heterogeneity. We then have

the following equations showing two different modal regression lines.

$$\begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 2) + X_{2i}sin(2\pi U_i) + 3Z_{1i} + 2Z_{2i}, \\ \text{Modal Regression Line 1:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}exp(2U_i - 1) + X_{2i}sin(2\pi U_i) + Z_{1i} + 2Z_{2i}, \\ \text{Modal Regression Line 2:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(exp(2U_i - 1) + 4) + X_{2i}sin(2\pi U_i) + 5Z_{1i} + 2Z_{2i}. \end{cases}$$

We consider data sample size $n \in \{200, 400, 600\}$ with 200 replications. Table S2 displays the simulation results, which shows that the proposed estimation method can estimate the SPLVC multimodal regression well with finite samples. With the approximate choice of the initial estimates, we can capture different modal regression lines for data with multiple modes. A set of the varying coefficient estimators is shown in Figure S2 (the black curves are the true varying coefficient functions, while the red and green curves represent the estimates for $exp(\cdot)$ and $sin(\cdot)$, respectively), which clearly indicates that the suggested estimation procedure can capture the true varying coefficients in SPLVC multimodal regression with finite samples, and the fitted performance improves with sample size increasing.

Table S2: The Results of Simulations for SPLVC Multimodal Regression

Method	n	β_1	$MSE(\beta_1)$	β_2	$MSE(\beta_2)$	$GMSE(\boldsymbol{\beta})$	$RASE(\boldsymbol{\alpha}(U_i))$
Mode 1	200	1.0025 (0.0671)	0.0045	2.0041 (0.0554)	0.0031	0.0061	0.5473
	400	0.9952 (0.0498)	0.0025	2.0065 (0.0438)	0.0020	0.0035	0.2564
	600	0.9970 (0.0409)	0.0017	1.9989 (0.0358)	0.0013	0.0034	0.2168
Mode 2	200	4.9770 (0.1374)	0.0193	2.0059 (0.1184)	0.0140	0.0272	0.5662
	400	4.9739 (0.1355)	0.0189	1.9951 (0.1061)	0.0112	0.0265	0.4519
	600	4.9410 (0.1102)	0.0156	1.9997 (0.0813)	0.0066	0.0230	0.4181
Mean	200	2.9585 (0.4820)	0.2329	2.0095 (0.4387)	0.1916	0.3741	1.2427
	400	2.9820 (0.3551)	0.1258	1.9716 (0.3229)	0.1046	0.1850	0.8675
	600	2.9700 (0.2996)	0.0902	2.0097 (0.2540)	0.0643	0.1204	0.7183

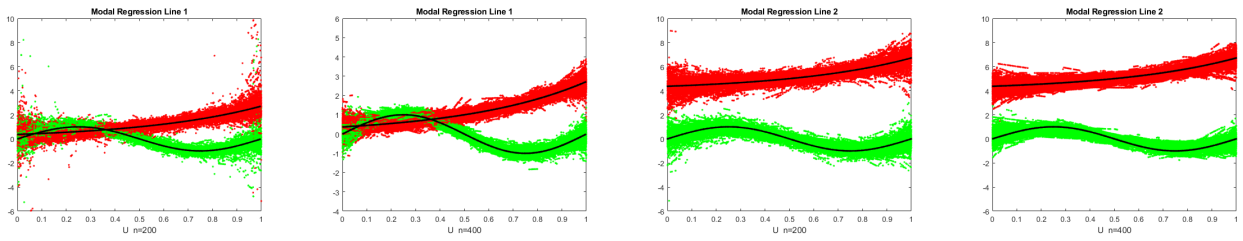

Figure S2: Fitted Varying Coefficient Functions

S3 Summary Statistics

Table S3: The Statistical Characteristics of Sample (Return to Education)

Variable	Mean	Standard Deviation	Min	Max
Born in Australian	0.8618	0.3452	0.0000	1.0000
Married	0.1823	0.3862	0.0000	1.0000
Government Employee	0.2861	0.4521	0.0000	1.0000
Union Member	0.4243	0.4944	0.0000	1.0000
Years of Education	11.7418	1.5277	16.0000	3.0000
Years of Experience	1.4552	1.5277	0.0000	8.0000
Log(Hourly Wage)	0.7950	0.1599	1.6767	-0.4260

Table S4: The Statistical Characteristics of Sample (Boston Housing Dataset)

Variable	Mean	Standard Deviation	Min	Max
CRIM	3.6135	8.6015	0.0063	88.9762
RM	6.2846	0.7026	3.5610	8.7800
TAX	408.2372	168.5371	187	711
NOX	0.5547	0.1159	0.3850	0.8710
PTRATIO	18.4555	2.1649	12.6000	22
AGE	68.5749	28.1489	2.9000	100
B	356.6740	91.2949	0.3200	396.9000
LSTAT	12.6531	7.1411	1.7300	37.9700
MEDV	22.5328	9.1971	5	50

S4 Monte Carlo Experiment (DGP 2)

To further illustrate the applicability of the proposed SPLVC modal regression, we generate random samples from the following DGP with different levels of skewness of density

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i)\epsilon_i, \tag{S.7}$$

where $\alpha_1(U_i) = 8U_i(1-U_i)$, $\alpha_2(U_i) = 2sin(2\pi U_i)$, $\boldsymbol{\beta} = (1, 1, 0.5)^T$, and $\epsilon_i \sim 0.5Ga(k_1, \theta) + 0.5Ga(k_2, \theta)$ in which Ga represents the Gamma distribution, $k_s \in \mathbb{N}_{>0}, s = 1$ or 2, is the shape parameter that can adjust the skewness of v_{it} (coefficient of skewness$= \sqrt{4/k}$), and $\theta \in \mathbb{N}_{>0}$ is the scale parameter (Ullah et al., 2021). Note that $\mathbb{E}(\epsilon_i) = 0.5(k_1 + k_2)\theta$ and $Mode(\epsilon_i) = 0.5$

$(k_1+k_2-1)\theta$. To gain an idea of the effect of the different skewness on estimations, we employ two different schemes to generate the distributions of ϵ_i, where we set $k_1 = 1$ or 7, $k_2 = 2$, and $\theta = 0.5$. The index variable U_i is simulated from the uniform distribution $U[0,1]$. The covariate vector $(\boldsymbol{X}_i^T, \boldsymbol{Z}_i^T)^T$ is normally distributed with mean 0, variance $I_{5\times5}$, and correlation $0.2^{|r-j|}$, where $r, j = 1, \cdots, 5$. To compare with mean regression, we consider three cases, where in case 1 we let $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = X_{1i} + Z_{1i}$, in case 2 we define $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = X_{1i}$, and in case 3 we allow $\sigma(\boldsymbol{X}_i, \boldsymbol{Z}_i) = Z_{1i}$. We then have the following equations.

More Skewed $k_1 = 1$

$$\text{Case 1}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+0.75) + X_{2i}(2sin(2\pi U_i)) + 1.75Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+0.5) + X_{2i}(2sin(2\pi U_i)) + 1.5Z_{1i} + Z_{2i} + 0.5Z_{3i}; \end{cases}$$

$$\text{Case 2}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+0.75) + X_{2i}(2sin(2\pi U_i)) + Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+0.5) + X_{2i}(2sin(2\pi U_i)) + Z_{1i} + Z_{2i} + 0.5Z_{3i}; \end{cases}$$

$$\text{Case 3}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)) + X_{2i}(2sin(2\pi U_i)) + 1.75Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)) + X_{2i}(2sin(2\pi U_i)) + 1.5Z_{1i} + Z_{2i} + 0.5Z_{3i}. \end{cases}$$

Less Skewed $k_1 = 7$

$$\text{Case 1}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+2.25) + X_{2i}(2sin(2\pi U_i)) + 3.25Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+2) + X_{2i}(2sin(2\pi U_i)) + 3Z_{1i} + Z_{2i} + 0.5Z_{3i}; \end{cases}$$

$$\text{Case 2}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+2.25) + X_{2i}(2sin(2\pi U_i)) + Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)+2) + X_{2i}(2sin(2\pi U_i)) + Z_{1i} + Z_{2i} + 0.5Z_{3i}; \end{cases}$$

$$\text{Case 3}: \begin{cases} \text{Mean Regression:} \\ \quad \mathbb{E}(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)) + X_{2i}(2sin(2\pi U_i)) + 3.25Z_{1i} + Z_{2i} + 0.5Z_{3i}, \\ \text{Modal Regression:} \\ \quad Mode(Y_i|\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = X_{1i}(8U_i(1-U_i)) + X_{2i}(2sin(2\pi U_i)) + 3Z_{1i} + Z_{2i} + 0.5Z_{3i}. \end{cases}$$

The estimation results of more skewed and less skewed settings are shown in Tables S5-S6, respectively, containing the estimates and their standard errors (in parentheses), the $MSEs$,

the $GMSEs$, and the $RASEs$. The results for both of these two settings indicate that modal estimators behave well in finite sample situations. For the more skewed case, same as the results of DGP 1, we can observe that the modal and mean estimators have comparable bias while the modal estimators have smaller $GMSEs$ and $RASEs$, indicating some finite sample efficiency gains of the modal estimators in this example. For the less skewed case, the finite sample performance of the modal estimators is better than the corresponding mean estimators in terms of $MSEs$, though the efficiency gain is not very large compared to the more skewed case. However, in the case of less skewed error, the modal estimators are less accurate than the mean estimators in terms of $GMSEs$ and $RASEs$. Figure S3 depicts a set of varying coefficient estimators with excellent fitting performances.

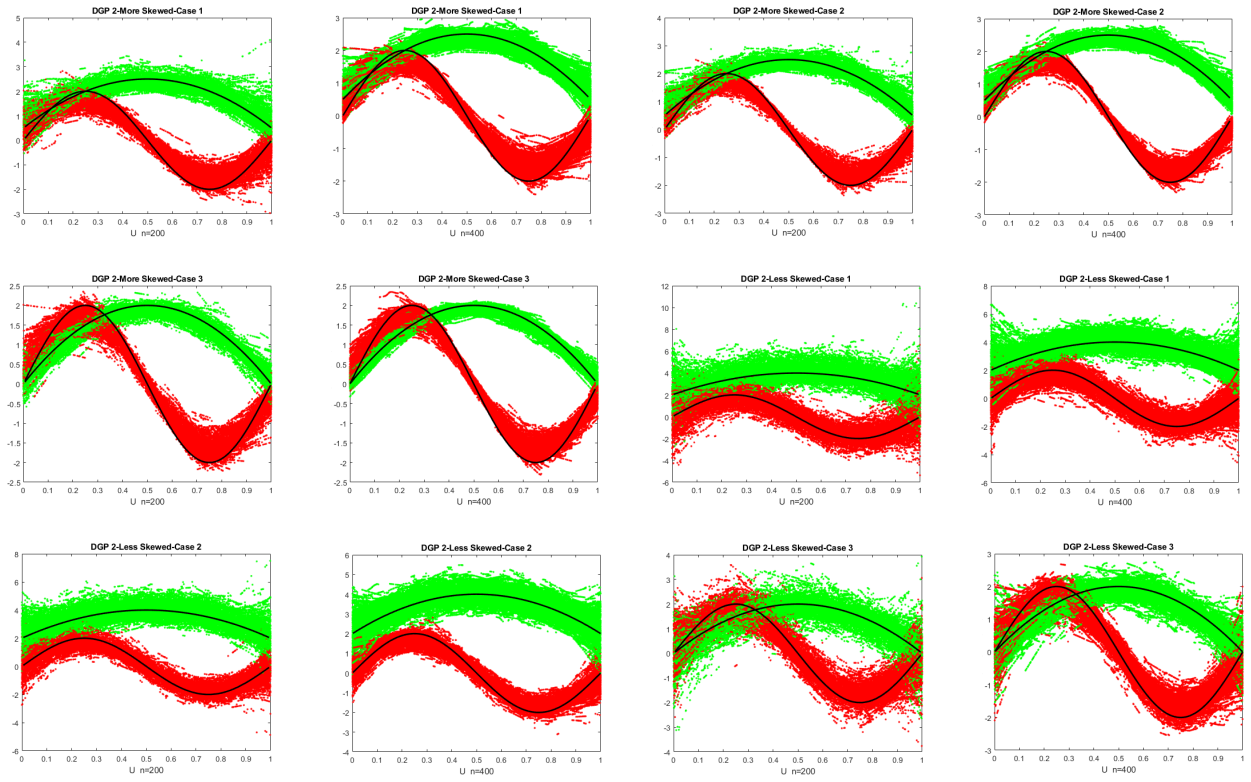

Figure S3: Fitted Varying Coefficient Functions with n=200 or 400 (DGP 2)

Note: The meanings of the different lines in the figure are the same as in DGP 1.

Similar to DGP 1, we report the shape of the empirical density of the standardized parameter estimate. Figure S4 indicates that the asymptotic results provide reasonable approximations in finite samples, and the whole distribution converges to the standard normal as sample size n increases. To evaluate the predictive capabilities of the modal regressions, different from DGP 1, we report the in-sample prediction performance with 0.1σ, 0.2σ, and 0.5σ length of intervals to see how effective the estimation procedure is in reproducing data. Figure S5 shows that the modal regression estimator has better coverage probabilities than the mean regression estimator even for in-sample prediction.

Table S5: The Results of Simulations for DGP 2: More Skewed

Method	Case	n	β_1	$MSE(\beta_1)$	β_2	$MSE(\beta_2)$	β_3	$MSE(\beta_3)$	$GMSE(\boldsymbol{\beta})$	$RASE(\boldsymbol{\alpha}(U_i))$
Mode	Case 1 $\beta_{1,mode}=1.5$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	1.6243 (0.1540)	0.0390	1.0042 (0.1460)	0.0212	0.5017 (0.1420)	0.0201	0.0665	0.6664
		200	1.5737 (0.1146)	0.0185	0.9964 (0.0749)	0.0056	0.5069 (0.0764)	0.0059	0.0241	0.4961
		400	1.5267 (0.0813)	0.0073	1.0040 (0.0449)	0.0020	0.4969 (0.0497)	0.0025	0.0095	0.4142
		600	1.5019 (0.0676)	0.0045	0.9961 (0.0370)	0.0014	0.4989 (0.0359)	0.0013	0.0059	0.3825
		1000	1.5070 (0.0528)	0.0028	0.9966 (0.0255)	0.00066	0.5071 (0.0265)	0.00075	0.0053	0.3553
	Case 2 $\beta_{1,mode}=1$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	0.9965 (0.0799)	0.0064	1.0087 (0.0826)	0.0069	0.4975 (0.0785)	0.0061	0.0143	0.4948
		200	1.0021 (0.0481)	0.0023	1.0030 (0.0461)	0.0021	0.5003 (0.0459)	0.0021	0.0045	0.3975
		400	1.0001 (0.0293)	0.00085	1.0038 (0.0296)	0.00089	0.5001 (0.0293)	0.00086	0.0016	0.3294
		600	1.0003 (0.0231)	0.00053	0.9994 (0.0240)	0.00058	0.5002 (0.0224)	0.0005	0.0011	0.3028
		1000	1.0012 (0.0157)	0.00025	0.9981 (0.0155)	0.00024	0.5030 (0.0157)	0.00025	0.0005	0.2749
	Case 3 $\beta_{1,mode}=1.5$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	1.5869 (0.1175)	0.0213	1.0037 (0.0784)	0.0061	0.4975 (0.0845)	0.0071	0.0288	0.4739
		200	1.5363 (0.0934)	0.0100	1.0003 (0.0474)	0.0022	0.5062 (0.0531)	0.0028	0.0139	0.3760
		400	1.4913 (0.0633)	0.0041	1.0030 (0.0309)	0.00096	0.4993 (0.0303)	0.00092	0.0048	0.3101
		600	1.4789)0.0538)	0.0033	0.9979 (0.0247)	0.00061	0.4976 (0.0225)	0.00051	0.0041	0.2830
		1000	1.5048 (0.0406)	0.0017	0.9981 (0.0174)	0.00030	0.5035 (0.0167)	0.00026	0.0018	0.2532
Mean	Case 1 $\beta_{1,mean}=1.75$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	1.7541 (0.2179)	0.0472	1.0184 (0.2072)	0.0431	0.4957 (0.2045)	0.0416	0.0996	0.6828
		200	1.7616 (0.1660)	0.0276	0.9862 (0.1374)	0.0190	0.4949 (0.1379)	0.0190	0.0492	0.5078
		400	1.7538 (0.0951)	0.0090	1.0081 (0.0899)	0.0081	0.4928 (0.0934)	0.0087	0.0184	0.4266
		600	1.7481 (0.0907)	0.0082	1.0027 (0.0814)	0.0066	0.4962 (0.0711)	0.0050	0.0144	0.3996
		1000	1.7551 (0.0706)	0.0050	1.0041 (0.0551)	0.0030	0.4988 (0.0640)	0.0041	0.0096	0.3744
	Case 2 $\beta_{1,mean}=1$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	1.0070 (0.1142)	0.0130	1.0102 (0.1235)	0.0153	0.4925 (0.1186)	0.0141	0.0312	0.5101
		200	1.0038 (0.0890)	0.0079	0.9876 (0.0759)	0.0059	0.4950 (0.0762)	0.0058	0.0142	0.4149
		400	1.0023 (0.0486)	0.0024	1.0019 (0.0531)	0.0028	0.4981 (0.0520)	0.0027	0.0053	0.3776
		600	0.9994 (0.0456)	0.0021	1.0009 (0.0467)	0.0022	0.4993 (0.0421)	0.0018	0.0042	0.3652
		1000	1.0038 (0.0371)	0.0014	1.0010 (0.0327)	0.0011	0.4988 (0.0361)	0.0013	0.0029	0.3520
	Case 3 $\beta_{1,mean}=1.75$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	1.7511 (0.1435)	0.0205	1.0063 (0.1157)	0.0134	0.5000 (0.1218)	0.0148	0.0385	0.4906
		200	1.7585 (0.1071)	0.0115	0.9971 (0.0821)	0.0067	0.5008 (0.0855)	0.0073	0.0195	0.4093
		400	1.7517 (0.0654)	0.0043	1.0060 (0.0503)	0.0026	0.4936 (0.0551)	0.0031	0.0073	0.3690
		600	1.7478 (0.0591)	0.0035	1.0005 (0.0451)	0.0020	0.4972 (0.0401)	0.0016	0.0054	0.3555
		1000	1.7519 (0.0476)	0.0023	1.0030 (0.0317)	0.0010	0.5001 (0.0366)	0.0013	0.0036	0.3510

Note: For case 1 and case 2, $\alpha_{1,mode}(U_i) = 8U_i(1-U_i)+0.5$, $\alpha_{2,mode}(U_i) = 2sin(2\pi U_i)$, $\alpha_{1,mean}(U_i) = 8U_i(1-U_i)+0.75$, and $\alpha_{2,mean}(U_i) = 2sin(2\pi U_i)$; for case 3, $\alpha_{1,mode}(U_i) = \alpha_{1,mean}(U_i) = 8U_i(1-U_i)$ and $\alpha_{2,mode}(U_i) = \alpha_{2,mean}(U_i) = 2sin(2\pi U_i)$.

Table S6: The Results of Simulations for DGP 2: Less Skewed

Method	Case	n	β_1	$MSE(\beta_1)$	β_2	$MSE(\beta_2)$	β_3	$MSE(\beta_3)$	$GMSE(\boldsymbol{\beta})$	$RASE(\boldsymbol{\alpha}(U_i))$
Mode	Case 1 $\beta_{1,mode}=3$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	3.2666 (0.5561)	0.3788	1.0045 (0.4961)	0.2449	0.4895 (0.4644)	0.2147	0.6171	1.8332
		200	3.1734 (0.3299)	0.1387	1.0306 (0.2986)	0.0896	0.5289 (0.2679)	0.0723	0.2451	1.1646
		400	3.2488 (0.2362)	0.1174	1.0041 (0.1823)	0.0331	0.4983 (0.1751)	0.0305	0.1558	0.8036
		600	3.2331 (0.2026)	0.0952	1.0145 (0.1437)	0.0208	0.4829 (0.1490)	0.0224	0.1148	0.6572
		1000	3.2038 (0.1390)	0.0608	1.0057 (0.0862)	0.0074	0.4929 (0.0846)	0.0072	0.0671	0.5161
	Case 2 $\beta_{1,mode}=1$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	1.0168 (0.2425)	0.0588	0.9887 (0.2315)	0.0534	0.5068 (0.2315)	0.0534	0.1210	1.0319
		200	0.9695 (0.1472)	0.0225	1.0145 (0.1294)	0.0169	0.5115 (0.1329)	0.0177	0.0412	0.7598
		400	1.0095 (0.0858)	0.0074	1.0036 (0.0794)	0.0063	0.5000 (0.0766)	0.0058	0.0151	0.5710
		600	1.0072 (0.0597)	0.0036	1.0020 (0.0629)	0.0039	0.4944 (0.0598)	0.0036	0.0080	0.5278
		1000	0.9961 (0.0491)	0.0024	1.0015 (0.0442)	0.0019	0.4973 (0.0419)	0.0018	0.0045	0.4683
	Case 3 $\beta_{1,mode}=3$ $\beta_{2,mode}=1$ $\beta_{3,mode}=0.5$	100	3.2214 (0.3697)	0.1850	1.0139 (0.2523)	0.0635	0.4930 (0.2476)	0.0610	0.2494	0.9257
		200	3.1705 (0.2467)	0.0896	1.0059 (0.1486)	0.0220	0.5060 (0.1351)	0.0182	0.1149	0.6072
		400	3.1630 (0.1888)	0.0620	0.9960 (0.0902)	0.0081	0.49996 (0.0795)	0.0063	0.0706	0.4393
		600	3.1326 (0.1849)	0.0516	0.9974 (0.0676)	0.0046	0.4962 (0.0710)	0.0050	0.0539	0.3706
		1000	3.0853 (0.1561)	0.0316	0.9999 (0.0444)	0.0020	0.5026 (0.0456)	0.0021	0.0348	0.3151
Mean	Case 1 $\beta_{1,mean}=3.25$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	3.3018 (0.5701)	0.3261	1.0021 (0.5206)	0.2697	0.4905 (0.4933)	0.2423	0.6050	1.6082
		200	3.1939 (0.3426)	0.1199	1.0303 (0.3279)	0.1079	0.5278 (0.2978)	0.0890	0.2344	1.1027
		400	3.2955 (0.2668)	0.0729	0.9966 (0.2398)	0.0572	0.4877 (0.2127)	0.0452	0.1334	0.7810
		600	3.2618 (0.2119)	0.0448	1.0161 (0.1994)	0.0398	0.4890 (0.1952)	0.0380	0.0826	0.6363
		1000	3.2343 (0.1645)	0.0272	1.0179 (0.1449)	0.0212	0.4887 (0.1467)	0.0215	0.0488	0.4943
	Case 2 $\beta_{1,mean}=1$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	1.0215 (0.2782)	0.0775	0.9923 (0.2867)	0.0818	0.5011 (0.2646)	0.0697	0.1675	0.9593
		200	0.9559 (0.1814)	0.0347	1.0207 (0.1745)	0.0307	0.5217 (0.1710)	0.0296	0.0656	0.6713
		400	1.0270 (0.1354)	0.0190	0.9988 (0.1261)	0.0158	0.4936 (0.1244)	0.0154	0.0378	0.4916
		600	1.0084 (0.1064)	0.0113	1.0106 (0.1099)	0.0121	0.4922 (0.1098)	0.0121	0.0231	0.4090
		1000	0.9969 (0.0850)	0.0072	1.0066 (0.0754)	0.0057	0.4934 (0.0814)	0.0066	0.0131	0.3305
	Case 3 $\beta_{1,mean}=3.25$ $\beta_{2,mean}=1$ $\beta_{3,mean}=0.5$	100	3.2785 (0.3633)	0.1321	1.0108 (0.2823)	0.0794	0.4882 (0.2749)	0.0753	0.2149	0.8753
		200	3.2376 (0.2296)	0.0526	1.0098 (0.1878)	0.0352	0.5074 (0.1577)	0.0248	0.0877	0.5930
		400	3.2682 (0.1737)	0.0303	0.9979 (0.1348)	0.0181	0.4944 (0.1130)	0.0127	0.0472	0.4329
		600	3.2535 (0.1414)	0.0199	1.0052 (0.1074)	0.0115	0.4966 (0.1047)	0.0109	0.0305	0.3598
		1000	3.2377 (0.1033)	0.0108	1.0113 (0.0827)	0.0069	0.4951 (0.0808)	0.0065	0.0180	0.2924

Note: For case 1 and case 2, $\alpha_{1,mode}(U_i) = 8U_i(1-U_i) + 2.25$, $\alpha_{2,mode}(U_i) = 2sin(2\pi U_i)$, $\alpha_{1,mean}(U_i) = 8U_i(1-U_i) + 2$, and $\alpha_{2,mean}(U_i) = 2sin(2\pi U_i)$; for case 3, $\alpha_{1,mode}(U_i) = \alpha_{1,mean}(U_i) = 8U_i(1-U_i)$ and $\alpha_{2,mode}(U_i) = \alpha_{2,mean}(U_i) = 2sin(2\pi U_i)$.

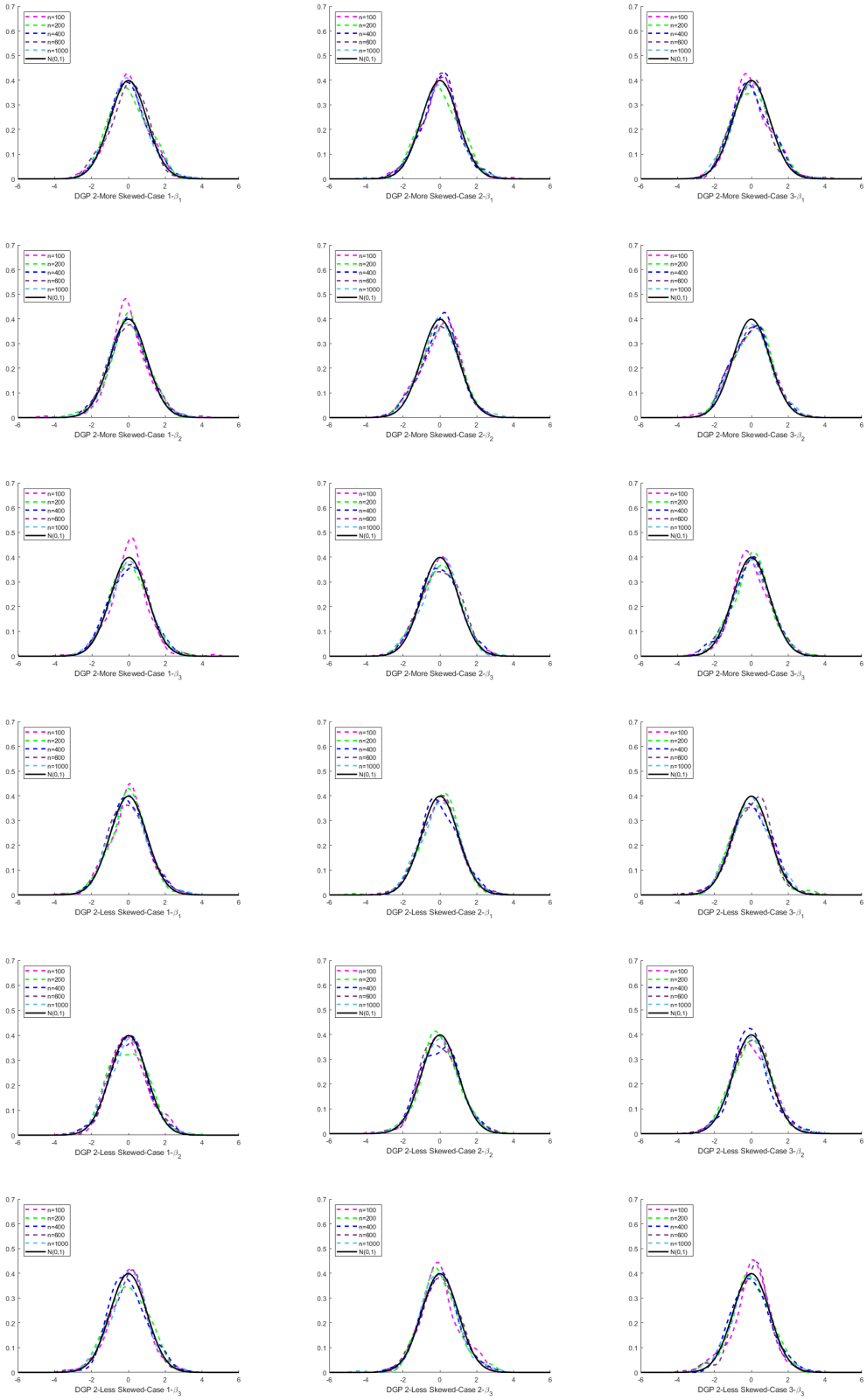

Figure S4: Empirical Density of the Standardized Estimate

Note: The first three rows are for the more skewed case, while the last three rows are for the less skewed case.

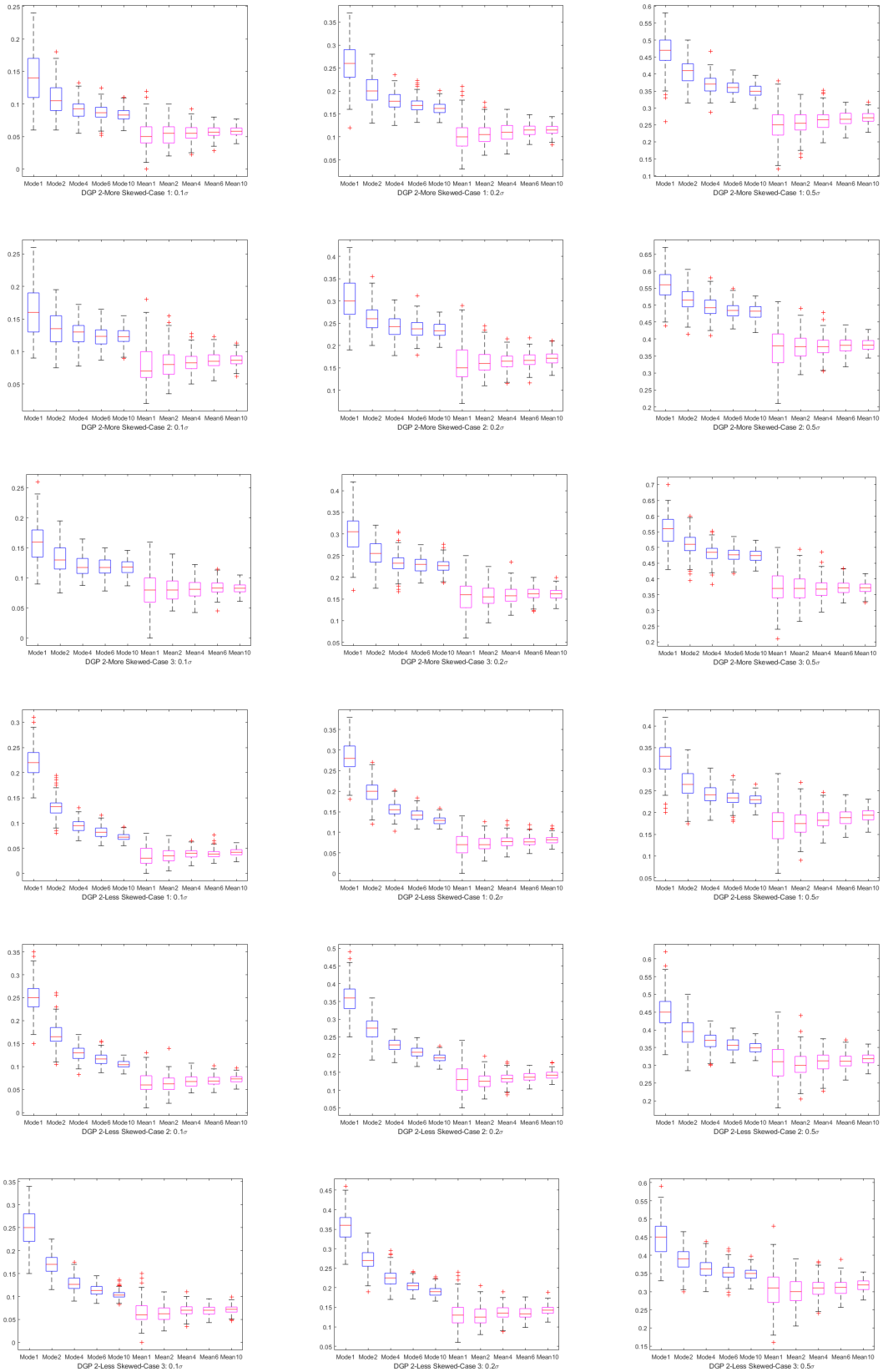

Figure S5: Boxplot of Average of Coverage Probability: $\sigma^{more}_{\text{skewed}} \approx 0.433$ and $\sigma^{less}_{\text{skewed}} \approx 0.75$

Note: The notations in each plot are the same as those of Figure 4.

S5 Proofs of Theorems

For convenience and simplicity, throughout the following parts of this supplementary note, we use δ_n to denote a variable associated with bandwidths and sample size n, and use c to represent a positive constant, which may take different forms at different places.

S5-1: Proof of Theorem 2.1

Recall that $Y_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) + \boldsymbol{Z}_i^T\boldsymbol{\beta} + \epsilon_i$. Defining $\boldsymbol{X}_i^{*T} = (\boldsymbol{X}_i^T, \boldsymbol{X}_i^T(U_i-u)/h_2, \boldsymbol{Z}_i^T)$, $\boldsymbol{\theta} = (\boldsymbol{\alpha}(u)^T, \boldsymbol{b}(u)^T, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0(u)^T, \boldsymbol{b}_0(u)^T, \boldsymbol{\beta}_0^T)^T$, $H = \text{diag}(\underbrace{1,\cdots,1}_{p}, \underbrace{h_2,\cdots,h_2}_{p}, \underbrace{1,\cdots,1}_{d})$, $\boldsymbol{\theta}_1 = H\boldsymbol{\theta}$, and $\boldsymbol{\theta}_{10} = H\boldsymbol{\theta}_0$, we then achieve

$$Q_n(\boldsymbol{\theta}) = \frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) - \boldsymbol{X}_i^T(\boldsymbol{\alpha}(u) + \boldsymbol{b}(u)(U_i-u))}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right).$$

Defining $\delta_n = h_1^2 + h_2^2 + \sqrt{(nh_1^3h_2)^{-1}}$, it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_{10}\right)\right\} \geq 1-\eta,$$

where $\boldsymbol{\theta}_{10}$ is the true value of the parameter. The above equation implies that with probability tending to one, there is a local maximum in the ball $\{\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu} : \|\mu\| \leq c\}$. Applying Taylor expansion, it follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\theta}_{10}\right)\\
=&\frac{1}{nh_1h_2}\sum_{i=1}^{n}\left[\phi\left(\frac{\epsilon_i + R(\boldsymbol{X}_i,U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\right.\\
&\left.-\frac{1}{nh_1h_2}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + R(\boldsymbol{X}_i,U_i)}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\right]\\
=&\frac{1}{nh_1h_2}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\epsilon_i + R(\boldsymbol{X}_i,U_i)}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\right.\\
&+\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + R(\boldsymbol{X}_i,U_i)}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)^2K\left(\frac{U_i-u}{h_2}\right)\\
&\left.-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)^3K\left(\frac{U_i-u}{h_2}\right)\right]\\
=&I_1 + I_2 + I_3,
\end{aligned}$$

where ϵ_i^* is between $\epsilon_i + R(\boldsymbol{X}_i,U_i)$ and $\epsilon_i + R(\boldsymbol{X}_i,U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*$, and $R(\boldsymbol{X}_i,U_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)$

$-\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(u)+\boldsymbol{b}_0(u)(U_i-u))$. Based on the result $T_n=\mathbb{E}\left(T_n\right)+O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is $I_1=\frac{1}{nh_1h_2}\sum_{i=1}^n\left(-\phi^{(1)}\left(\frac{\epsilon_i+R(\boldsymbol{X}_i,U_i)}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\right)$, by Taylor expansion, we can re-write it as

$$\begin{aligned}
\mathbb{E}(I_1)&=\frac{-\delta_n}{h_1h_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i+R(\boldsymbol{X}_i,U_i)}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)\right)\\
&=\frac{-\delta_n}{h_1h_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)+\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{R(\boldsymbol{X}_i^*,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1^2}K\left(\frac{U_i-u}{h_2}\right)\right.\\
&\qquad\left.+\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{**}}{h_1}\right)\frac{R^2(\boldsymbol{X}_i^*,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\right)\\
&=I_{11}+I_{12}+I_{13},
\end{aligned}$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i+R(\boldsymbol{X}_i,U_i)$. Note that under the same conditions as Theorem 2.1, 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

$$\begin{aligned}
I_{11}=&\frac{-\delta_n}{h_1h_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=&\frac{-\delta_n}{h_1h_2}\iiint\phi^{(1)}\left(\frac{\epsilon}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}^*}{h_1}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{\delta_n}{h_1}\iiint\phi\left(\tau\right)\tau\boldsymbol{\mu}^T\boldsymbol{X}^*f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K\left(w\right)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=&O_p(\delta_nch_1^2).
\end{aligned}$$

$$\begin{aligned}
I_{12}=&\frac{-\delta_n}{h_1h_2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)\frac{R(\boldsymbol{X}_i^*,U_i)}{h_1}\right)\\
=&\frac{-\delta_n}{h_1h_2}\iiint\phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}^*}{h_1}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)\frac{R(\boldsymbol{X}^*,U)}{h_1}f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{-\delta_n}{h_1}\iiint\phi\left(\tau\right)(\tau^2-1)\boldsymbol{\mu}^T\boldsymbol{X}^*f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K\left(w\right)\frac{R(\boldsymbol{X}^*,U)}{h_1}f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=&O_p(\delta_nch_2^2).
\end{aligned}$$

$$\begin{aligned}
I_{13}\approx&\frac{-\delta_n}{h_1h_2}\mathbb{E}\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{R^2(\boldsymbol{X}_i^*,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=&\frac{-\delta_n}{2h_1h_2}\iiint\phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{R^2(\boldsymbol{X}^*,U)\boldsymbol{\mu}^T\boldsymbol{X}^*}{h_1^3}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{-\delta_nh_2^4}{2}\iiint\phi(\tau)(3\tau-\tau^3)\frac{(\boldsymbol{X}^T\boldsymbol{\alpha}^{(2)}(u))^2\boldsymbol{\mu}^T\boldsymbol{X}^*}{4h_1^3}f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)w^4\\
&f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\{1+o_p(1)\}=o_p(\delta_nch_2^2).
\end{aligned}$$

Meanwhile, with the condition $h_2^2/h_1 \to 0$ held, we obtain

$$\begin{aligned}
&\frac{\delta_n^2}{h_1^2 h_2^2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)\right)^2\\
=&\frac{\delta_n^2}{h_1^2 h_2^2}\iiint \phi^{(1)2}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2}{h_1^2}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{\delta_n^2}{h_1^3 h_2}\iiint \phi^2(\tau)\tau^2(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2 f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K^2(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=&O_p(\delta_n^2 c^2(h_1^3h_2)^{-1}).
\end{aligned}$$

$$\begin{aligned}
&\frac{\delta_n^2}{h_1^2 h_2^2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}K\left(\frac{U_i-u}{h_2}\right)\frac{R(\boldsymbol{X}_i^*,U_i)}{h_1}\right)^2\\
=&\frac{\delta_n^2}{h_1^2 h_2^2}\iiint \phi^{(2)2}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2}{h_1^2}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_2}\right)\frac{R^2(\boldsymbol{X}^*,U)}{h_1^2}f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{\delta_n^2 h_2^3}{h_1^5}\iiint \phi^2(\tau)(\tau^2-1)^2(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2 f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})w^4K^2(w)\frac{(\boldsymbol{X}^T\boldsymbol{\alpha}^{(2)}(u))^2}{4}\\
&f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\{1+o_p(1)\}=o_p(\delta_n^2c^2(h_1^3h_2)^{-1}).
\end{aligned}$$

These indicate $I_1 = O_p(\delta_n c(h_1^2+h_2^2)) + O_p(\sqrt{\delta_n^2c^2(nh_1^3h_2)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_2 = \frac{1}{nh_1h_2}\sum_{i=1}^n\left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i+R(\boldsymbol{X}_i,U_i)}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)^2K\left(\frac{U_i-u}{h_2}\right)\right)$, we can re-write it as

$$\begin{aligned}
\mathbb{E}(I_2) =&\frac{\delta_n^2}{2h_2h_1}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i+R(\boldsymbol{X}_i,U_i)}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^2}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=&\frac{\delta_n^2}{2h_2h_1}\mathbb{E}\Bigg(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^2}K\left(\frac{U_i-u}{h_2}\right)+\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{R(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\\
&+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{**}}{h_1}\right)\frac{R^2(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^4}K\left(\frac{U_i-u}{h_2}\right)\Bigg)\\
=&I_{21}+I_{22}+I_{23}.
\end{aligned}$$

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 direct calculations for each part, we can obtain

$$\begin{aligned}
I_{21} =&\frac{\delta_n^2}{2h_2h_1}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^2}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=&\frac{\delta_n^2}{2h_2h_1}\iiint \phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2}{h_1^2}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{\delta_n^2}{2h_1^2}\iiint \phi(\tau)(\tau^2-1)(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2 f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=&O_p((\delta_n c)^2).
\end{aligned}$$

$$
\begin{aligned}
I_{22} =& \frac{\delta_n^2}{2h_2h_1}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{R(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=& \frac{\delta_n^2}{2h_2h_1}\iiint \phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{R(\boldsymbol{X},U)(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2}{h_1^3}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=& \frac{\delta_n^2h_2^2}{2h_1^3}\iiint \phi(\tau)(3\tau-\tau^3)\frac{\boldsymbol{X}^T\boldsymbol{\alpha}^{(2)}(u)}{2}(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})w^2K(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
& \{1+o_p(1)\}=o_p((\delta_nc)^2).
\end{aligned}
$$

Meanwhile, we can prove that $I_{23}=o_p((\delta_nc)^2)$. Following the same steps in (i), we obtain the following result

$$
\begin{aligned}
& \frac{\delta_n^4}{4h_2^2h_1^2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{h_1^2}K\left(\frac{U_i-u}{h_2}\right)\right)^2\\
=& \frac{\delta_n^4}{4h_2^2h_1^2}\iiint \phi^{(2)2}\left(\frac{\epsilon}{h_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^4}{h_1^4}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=& \frac{\delta_n^4}{4h_2h_1^2}\iiint \phi^2(\tau)(\tau^2-1)^2\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^4}{h_1^4}f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K^2(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=& O_p((\delta_nc)^4(h_2h_1^5)^{-1}).
\end{aligned}
$$

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) The same way to calculate the third part. As the order of ϵ_i^* is the same as the order of ϵ_i, which indicates we can obtain $I_3\approx\frac{1}{nh_1h_2}\sum_{i=1}^n\left(-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{h_1}\right)^3K\left(\frac{U_i-u}{h_2}\right)\right)$. By directly calculating, we arrive at

$$
\begin{aligned}
& \frac{\delta_n^3}{6h_2h_1}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{i}}^*\right)^3}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\right)\\
=& \frac{\delta_n^3}{6h_2h_1}\iiint \phi^{(3)}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}^*\right)^3}{h_1^3}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=& \frac{\delta_n^3}{6}\iiint \phi(\tau)(3\tau-\tau^3)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}^*\right)^3}{h_1^3}f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=& O_p(\delta_n^3c^3).
\end{aligned}
$$

$$
\begin{aligned}
& \frac{\delta_n^6}{36h_2^2h_1^2}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_1}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{i}}^*\right)^3}{h_1^3}K\left(\frac{U_i-u}{h_2}\right)\right)^2\\
=& \frac{\delta_n^6}{36h_2^2h_1^2}\iiint \phi^{(3)2}\left(\frac{\epsilon}{h_1}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}^*\right)^6}{h_1^6}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=& \frac{\delta_n^6}{36h_2h_1}\iiint \phi^2(\tau)(3\tau-\tau^3)^2\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}^*\right)^6}{h_1^6}f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K^2(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=& O_p(\delta_n^6c^6(h_2h_1^7)^{-1}).
\end{aligned}
$$

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, $P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_{10}+\delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_{10}\right)\right\} \geq 1-\eta$ holds. Hence with the probability approaching one, there exists local maximizers $\hat{\boldsymbol{\alpha}}(u)$, $\hat{\boldsymbol{b}}(u)$ and $\hat{\boldsymbol{\beta}}$ such that

$$\|\hat{\boldsymbol{\alpha}}(u)-\hat{\boldsymbol{\alpha}}_0(u)\| \leq \delta_n c, \quad \|\hat{\boldsymbol{b}}(u)h_2-\hat{\boldsymbol{b}}_0(u)h_2\| \leq \delta_n c, \quad \text{and} \quad \|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| \leq \delta_n c.$$

□

S5-2: Proof of Theorem 2.2

Following the same steps as proving Theorem 2.1, recall that

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i)+\boldsymbol{Z}_i^T\boldsymbol{\beta}+\epsilon_i = \boldsymbol{X}_i^T(\boldsymbol{\alpha}(u)+\boldsymbol{b}(u)(U_i-u))+\boldsymbol{Z}_i^T\boldsymbol{\beta}+\epsilon_i+R_1(\boldsymbol{X}_i,U_i),$$

where $R_1(\boldsymbol{X}_i,U_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i)-\boldsymbol{X}_i^T(\boldsymbol{\alpha}(u)+\boldsymbol{b}(u)(U_i-u))$. Defining $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}(u)^T, \hat{\boldsymbol{b}}(u)^T, \hat{\boldsymbol{\beta}}^T)^T$ and $\hat{\boldsymbol{\theta}}_1 = H\hat{\boldsymbol{\theta}}$, then $\hat{\boldsymbol{\theta}}_1$ must satisfy the following equation

$$-\frac{1}{nh_1^2h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i+R(\boldsymbol{X}_i,U_i,\boldsymbol{Z}_i)}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^* = 0,$$

where $R(\boldsymbol{X}_i,U_i,\boldsymbol{Z}_i) = R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})$.

By taking Taylor expansion, we can obtain

$$\begin{aligned}
&-\frac{1}{nh_1^2h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^* \\
&+\frac{1}{nh_1^3h_2}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*(R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})) \\
&-\frac{1}{nh_1^4h_2}\sum_{i=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^*}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\left(R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})\right)^2 = 0,
\end{aligned}$$

where $\tilde{\epsilon}_i^*$ is between ϵ_i and $\epsilon_i+R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})$. From Theorem 2.1, we know $\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}\| = O_p(\delta_n)$, which indicates that

$$\begin{aligned}
\sup_{i:|U_i-u|/h_2\leq 1}|R(\boldsymbol{X}_i,U_i,\boldsymbol{Z}_i)| &\leq \sup_{i:|U_i-u|/h_2\leq 1}\{|R(\boldsymbol{X}_i,U_i)|+|\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})|\} \\
&= O_p(\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})\|) = O_p(\delta_n).
\end{aligned}$$

Combining this with the Proof of Theorem 2.1, we can see that the third part which is associated

with $(R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10}))^2$ is dominated by the second part which is associated with $R(\boldsymbol{X}_i,U_i)-\boldsymbol{X}_i^{*T}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_{10})$. We then mainly focus on the first two parts of the above equation.

Considering $-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*+\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*R(\boldsymbol{X}_i,U_i)$, by some direct calculations, we can obtain

$$
\begin{aligned}
&\mathbb{E}\bigg(-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\\
&\quad+\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*(R(\boldsymbol{X}_i,U_i))\bigg)\\
=&-\frac{1}{h_1^2h_2}\iiint\phi^{(1)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
&\quad+\frac{1}{h_1^3h_2}\iiint\phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)R(\boldsymbol{X},U)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{1}{h_1}\iiint\phi(\tau)\tau\boldsymbol{X}^*f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
&\quad-\frac{1}{h_1^2}\iiint\phi(\tau)(\tau^2-1)\boldsymbol{X}^*f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)R(\boldsymbol{X},U)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})\\
=&\frac{h_1^2}{2}f_U(u)\begin{bmatrix}\mathbb{E}(\boldsymbol{X}f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\\ \mathbb{E}(\boldsymbol{Z}f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}-\left(\frac{h_2^2\boldsymbol{\alpha}^{(2)}(u)}{2}f_U(u)\begin{bmatrix}\mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\\ \mu_2\mathbb{E}(\boldsymbol{Z}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}\right)\{1+o_p(1)\}.
\end{aligned}
$$

Considering $\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T}$, by directly calculating, we have

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T}\right)\\
=&\frac{1}{h_1^3h_2}\iiint\phi^{(2)}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{1}{h_1^2}\iiint\phi(\tau)(\tau^2-1)\boldsymbol{X}^*\boldsymbol{X}^{*T}f_\epsilon(\tau h_1|\hat{\boldsymbol{X}})K(w)f_U(wh_2+u)dwd\tau dF(\hat{\boldsymbol{X}})(1+o_p(1))\\
=&f_U(u)\begin{bmatrix}\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & \mathbb{E}(\boldsymbol{X}\boldsymbol{Z}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0} & \mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ \mathbb{E}(\boldsymbol{Z}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & \mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}.
\end{aligned}
$$

Meanwhile, with the condition $h_2^2/h_1\to 0$ held, we can get

$$
\begin{aligned}
&\mathrm{Var}\bigg(-\frac{1}{nh_1^2h_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\\
&+\frac{1}{nh_1^3h_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*(R(\boldsymbol{X}_i,U_i))\bigg)
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}\left(-\frac{1}{nh_1^2h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\right)\left(-\frac{1}{nh_1^2h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*\right)^T\\
&(1+o_p(1))\\
=&\frac{1}{nh_1^4h_2^2}\iiint\phi^{(1)2}\left(\frac{\epsilon}{h_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})(1+o_p(1))\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_1^3h_2}f_U(u)\begin{bmatrix}v_0\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & v_0\mathbb{E}(\boldsymbol{X}\boldsymbol{Z}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0} & v_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ v_0\mathbb{E}(\boldsymbol{Z}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & v_0\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}\\
&(1+o_p(1)).
\end{aligned}
$$

Define $W_n=\frac{1}{nh_1^2h_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)K\left(\frac{U_i-u}{h_2}\right)\boldsymbol{X}_i^*$. To show Theorem 2.2, it is sufficient to show that

$$T_n=\sqrt{nh_2h_1^3}W_n\xrightarrow{d}\mathcal{N}(0,T),$$

where

$$T=\int\tau^2\phi^2(\tau)d\tau f_U(u)\begin{bmatrix}v_0\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & v_0\mathbb{E}(\boldsymbol{X}\boldsymbol{Z}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0} & v_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ v_0\mathbb{E}(\boldsymbol{Z}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0} & v_0\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}.$$

Then, by Slutsky's theorem and the above two equations, we can achieve Theorem 2.2. To show the preceding equation, we prove that for any unit vector $\boldsymbol{d}\in\mathbb{R}^p$,

$$\left\{\boldsymbol{d}^T\operatorname{Cov}(T_n)\boldsymbol{d}\right\}^{-1/2}\left\{\boldsymbol{d}^TT_n-\boldsymbol{d}^T\mathbb{E}(T_n)\right\}\xrightarrow{d}N(0,1).$$

We then check Lyapunov's condition. Let

$$\xi_i=\sqrt{h_2h_1^3/n}K\left(\frac{U-u}{h_2}\right)\frac{1}{h_1h_2}\phi^{(1)}\left(\frac{\epsilon_i}{h_1}\right)\boldsymbol{d}^T\boldsymbol{X}_i^*,$$

we need to prove $n\mathbb{E}|\xi_1|^3\to 0$. 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

$$n\mathbb{E}|\xi|^3\le O\left(n^{-1/2}h_2^{-3/2}h_1^{3/2}\right)\mathbb{E}\left|K^3\left(\frac{U-u}{h_2}\right)\phi^{(1)3}\left(\frac{\epsilon_i}{h_1}\right)\boldsymbol{d}^T\boldsymbol{X}_i^*\right|\to 0.$$

Thus, the asymptotic normality for T_n holds.

□

S5-3: Proof of Theorem 2.3

The proof is similar to Theorem 2.1, except that we need to take the estimation error from the-

first stage into consideration. Hence, we provide a sketch of the proof here. Recall that

$$Q_n(\boldsymbol{\beta}) = \frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{h_3}\right) = \frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right).$$

Define $\delta_n = h_3^2 + \sqrt{(nh_3^3)^{-1}}$. It is sufficient to show that for any given η, there exists a large number constant c such that $P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\beta}_0\right)\right\} \geq 1-\eta$, where $\boldsymbol{\beta}_0$ is the true value of the parameter. Using Taylor expansion, it follows that

$$\begin{aligned}
&Q_n\left(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\beta}_0\right)\\
=&\frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)\\
&\quad - \frac{1}{nh_3}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\\
=&\frac{1}{nh_3}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)\right.\\
&\quad \left.+\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)^2 - \frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{h_3}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)^3\right]\\
=&I_4 + I_5 + I_6,
\end{aligned}$$

where ϵ_i^* is between $\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)$ and $\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i$. Based on the result $T_n = \mathbb{E}\left(T_n\right) + O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we consider each part of the above Taylor expansion.

(i) For the first part, which is $I_4 = \frac{1}{nh_3}\sum_{i=1}^{n}\left(-\phi^{(1)}\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)\right)$, we can re-write it as

$$\begin{aligned}
\mathbb{E}(I_4) =&\frac{-\delta_n}{h_3}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)\\
=&\frac{-\delta_n}{h_3}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3} + \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3^2}\right.\\
&\quad \left.+\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{***}}{h_3}\right)\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))^2\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3^3}\right)\\
=&I_{41} + I_{42} + I_{43},
\end{aligned}$$

where ϵ_i^{***} is between ϵ_i and $\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)$. Notice that as the order of ϵ_i^{***} is the same as that of ϵ_i, when we do the calculations associated with I_{43}, we instead use ϵ_i directly. By some direct calculations for each part, we can get

$$I_{41} = \frac{-\delta_n}{h_3}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right) = \frac{-\delta_n}{h_3}\iint\phi^{(1)}\left(\frac{\epsilon}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}}{h_3}f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p(\delta_n c h_3^2).$$

$$
\begin{aligned}
I_{42} =& \frac{-\delta_n}{h_3}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))}{h_3}\right)\\
=& \frac{-\delta_n}{h_3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}}{h_3}f_\epsilon(\epsilon|\boldsymbol{X})\frac{\boldsymbol{X}^T\boldsymbol{\alpha}_0(U)-\boldsymbol{X}^T\hat{\boldsymbol{\alpha}}(U)}{h_3}d\epsilon dF(\boldsymbol{X}) = O_p(\delta_n c(h_1^2+h_2^2)).
\end{aligned}
$$

With the conditions that $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$, it can be seen that I_{41} dominates I_{42}. Meanwhile, according to the result in the Proof of Theorem 2.1, we could easily prove that I_{42} dominates I_{43}. We then obtain

$$
\frac{\delta_n^2}{h_3^2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)^2 = \frac{\delta_n^2}{h_3^2}\iint \phi^{(1)2}\left(\frac{\epsilon}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X})^2}{h_3^2}f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p(\delta_n^2 c^2 h_3^{-3}).
$$

These indicate $I_4 = O_p(\delta_n c h_3^2) + O_p(\sqrt{\delta_n^2 c^2 (nh_3^3)^{-1}}) = O_p(\delta_n^2 c)$.

(ii) For the second part, which is $I_5 = \frac{1}{nh_3}\sum_{i=1}^n\left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i+\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i}{h_3}\right)^2\right)$, we can re-write it as

$$
\begin{aligned}
\mathbb{E}(I_5) =& \frac{\delta_n^2}{2h_3}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i+\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i)}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^2}\right)\\
=& \frac{\delta_n^2}{2h_3}\mathbb{E}\Bigg(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^2} + \phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^3}\\
&+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{***}}{h_3}\right)\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))^2(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^4}\Bigg)\\
=& I_{51}+I_{52}+I_{53},
\end{aligned}
$$

As the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{53}, we instead use ϵ_i directly. By some calculations for each part, we can achieve

$$
I_{51} = \frac{\delta_n^2}{2h_3}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^2}\right) = \frac{\delta_n^2}{2h_3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X})^2}{h_3^2}f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p((\delta_n c)^2).
$$

$$
\begin{aligned}
I_{52} =& \frac{\delta_n^2}{2h_3}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i)-\boldsymbol{X}_i^T\hat{\boldsymbol{\alpha}}(U_i))(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^3}\right)\\
=& \frac{\delta_n^2}{2h_3}\iint \phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{X}^T\boldsymbol{\alpha}_0(U)-\boldsymbol{X}^T\hat{\boldsymbol{\alpha}}(U))(\boldsymbol{\mu}^T\boldsymbol{X})^2}{h_3^3}f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = o_p((\delta_n c)^2).
\end{aligned}
$$

Meanwhile, we can prove that $I_{53} = o_p((\delta_n c)^2)$ and get the following result

$$
\frac{\delta_n^4}{4h_3^2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i)^2}{h_3^2}\right)^2 = \frac{\delta_n^4}{4h_3^2}\iint \phi^{(2)2}\left(\frac{\epsilon}{h_3}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X})^4}{h_3^4}f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p((\delta_n c)^4 h_3^{-5}).
$$

These imply 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_i^{***} is the same as the order of ϵ_i, we can obtain $I_6 \approx \frac{1}{nh_3}\sum_{i=1}^{n} -\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\left(\frac{\delta_n \boldsymbol{\mu}^T \boldsymbol{X}_i}{h_3}\right)^3$. By directly calculating, we get

$$\frac{\delta_n^3}{6h_3}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X_i}\right)^3}{h_3^3}\right) = \frac{\delta_n^3}{6h_3}\iint \phi^{(3)}\left(\frac{\epsilon}{h_3}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}\right)^3}{h_3^3} f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p(\delta_n^3).$$

$$\frac{\delta_n^6}{36h_3^2}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{h_3}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X_i}\right)^3}{h_3^3}\right)^2 = \frac{\delta_n^6}{36h_3^2}\iint \phi^{(3)2}\left(\frac{\epsilon}{h_3}\right)\frac{\left(\boldsymbol{\mu}^{\boldsymbol{T}}\boldsymbol{X}\right)^6}{h_3^6} f_\epsilon(\epsilon|\boldsymbol{X})d\epsilon dF(\boldsymbol{X}) = O_p(\delta_n^6 h_3^{-7}).$$

These demonstrate that the second part dominates the third part.

Based on these, we can choose c bigger enough such that the second term dominates the other two terms with probability $1-\eta$. Because the second term is negative, $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}) < Q_n(\boldsymbol{\beta}_0)\} \geq 1-\eta$ holds. Hence with the probability approaching one, there exists a maximizer $\tilde{\boldsymbol{\beta}}$ such that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq \delta_n c$.

□

S5-4: Proof of Theorem 2.4

Following the same steps as proving Theorem 2.3, since $\tilde{\boldsymbol{\beta}}$ maximizes $Q_n(\boldsymbol{\beta})$, we can take the derivative of $Q_n(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ to obtain

$$\frac{dQ_n(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = -\frac{1}{nh_3^2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i + \boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)}{h_3}\right)\boldsymbol{Z}_i = 0.$$

By taking Taylor expansion, we get

$$\begin{aligned}-\frac{1}{nh_3^2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i + \frac{1}{nh_3^3}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))\\ - \frac{1}{nh_3^4}\sum_{i=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^{**}}{h_3}\right)\boldsymbol{Z}_i(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^2 = 0,\end{aligned}$$

where $\tilde{\epsilon}_i^{**}$ is between ϵ_i and $\epsilon_i + \boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. From Theorem 2.3, we know $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\delta_n)$, which indicates that

$$\begin{aligned}|\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| &\leq \{|\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i))| + |\boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|\}\\ &= O_p(\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|) = O_p(\delta_n).\end{aligned}$$

It can be seen that the third part which is associated with $(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))^2$ is dominated by the second part which is associated with $\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i) - \hat{\boldsymbol{\alpha}}(U_i)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$.

We then mainly focus on the first two parts of the above equation.

Considering $-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i+\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i)-\hat{\boldsymbol{\alpha}}(U_i)))$, with the conditions that $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$, by some direct calculations, we can obtain

$$
\begin{aligned}
&\mathbb{E}\left(-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i+\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i)-\hat{\boldsymbol{\alpha}}(U_i)))\right)\\
=&-\frac{1}{h_3^2}\iint \phi^{(1)}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{Z}f_\epsilon(\epsilon|\boldsymbol{Z})d\epsilon dF(\boldsymbol{Z})\\
&+\frac{1}{h_3^3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{Z}f_\epsilon(\epsilon|\boldsymbol{Z})(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i)-\hat{\boldsymbol{\alpha}}(U_i)))d\epsilon dF(\hat{\boldsymbol{X}})\\
=&\frac{1}{h_3}\iint \phi(\tau)\tau\boldsymbol{Z}f_\epsilon(\tau h_3|\boldsymbol{Z})d\tau dF(\boldsymbol{Z})\\
&-\frac{1}{h_3^2}\iint \phi(\tau)(\tau^2-1)\boldsymbol{Z}f_\epsilon(\tau h_3|\boldsymbol{Z})(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i)-\hat{\boldsymbol{\alpha}}(U_i)))d\tau dF(\hat{\boldsymbol{X}})\\
=&\frac{h_3^2}{2}\mathbb{E}(\boldsymbol{Z}f_\epsilon^{(3)}(0|\boldsymbol{Z})|u)\{1+o_p(1)\}.
\end{aligned}
$$

Considering $\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i\boldsymbol{Z}_i^T$, by directly calculating, we have

$$
\begin{aligned}
&\mathbb{E}\left(\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i\boldsymbol{Z}_i^T\right)=\frac{1}{h_3^3}\iint \phi^{(2)}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(\epsilon|\boldsymbol{Z})d\epsilon dF(\boldsymbol{Z})\\
=&\frac{1}{h_3^2}\iint \phi(\tau)(\tau^2-1)\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(\tau h_3|\boldsymbol{Z})d\tau dF(\boldsymbol{Z})=\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon^{(2)}(0|\boldsymbol{Z})|u).
\end{aligned}
$$

Based on the above two equations, we can get

$$
\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=\frac{h_3^2}{2}\left(\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon^{(2)}(0|\boldsymbol{Z})|u)\right)^{-1}\mathbb{E}(\boldsymbol{Z}f_\epsilon^{(3)}(0|\boldsymbol{Z})|u)(1+o_p(1)).
$$

Meanwhile, with the conditions $h_1/h_3 \to 0$ and $h_2/h_3 \to 0$ held, we can obtain

$$
\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{nh_3^2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i+\frac{1}{nh_3^3}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{h_3}\right)\boldsymbol{Z}_i(\boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(U_i)-\hat{\boldsymbol{\alpha}}(U_i)))\right)\\
=&\frac{1}{nh_3^4}\iint \phi^{(1)2}\left(\frac{\epsilon}{h_3}\right)\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(\epsilon|\boldsymbol{Z})d\epsilon dF(\boldsymbol{Z})(1+o_p(1))\\
=&\frac{\int \tau^2\phi^2(\tau)d\tau}{nh_3^3}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^T f_\epsilon(0|\boldsymbol{Z}))\{1+o_p(1)\}.
\end{aligned}
$$

For the remaining part, we can follow the same idea in the Proof of Theorem 2.2 to easily achieve the result.

□

S5-5: Proof of Theorem 2.5

The proof is similar to Theorem 2.1, except that we need to take the estimation errors from the previous stages into consideration. Define $\tilde{\boldsymbol{X}}_i^T = (\boldsymbol{X}_i^T, \boldsymbol{X}_i^T(U_i - u)/h_5)$, $\boldsymbol{\theta}^* = (\boldsymbol{\alpha}(u)^T, \boldsymbol{b}(u)^T)^T$, $\boldsymbol{\theta}_0^* = (\boldsymbol{\alpha}_0(u)^T, \boldsymbol{b}_0(u)^T)^T$, $H = \text{diag}(\underbrace{1, \cdots, 1}_{p}, \underbrace{h_5, \cdots, h_5}_{p})$, $\tilde{\boldsymbol{\theta}}_1 = H\boldsymbol{\theta}^*$, and $\tilde{\boldsymbol{\theta}}_{10} = H\boldsymbol{\theta}_0^*$.

Let $\delta_n = h_4^2 + h_5^2 + \sqrt{(nh_4^3h_5)^{-1}}$. It is sufficient to show that for any given η, there exists a large number constant c such that $P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\tilde{\boldsymbol{\theta}}_{10} + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\tilde{\boldsymbol{\theta}}_{10}\right)\right\} \geq 1 - \eta$, where $\tilde{\boldsymbol{\theta}}_{10}$ is the true value of the parameter. Using Taylor expansion, it follows that

$$
\begin{aligned}
& Q_n\left(\tilde{\boldsymbol{\theta}}_{10} + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\tilde{\boldsymbol{\theta}}_{10}\right) \\
=& \frac{1}{nh_4h_5}\sum_{i=1}^{n}\left[\phi\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)K\left(\frac{U_i - u}{h_5}\right)\right. \\
& \left. - \frac{1}{nh_4h_5}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{h_4}\right)K\left(\frac{U_i - u}{h_5}\right)\right] \\
=& \frac{1}{nh_4h_5}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{h_4}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)K\left(\frac{U_i - u}{h_5}\right)\right. \\
& + \frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{h_4}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)^2 K\left(\frac{U_i - u}{h_5}\right) \\
& \left. - \frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^{\Delta}}{h_4}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)^3 K\left(\frac{U_i - u}{h_5}\right)\right] \\
=& I_7 + I_8 + I_9,
\end{aligned}
$$

where ϵ_i^{Δ} is between $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)$ and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i$ in which $\tilde{R}(\boldsymbol{X}_i, U_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(u) + \boldsymbol{b}_0(u)(U_i - u)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$. Based on the result $T_n = \mathbb{E}(T_n) + O_p(\sqrt{\text{Var}(T_n)})$, we could consider each part of the above Taylor expansion.

(i) For the first part, which is $I_7 = \frac{1}{nh_4h_5}\sum_{i=1}^{n}\left(-\phi^{(1)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{h_4}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)K\left(\frac{U_i - u}{h_5}\right)\right)$, we can get

$$
I_7 = O_p(\delta_n c(h_4^2 + h_5^2)) + O_p(\sqrt{\delta_n^2c^2(nh_4^3h_5)^{-1}}) = O_p(\delta_n^2 c)
$$

by combining the results obtained from the Proofs of Theorem 2.1 and 2.3 and the assumptions $h_3/h_5 \to 0$ and $h_5^2/h_4 \to 0$.

(ii) For the second part, $I_8 = \frac{1}{nh_4h_5}\sum_{i=1}^{n}\left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{h_4}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\tilde{\boldsymbol{X}}_i}{h_4}\right)^2 K\left(\frac{U_i - u}{h_5}\right)\right)$, by combining the results obtained from the Proof of Theorem 2.1 and assumptions $h_3/h_5 \to 0$ and $nh_4^5h_5 \to \infty$, we can see that it will dominate the first part when we choose c big enough with $\mathbb{E}(I_8) = O_p((\delta_n c)^2)$.

(iii) The same way to calculate the third part. As ϵ_i^{Δ} is between ϵ_i and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)$

$-\delta_n \boldsymbol{\mu}^T \tilde{\boldsymbol{X}}_i$, the order of ϵ_i^{Δ} is the same as the order of ϵ_i, which indicates that we can obtain $I_9 \approx \frac{1}{nh_4h_5}\sum_{i=1}^{n}\left(-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i}{h_4}\right)\left(\frac{\delta_n \boldsymbol{\mu}^T \tilde{\boldsymbol{X}}_i}{h_4}\right)^3 K\left(\frac{U_i-u}{h_5}\right)\right)$. Combining the results obtained from the Proof of Theorem 2.1, we can get that the second part dominates the third part.

Based on these, we can choose c bigger enough such that the second term dominates the other two terms with probability $1-\eta$. Because the second term is negative, $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}) < Q_n(\boldsymbol{\theta}_{10})\} \geq 1-\eta$ holds. Hence with the probability approaching one, there exists local maximizers $\tilde{\boldsymbol{\alpha}}(u)$ and $\tilde{\boldsymbol{b}}(u)$ such that

$$\|\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u)\| \leq \delta_n c \quad \text{and} \quad \|\tilde{\boldsymbol{b}}(u)h_5 - \boldsymbol{b}_0(u)h_5\| \leq \delta_n c.$$

□

S5-6: Proof of Theorem 2.6

Following the same steps as proving Theorem 2.4, since $\tilde{\boldsymbol{\theta}}_1$ maximizes $Q_n(\boldsymbol{\theta}_1)$, we can take the derivative of $Q_n(\boldsymbol{\theta}_1)$ with respect to $\boldsymbol{\theta}_1$ to obtain

$$-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i = 0,$$

where $\tilde{R}(\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = \tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})$. By taking Taylor expansion, we can obtain

$$\begin{aligned}
&-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i \\
&+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i(\tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})) \\
&-\frac{1}{nh_4^4h_5}\sum_{i=1}^{n}\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^{\Delta *}}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\left(\tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})\right)^2 = 0,
\end{aligned}$$

where $\tilde{\epsilon}_i^{\Delta *}$ is between ϵ_i and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})$. From Theorem 2.5, we know $\|\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10}\| = O_p(\delta)$, which indicates that

$$|\tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})| \leq \{|\tilde{R}(\boldsymbol{X}_i, U_i)| + |\tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})|\} = O_p(\|\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10}\|) = O_p(\delta).$$

It can be seen that the third part which is associated with $\tilde{\boldsymbol{X}}_i(\tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10}))^2$ is dominated by the second part which is associated with $\tilde{R}(\boldsymbol{X}_i, U_i) - \tilde{\boldsymbol{X}}_i^T(\tilde{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_{10})$. We then mainly focus on the first two parts of the above equation.

Considering $-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i + \frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right) K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\tilde{R}(\boldsymbol{X}_i,$

U_i) and combining the results obtained from the Proof of Theorem 2.2, with the assumption that $h_3/h_5 \to 0$ held, we get

$$\begin{aligned}
&\mathbb{E}\left(-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\tilde{R}(\boldsymbol{X}_i,U_i)\right)\\
=&\frac{h_4^2}{2}f_U(u)\begin{bmatrix}\mathbb{E}(\boldsymbol{X}f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\end{bmatrix}-\frac{h_5^2\boldsymbol{\alpha}^{(2)}(u)}{2}f_U(u)\begin{bmatrix}\mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\end{bmatrix}\{1+o_p(1)\}.
\end{aligned}$$

Considering $\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T$, by directly calculating, we have

$$\begin{aligned}
&\mathbb{E}\left(\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{X}}_i^T\right)\\
=&f_U(u)\begin{bmatrix}\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ \mathbf{0} & \mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}.
\end{aligned}$$

Based on the above two equations, we can achieve

$$\begin{aligned}
\tilde{\boldsymbol{\theta}}_1-\tilde{\boldsymbol{\theta}}_{10}=&\begin{bmatrix}\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ \mathbf{0} & \mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}^{-1}\\
&\left(\frac{h_4^2}{2}f_U(u)\begin{bmatrix}\mathbb{E}(\boldsymbol{X}f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\end{bmatrix}-\frac{h_5^2\boldsymbol{\alpha}^{(2)}(u)}{2}f_U(u)\begin{bmatrix}\mu_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\\ \mathbf{0}\end{bmatrix}\{1+o_p(1)\}\right).
\end{aligned}$$

Meanwhile, with the condition $h_5^2/h_4 \to 0$ held, we obtain

$$\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{nh_4^2h_5}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i+\frac{1}{nh_4^3h_5}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{h_4}\right)K\left(\frac{U_i-u}{h_5}\right)\tilde{\boldsymbol{X}}_i\tilde{R}(\boldsymbol{X}_i,U_i)\right)\\
=&\frac{1}{nh_4^4h_5^2}\iiint\phi^{(1)2}\left(\frac{\epsilon}{h_4}\right)\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^Tf_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{h_5}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})(1+o_p(1))\\
=&\frac{\int\tau^2\phi^2(\tau)d\tau}{nh_4^3h_5}f_U(u)\begin{bmatrix}v_0\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & \mathbf{0}\\ \mathbf{0} & v_2\mathbb{E}(\boldsymbol{X}\boldsymbol{X}^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}(1+o_p(1)).
\end{aligned}$$

For the remaining part, we can follow the same idea in the Proof of Theorem 2.2 to easily obtain the result.

□

S5-7: Proof of Theorem 2.7

Following the result in Theorem 2.2, under the null hypothesis, we can prove

$$L(\mathcal{H}_0) = \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\boldsymbol{\alpha}^* - \boldsymbol{Z}_i^T\boldsymbol{\beta}^*}{h_4}\right) = \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \boldsymbol{X}_i^T(\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}^*) + \boldsymbol{Z}_i^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta}^*))}{h_4}\right)$$
$$= \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i}{h_4}\right) + o_p(1).$$

Similarly, by Theorem 2.6, we can obtain

$$L(\mathcal{H}_1) = \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{Y_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}}{h_4}\right) = \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)}{h_4}\right)$$
$$= \frac{1}{h_4}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i}{h_4}\right) + o_p(1),$$

where $\tilde{R}(\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i) = \boldsymbol{X}_i^T\boldsymbol{\alpha}_0(U_i) - \boldsymbol{X}_i^T(\boldsymbol{\alpha}_0(u) + \boldsymbol{b}_0(u)(U_i - u)) - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \boldsymbol{X}_i^T[\tilde{\boldsymbol{\alpha}}(u) - \boldsymbol{\alpha}_0(u) + (\tilde{\boldsymbol{b}}(u) - \boldsymbol{b}_0(u))(U_i - u)]$. Thus, we have $L(\mathcal{H}_1) - L(\mathcal{H}_0) \to 0$. Following the similar steps, we could show $T_0 \overset{\text{def}}{=} L(\mathcal{H}_1) - L(\mathcal{H}_0) > 0$ if $\inf_{\alpha_l \in R}\|\alpha_l(.) - \alpha_l\| > 0$.

□

S5-8: Proof of Theorem 2.8

Notice that, under $\mathcal{H}_0$, we have

$$T_0 \overset{\text{def}}{=} L(\mathcal{H}_1) - L(\mathcal{H}_0) = \sum_{i=1}^{n}\phi_{h_4}\left(Y_i - \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}\right) - \sum_{i=1}^{n}\phi_{h_4}\left(Y_i - \boldsymbol{X}_i^T\boldsymbol{\alpha}^* - \boldsymbol{Z}_i^T\boldsymbol{\beta}^*\right)$$
$$= \left\{\sum_{i=1}^{n}\phi_{h_4}\left(\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0 - \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) + \boldsymbol{Z}_i^T\boldsymbol{\beta}_0 - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}\right) - \sum_{i=1}^{n}\phi_{h_4}\left(\epsilon_i\right)\right\}$$
$$- \left\{\sum_{i=1}^{n}\phi_{h_4}\left(\epsilon_i + \boldsymbol{X}_i^T\boldsymbol{\alpha}_0 - \boldsymbol{X}_i^T\boldsymbol{\alpha}^* + \boldsymbol{Z}_i^T\boldsymbol{\beta}_0 - \boldsymbol{Z}_i^T\boldsymbol{\beta}^*\right) - \sum_{i=1}^{n}\phi_{h_4}\left(\epsilon_i\right)\right\}$$
$$= Z_1 - Z_2.$$

By Taylor expansion, under the null hypothesis, we can show that Z_2 follows $\chi^2(p)$ asymptotically. Therefore, $Z_2 = O_p(1)$. We then mainly focus on the asymptotic distribution of Z_1.

Considering Z_1, by Taylor expansion, we have

$$Z_1 = -\sum_{i=1}^{n}\phi_{h_4}^{(1)}\left(\epsilon_i\right)\left(\boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{X}_i^T\boldsymbol{\alpha}_0\right) - \sum_{i=1}^{n}\phi_{h_4}^{(1)}\left(\epsilon_i\right)\left(\boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}} - \boldsymbol{Z}_i^T\boldsymbol{\beta}_0\right)$$
$$+\frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}\left(\epsilon_i\right)\left[\boldsymbol{X}_i^T\boldsymbol{\alpha}_0 - \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i) + \boldsymbol{Z}_i^T\boldsymbol{\beta}_0 - \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}\right]^2 + o_p(1)$$
$$= Z_{11} + Z_{12} + o_p(1).$$

Based on Theorems 2.4 and 2.6, we know that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(h_3^2 + (nh_3^3)^{-1/2})$ and

$$\begin{aligned}
&\tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{\alpha}_0 \\
&= \left\{ \frac{h_5^2}{2} \int u^2 K(u) du \boldsymbol{\alpha}_0^{(2)}(U_i) + \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X}, U, \boldsymbol{Z})} \sum_{j=1}^n \boldsymbol{X}_j K_{h_5}(U_j - U_i) \phi_{h_4}^{(1)}(\epsilon_j) \right\} (1 + o_p(1)) \\
&= (R_1(U_i) + R_2(U_i))(1 + o_p(1))
\end{aligned}$$

according to the Bahadur representation of the estimator. We can then re-write

$$\begin{aligned}
Z_{11} &= -\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) (\boldsymbol{X}_i^T \tilde{\boldsymbol{\alpha}}(U_i) - \boldsymbol{X}_i^T \boldsymbol{\alpha}_0) + O_p(nh_3^2 h_4^2) \\
&= -\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T (R_1(U_i) + R_2(U_i))(1 + o_p(1)) \\
&= -\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T R_1(U_i)(1 + o_p(1)) - \sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T R_2(U_i)(1 + o_p(1)) \\
&= (Z_{11,1} + Z_{11,2})(1 + o_p(1)).
\end{aligned}$$

According to the result that $S_n = \mathbb{E}(S_n) + O_p(\sqrt{Var(S_n)})$ and the Strong Law of Large Number theory, by directly calculating, we can have

$$\mathbb{E}\Big(-\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T R_1(U_i) \Big) = \frac{3nh_5^2 h_4^2}{2} f_\epsilon(0|\boldsymbol{X}, U, \boldsymbol{Z}) \boldsymbol{X}_i^T \int u^2 K(u) du \boldsymbol{\alpha}_0^{(2)}(U_i).$$

Thus, $Z_{11,1} = O_p(nh_5^2 h_4^2) + O_p(\sqrt{n} h_5^2 h_4^{-3/2})$. As to $Z_{11,2}$, note that

$$\begin{aligned}
Z_{11,2} &= -\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T R_2(U_i) \\
&= -\sum_{i=1}^n \phi_{h_4}^{(1)}(\epsilon_i) \boldsymbol{X}_i^T \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X}, U, \boldsymbol{Z})} \sum_{j=1}^n \boldsymbol{X}_j K_{h_5}(U_j - U_i) \phi_{h_4}^{(1)}(\epsilon_j) \\
&= -\frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X}, U)} \sum_{i=1}^n (\phi_{h_4}^{(1)}(\epsilon_i))^2 \boldsymbol{X}_i^T \boldsymbol{X}_i K_{h_5}(0) \\
&\quad - \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X}, U)} \sum_{i=1}^n \sum_{j \neq i}^n \phi_{h_4}^{(1)}(\epsilon_i) \phi_{h_4}^{(1)}(\epsilon_j) \boldsymbol{X}_i^T \boldsymbol{X}_j K_{h_5}(U_j - U_i) = Z_{11,21} + Z_{11,22}.
\end{aligned}$$

We can obtain $\mathbb{E}(Z_{11,21}) = -\frac{f_U^{-1}(U)}{h_4^3 h_5 f_\epsilon^{(2)}(0|\boldsymbol{X}, U, \boldsymbol{Z})} f_\epsilon(0|\boldsymbol{X}, U, \boldsymbol{Z}) \int t^2 \phi^2(t) dt \boldsymbol{X}_i^T \boldsymbol{X}_i K(0)$. Thus, we have

$$Z_{11,2} = -\frac{f_U^{-1}(U)}{h_4^3 h_5 f_\epsilon^{(2)}(0|\boldsymbol{X}, U, \boldsymbol{Z})} f_\epsilon(0|\boldsymbol{X}, U, \boldsymbol{Z}) \int t^2 \phi^2(t) dt \boldsymbol{X}_i^T \boldsymbol{X}_i K(0)$$

$$- \frac{f_U^{-1}(U)}{nf_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)\,\boldsymbol{X}_i^T\boldsymbol{X}_jK_{h_5}(U_j-U_i)+O_p(n^{-1/2}h_4^{-7/2}h_5^{-1}).$$

Combining the above equations, we get

$$\begin{aligned}Z_{11} =& -\frac{f_U^{-1}(U)}{h_4^3h_5f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})\int t^2\phi^2(t)dt\boldsymbol{X}_i^T\boldsymbol{X}_iK(0)\\ &-\frac{f_U^{-1}(U)}{nf_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)\,\boldsymbol{X}_i^T\boldsymbol{X}_jK_{h_5}(U_j-U_i)\\ &+O_p(nh_5^2h_4^2)+O_p(\sqrt{n}h_5^2h_4^{-3/2})+O_p(n^{-1/2}h_4^{-7/2}h_5^{-1}).\end{aligned}$$

We then consider Z_{12} and obtain

$$\begin{aligned}Z_{12} &= \frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\left[\boldsymbol{X}_i^T\boldsymbol{\alpha}_0-\boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i)+\boldsymbol{Z}_i^T\boldsymbol{\beta}_0-\boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}}\right]^2+o_p(1)\\ &= \frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\tilde{\boldsymbol{\alpha}}(U_i)-\boldsymbol{\alpha}_0)^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\tilde{\boldsymbol{\alpha}}(U_i)-\boldsymbol{\alpha}_0)+\frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\boldsymbol{Z}_i\boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\ &+\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\boldsymbol{X}_i^T\boldsymbol{\alpha}_0-\boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}}(U_i))(\boldsymbol{Z}_i^T\boldsymbol{\beta}_0-\boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}})+o_p(1).\end{aligned}$$

Due to the faster convergence rate of $\tilde{\boldsymbol{\beta}}$, we can re-write Z_{12} as

$$\begin{aligned}Z_{12} &= \frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\tilde{\boldsymbol{\alpha}}(U_i)-\boldsymbol{\alpha}_0)^T\boldsymbol{X}_i\boldsymbol{X}_i^T(\tilde{\boldsymbol{\alpha}}(U_i)-\boldsymbol{\alpha}_0)+o_p(1)\\ &= \frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,R_1^T(U_i)\boldsymbol{X}_i\boldsymbol{X}_i^TR_1(U_i)+\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,R_1^T(U_i)\boldsymbol{X}_i\boldsymbol{X}_i^TR_2(U_i)\\ &+\frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,R_2^T(U_i)\boldsymbol{X}_i\boldsymbol{X}_i^TR_2(U_i)+o_p(1)=Z_{12,1}+Z_{12,2}+Z_{12,3}+o_p(1).\end{aligned}$$

Using the same procedure as that used for $Z_{11,1}$ and $Z_{11,2}$, respectively, we obtain

$$\mathbb{E}\Big(\frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,R_1^T(U_i)\boldsymbol{X}_i\boldsymbol{X}_i^TR_1(U_i)\Big)=\frac{nh_5^4h_4^{-2}}{8}f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})\Big(\int u^2K(u)du\boldsymbol{X}_i^T\boldsymbol{\alpha}_0^{(2)}(U_i)\Big)^2.$$

Thus, $Z_{12,1}=O_p(nh_5^4h_4^{-2})+O_p(\sqrt{n}h_5^4h_4^{-5/2})$. By directly calculating, we can have

$$\begin{aligned}Z_{12,2} &= \frac{h_5^2}{2n}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\boldsymbol{\alpha}_0^{(2)}(U_i))^T\boldsymbol{X}_i\boldsymbol{X}_i^T\\ &\int u^2K(u)du\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\sum_{j=1}^{n}\boldsymbol{X}_jK_{h_5}(U_j-U_i)\phi_{h_4}^{(1)}(\epsilon_j)\end{aligned}$$

$$= \frac{h_5^2}{2n}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_i)(\boldsymbol{\alpha}_0^{(2)}(U_i))^T\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_i\int u^2K(u)du\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(0)$$
$$+ \frac{h_5^2}{2n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)(\boldsymbol{\alpha}_0^{(2)}(U_i))^T\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_j\int u^2K(u)du\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(U_j-U_i)$$
$$= O_p(h_5h_4^{-1}) + O_p(n^{-1/2}h_5h_4^{-9/2}) + O_p(n^{-1/2}h_5^{3/2}h_4^{-4}).$$

In terms of $Z_{12,3}$, it can be decomposed into three parts

$$Z_{12,3} = \frac{1}{2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\frac{f_U^{-1}(U)}{nf_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\sum_{j=1}^{n}\boldsymbol{X}_j^TK_{h_5}(U_j-U_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_i\boldsymbol{X}_i^T$$
$$\frac{f_U^{-1}(U)}{nf_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\sum_{j=1}^{n}\boldsymbol{X}_jK_{h_5}(U_j-U_i)\phi_{h_4}^{(1)}(\epsilon_j)$$
$$= \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\phi_{h_4}^{(1)}(\epsilon_j))^2\boldsymbol{X}_j^T\left[\frac{f_U^{-2}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^TK_{h_5}^2(U_j-U_i)\right]\boldsymbol{X}_j$$
$$+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T$$
$$\left[\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\boldsymbol{X}_i\boldsymbol{X}_i^T\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(U_j-U_i)K_{h_5}(0)\right]\boldsymbol{X}_i$$
$$+ \frac{1}{2n^2}\sum_{i\neq j,j\neq t,t\neq i}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_t)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T$$
$$\left[\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\boldsymbol{X}_i\boldsymbol{X}_i^T\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(U_t-U_i)K_{h_5}(U_j-U_i)\right]\boldsymbol{X}_t$$
$$= Z_{12,31} + Z_{12,32} + Z_{12,33}.$$

$Z_{12,31}$ can be rewritten as

$$Z_{12,31} = \frac{1}{2n^2}\sum_{i=1}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\phi_{h_4}^{(1)}(\epsilon_i))^2\boldsymbol{X}_i^T\left[\frac{f_U^{-2}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^TK_{h_5}^2(0)\right]\boldsymbol{X}_i$$
$$+ \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\phi_{h_4}^{(2)}(\epsilon_i)\,(\phi_{h_4}^{(1)}(\epsilon_j))^2\boldsymbol{X}_j^T\left[\frac{f_U^{-2}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^TK_{h_5}^2(U_j-U_i)\right]\boldsymbol{X}_j + o_p(1),$$

where we can obtain

$$Z_{12,31} = \frac{\int\phi^2(t)t^2dt}{2nh_4^6h_5^2}\frac{f_U^{-2}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})\int K^2(t)dt\boldsymbol{X}_i^T\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_iK^2(0)$$
$$+ O_p((n^3h_5^4h_4^{13})^{-1/2}) + O_p((n^3h_5^3h_4^{12})^{-1/2}) = O_p((nh_4^6h_5^2)^{-1}).$$

It is obvious that $\mathbb{E}(Z_{12,32}) = \mathbb{E}(Z_{12,33}) = 0$. Thus, $Z_{12,32} = O_p((n^3h_5^3h_4^{12})^{-1/2})$.

We can re-write $Z_{12,33}$ as

$$Z_{12,33} = \frac{1}{2n}\sum_{t,j=1,t\neq j}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\frac{1}{n}\sum_{i=1,i\neq j,i\neq t}\phi_{h_4}^{(2)}(\epsilon_i)$$
$$\boldsymbol{X}_j^T\left[\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\boldsymbol{X}_i\boldsymbol{X}_i^T\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(U_t-U_i)K_{h_5}(U_j-U_i)\right]\boldsymbol{X}_t,$$

where

$$\frac{1}{n}\sum_{i=1,i\neq j,i\neq t}\phi_{h_4}^{(2)}(\epsilon_i)\,\boldsymbol{X}_j^T\left[\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}\boldsymbol{X}_i\boldsymbol{X}_i^T\frac{f_U^{-1}(U)}{f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}K_{h_5}(U_t-U_i)K_{h_5}(U_j-U_i)\right]\boldsymbol{X}_t$$
$$=\frac{1}{h_4^2h_5}\boldsymbol{X}_j^T\frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_tf_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt$$
$$+O_p((nh_5^3h_4^5)^{-1/2}).$$

We then have

$$Z_{12,33} = \frac{1}{2n}\frac{1}{h_4^2h_5}\sum_{t,j=1,j\neq t}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T\frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_tf_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})$$
$$\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt+O_p((nh_5^3h_4^5)^{-1/2}).$$

Combining the above equations, we obtain

$$Z_{12,3} = \frac{1}{2n}\frac{1}{h_4^2h_5}\sum_{t,j=1,j\neq t}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T\frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_tf_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})$$
$$\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt+O_p((nh_5^3h_4^5)^{-1/2}).$$

Furthermore, we get

$$Z_{12} = \frac{1}{2n}\frac{1}{h_4^2h_5}\sum_{t,j=1,j\neq t}^{n}\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T\frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_tf_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})$$
$$\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt+O_p((nh_5^3h_4^5)^{-1/2})+O_p((n^3h_5^3h_4^{12})^{-1/2})+O_p((nh_4^6h_5^2)^{-1}).$$

Combining the above equations, with the bandwidth conditions imposed in the paper satisfied, we have

$$Z_1 = -\frac{f_U^{-1}(U)}{h_4^3h_5f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})}f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z})\int t^2\phi^2(t)dt\boldsymbol{X}_i^T\boldsymbol{X}_iK(0)$$

$$
\begin{aligned}
&- \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})} \sum_{i=1}^{n}\sum_{j\neq i}^{n} \phi_{h_4}^{(1)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)\,\boldsymbol{X}_i^T\boldsymbol{X}_j K_{h_5}(U_j-U_i) \\
&+ \frac{1}{2n}\frac{1}{h_4^2 h_5} \sum_{t,j=1,j\neq t}^{n} \phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T \frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_t f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z}) \\
&\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt + O_p(nh_5^4h_4^{-2}) + O_p(\sqrt{n}h_5^2h_4^{-3/2}).
\end{aligned}
$$

Based on the above calculations, we get

$$T_0 = \mu_n + W_n + d_n,$$

where $d_n = O_p(nh_5^4h_4^{-2}) + O_p(\sqrt{n}h_5^2h_4^{-3/2})$,

$$\mu_n = -\frac{f_U^{-1}(U)}{h_4^3 h_5 f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})} f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z}) \int t^2\phi^2(t)dt\boldsymbol{X}_i^T\boldsymbol{X}_i K(0), \text{ and}$$

$$
\begin{aligned}
W_n = &- \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})} \sum_{i=1}^{n}\sum_{j\neq i}^{n} \phi_{h_4}^{(1)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)\,\boldsymbol{X}_i^T\boldsymbol{X}_j K_{h_5}(U_j-U_i) \\
&+ \frac{1}{2n}\frac{1}{h_4^2 h_5} \sum_{t,j=1,j\neq t}^{n} \phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j)\boldsymbol{X}_j^T \frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_t f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z}) \\
&\int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt.
\end{aligned}
$$

It is noticed that $\mathbb{E}(W_n) = 0$. Let $W_n = \sum_{i,j=1,i\neq j}^{n} \omega_{i,j}$, where

$$
\begin{aligned}
\omega(i,j) = &- \frac{f_U^{-1}(U)}{n f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z})} \phi_{h_4}^{(1)}(\epsilon_i)\,\phi_{h_4}^{(1)}(\epsilon_j)\,\boldsymbol{X}_i^T\boldsymbol{X}_j K_{h_5}(U_j-U_i) + \frac{1}{2n}\frac{1}{h_4^2 h_5}\phi_{h_4}^{(1)}(\epsilon_i)\phi_{h_4}^{(1)}(\epsilon_j) \\
&\boldsymbol{X}_j^T \frac{f_U^{-1}(U)}{(f_\epsilon^{(2)}(0|\boldsymbol{X},U,\boldsymbol{Z}))^2}\boldsymbol{X}_i\boldsymbol{X}_i^T\boldsymbol{X}_t f_\epsilon(0|\boldsymbol{X},U,\boldsymbol{Z}) \int K(t)K\Big(t-\frac{U_t-U_j}{h_5}\Big)dt.
\end{aligned}
$$

As $\{\epsilon_i\}_{i=1}^n$ are independent with $\mathbb{E}(\epsilon_i) = 0$, we have

$$Var(W_n) = 2n(n-1)\mathbb{E}(\omega(1,2))^2.$$

Due to the complicated form, we here use $\sigma_n^2 = Var(W_n)$ to denote the variance of W_n. Next, we discuss the asymptotic distribution of W_n. Let $W_{i,j} = \omega(i,j) + \omega(j,i)$, we have

$$W_n = \sum_{i,j=1,i<j}^{n} \omega_{i,j}.$$

It is easy to show that W_n is clear (De Jong, 1987), and G_1, G_2, and G_4 are of lower than

$(Var(W_n))^2$, where

$$G_1 = \sum_{1\leqslant i<k\leqslant n} \mathbb{E}\left(\omega_{i,k}^4\right)$$
$$G_2 = \sum_{1\leqslant i<j<k\leqslant n} \mathbb{E}\left(\omega_{i,k}^2\omega_{i,j}^2\right) + E\left(\omega_{j,k}^2\omega_{j,i}^2\right) + \mathbb{E}\left(\omega_{k,i}^2\omega_{k,j}^2\right),$$
$$G_4 = \sum_{1\leqslant i<j<k<l\leqslant n} \mathbb{E}\left(\omega_{i,k}\omega_{i,j}\omega_{l,k}\omega_{l,j}\right) + \mathbb{E}\left(\omega_{i,j}\omega_{i,l}\omega_{k,j}\omega_{k,l}\right) + \mathbb{E}\left(\omega_{i,k}\omega_{i,l}\omega_{j,k}\omega_{j,l}\right).$$

By Proposition 3.2 of De Jong (1987), it can be shown that $\sigma_n^{-1}W_n \xrightarrow{d} N(0,1)$. This implies that

$$\sigma_n^{-1}(T_0 - \mu_n + d_n) \xrightarrow{d} N(0,1).$$

□

S5-9: Proof of Theorem 2.9

Let $\mathcal{L}_0^*$ and $\mathcal{L}_1^*$ be defined similarly as $\mathcal{L}_0$ and $\mathcal{L}_1$ based on a bootstrap sample $\{(Y_i^*, \boldsymbol{X}_i, U_i, \boldsymbol{Z}_i))\}_{i=1}^n$. We use the superscript $*$ of a quantity as its bootstrap analogue. Then,

$$T_0^* = \mathcal{L}_1^* - \mathcal{L}_0^*.$$

The proof mainly consists of the two steps. (1) Noting $Y_i^* = \boldsymbol{X}_i^T\tilde{\boldsymbol{\alpha}} + \boldsymbol{Z}_i^T\tilde{\boldsymbol{\beta}} + \tilde{\epsilon}_i^*$ and bandwidths satisfy the corresponding restrictions and using the same arguments as the Proof of Theorem 2.8, it follows that

$$\mathcal{L}_1 - \mathcal{L}_0 = \mu_n + W_n^* + d_n,$$

where W_n^* is defined similarly as W_n but with ϵ_i replaced by $\tilde{\epsilon}_i^*$. (2) We further use the arguments similar to that given in Theorem 2.8 to obtain that

$$\sigma_n^{-1}W_n^* \xrightarrow{d} N(0,1),$$

which completes the proof.

□

S5-10: Proof of Theorem 3.1

Following the steps to prove Theorem 2.3, we define $\delta_n = h_3^2 + \sqrt{(nh_3^3)^{-1}} + a_n$. Then, it is sufficient to show that for any given η, there exists a large number constant c such that

$$P\left\{\sup_{\|\boldsymbol{\mu}\|=c} \mathcal{L}_P(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}) < \mathcal{L}_P(\boldsymbol{\beta}_0)\right\} \geq 1 - \eta,$$

where $\boldsymbol{\mu}$ is a $k \times 1$ dimension vector. The above equation implies that with probability at least $1-\delta$, there exits a local maximum in the ball $\{\boldsymbol{\beta}_0 + \delta_n \boldsymbol{\mu} : \|\boldsymbol{\mu}\| \leq c\}$. Using $p_\lambda(0) = 0$ and Taylor expansion, it follows that

$$\begin{aligned}
\frac{1}{n}\left(\mathcal{L}_P(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}) - \mathcal{L}_P(\boldsymbol{\beta}_0)\right) =& Q_n(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}) - Q_n(\boldsymbol{\beta}_0) - \sum_{j=1}^{s}[p_{\lambda_j}\left(|\beta_{j0} + \delta_n\mu_j|\right) - p_{\lambda_j}\left(|\beta_{j0}|\right)] \\
=& \delta_n Q_n^{(1)}(\boldsymbol{\beta}_0)^T\boldsymbol{\mu} + \frac{1}{2}\delta_n^2\mu^T Q_n^{(2)}(\boldsymbol{\beta}_0)^T\mu + \frac{1}{6}\delta_n^3\mu^T Q_n^{(3)}(\boldsymbol{\beta}_0^*)^T\mu^T\mu \\
& - \sum_{j=1}^{s}\left[\delta_n p_\lambda^{(1)}\left(|\beta_{j0}|\right)\operatorname{sgn}\left(\beta_{j0}\right)\mu_j + \delta_n^2 p_\lambda^{(2)}(|\beta_{j0}|)\mu_j^2\left\{1 + o_p(1)\right\}\right] \\
=& M_1 + M_2 + M_3 + M_4,
\end{aligned}$$

where $\|\boldsymbol{\beta}_0^* - \boldsymbol{\beta}_0\| \leq c\delta_n$. From the Proof of Theorem 2.3, we know $M_1 = O_p(\delta_n^2 c)$, $M_2 = O_p(\delta_n^2 c^2)$, and $M_3 = O_p(\delta_n^3)$. By choosing bigger enough c, M_2 could domain M_1 and M_3 with probability $1-\eta$. Note that M_4 is bounded by

$$\sqrt{s}\delta_n \max\left\{p_\lambda^{(1)}\left(|\beta_{j0}|\right) : \beta_{j0} \neq 0\right\}\|\boldsymbol{\mu}\| + \delta_n^2 \max\left\{p_\lambda^{(2)}\left(|\beta_{j0}|\right) : \beta_{j0} \neq 0\right\}\|\boldsymbol{\mu}\|^2,$$

which is also dominated by M_2 as $\max\left\{p_\lambda^{(2)}\left(|\beta_{j0}|\right) : \beta_{j0} \neq 0\right\} \to 0$. Because $Q_n^{(2)}(\boldsymbol{\beta}_0) < 0$, we have $\mathcal{L}_P(\boldsymbol{\beta}_0 + \delta_n\boldsymbol{\mu}) < \mathcal{L}_P(\boldsymbol{\beta}_0)$ with probability $1-\eta$ for $\eta > 0$ by choosing a sufficiently large c.

□

S5-11: Proof of Theorem 3.2

By the property of SCAD penalty function, as $\lambda_{max} \to 0$, it can be shown that $a_n = 0$ for large n. Then, according to Theorem 3.1, it is sufficient to show that for any $\boldsymbol{\beta}^P$ that satisfies $\|\boldsymbol{\beta}^P - \boldsymbol{\beta}_0\| = O_p(\delta_n)$ and for some small $\epsilon = c\delta_n$ in which $\delta_n = h_3^2 + \sqrt{(nh_3^3)^{-1}}$, when $n \to \infty$, with probability tending to one, we have

$$\begin{aligned}
\frac{\partial\mathcal{L}_P(\boldsymbol{\beta})}{\partial\beta_j^P} < 0, \quad & \text{for } 0 < \beta_j^P < \epsilon,\ j = s+1, \ldots, k, \\
\frac{\partial\mathcal{L}_P(\boldsymbol{\beta})}{\partial\beta_j^P} > 0, \quad & \text{for } -\epsilon < \beta_j^P < 0,\ j = s+1, \ldots, k,
\end{aligned}$$

which indicates that the maximizer of $\mathcal{L}_P(\boldsymbol{\beta})$ gets at $\beta_j^P = 0, j = s+1, \ldots, k$. Similar to the proof of Theorem 3.1, as $Q_n^{(1)}(\boldsymbol{\beta}_0) = O_p(\delta_n)$ and $\|\boldsymbol{\beta}^P - \boldsymbol{\beta}_0\| = O_p(\delta_n)$, we obtain

$$\begin{aligned}
\frac{\partial\mathcal{L}_P(\boldsymbol{\beta})}{\partial\beta_j^P} =& nQ_n^{(1)}(\boldsymbol{\beta}) - np_\lambda^{(1)}\left(|\beta_j^P|\right)\operatorname{sgn}\beta_j^P \\
=& nQ_n^{(1)}(\boldsymbol{\beta}_0) + nQ_n^{(2)}(\boldsymbol{\beta}_0)(\beta_{0j} - \beta_j^P) + \frac{n}{2}Q_n^{(3)}(\boldsymbol{\beta}_0^*)(\beta_{0j} - \beta_j^P)^2 - np_\lambda^{(1)}\left(|\beta_j^P|\right)\operatorname{sgn}\beta_j^P
\end{aligned}$$

$$= - n\lambda \left\{ \lambda^{-1} p_\lambda^{(1)} \left(|\beta_j^P| \right) \operatorname{sgn} \beta_j^P + O_p \left(\delta_n / \lambda \right) \right\},$$

where $\boldsymbol{\beta}_0^*$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. As $\delta_n^{-1}\lambda \geq \delta_n^{-1}\lambda_{min} \to \infty$ when $n \to \infty$ and $\liminf_{n\to 0} \liminf_{\beta_j^P \to 0} p_\lambda^{(1)}(|\beta_j^P|)/\lambda > 0$, the sign of the derivation is completely determined by that of β_j^P. Then, the above two equations hold. This completes the proof.

□

S5-12: Proof of Theorem 3.3

From the Proof of Theorem 3.2, we know that for $j = 1, ..., s$, we have

$$\begin{aligned}
\frac{1}{n} \frac{\partial \mathcal{L}_P(\boldsymbol{\beta})}{\partial \beta_j} \Big|_{\boldsymbol{\beta}=((\hat{\boldsymbol{\beta}}_{0'}^P)^T, 0)^T} =& Q_n^{(1)}(\hat{\boldsymbol{\beta}}_{0'}^P) - p_\lambda^{(1)} \left(|\hat{\beta}_{0'j}^P| \right) \operatorname{sgn} \hat{\beta}_{0'j}^P \\
=& Q_n^{(1)}(\boldsymbol{\beta}_{0'}) + Q_n^{(2)}(\boldsymbol{\beta}_{0'})(\beta_{0'j} - \hat{\beta}_{0'j}^P) + \frac{1}{2} Q_n^{(3)}(\boldsymbol{\beta}_0^*)(\beta_{0'j} - \hat{\beta}_{0'j}^P)^2 \\
& - \left\{ p_\lambda^{(1)} \left(|\beta_{0'j}| \right) \operatorname{sgn} \beta_{0'j} + (p_\lambda^{(2)} \left(|\beta_{0'j}| \right) + o_p(1))(\hat{\beta}_{0'j}^P - \beta_{0'j}) \right\}.
\end{aligned}$$

Combining these equations, we have

$$\begin{aligned}
& Q_n^{(1)}(\boldsymbol{\beta}_{0'}) + Q_n^{(2)}(\boldsymbol{\beta}_{0'})(\beta_{0'j} - \hat{\beta}_{0'j}^P) + \frac{1}{2} Q_n^{(3)}(\boldsymbol{\beta}_0^*)(\beta_{0'j} - \hat{\beta}_{0'j}^P)^2 \\
& - \left\{ \Psi_\lambda + (\Phi_\lambda + o_p(1))(\hat{\beta}_{0'j}^P - \beta_{0'j}) \right\} = 0.
\end{aligned}$$

From Theorem 3.1, following by Slutskys theorem and the central limit theorem, we know

$$\frac{h_3^2}{2} M_{(1)} - J_{(1)}(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'}) - \left\{ \Psi_\lambda + (\Phi_\lambda + o_p(1))(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'}) \right\} = 0,$$

$$\sqrt{nh_3^3}(J_{(1)} + \Phi_\lambda) \left(\hat{\boldsymbol{\beta}}_{0'}^P - \boldsymbol{\beta}_{0'} + (J_{(1)} + \Phi_\lambda)^{-1} \left(\Psi_\lambda - \frac{h_3^2}{2} M_{(1)} \right) \right) \xrightarrow{d} \mathcal{N} \left(0, \int t^2 \phi^2(t) dt L_{(1)} \right),$$

where $J_{(1)}, M_{(1)}$ and $L_{(1)}$ are the submatrices of J, M and L.

□

S5-13: Proof of Theorem S1

Notice that the notations in this proof are independent of the notations in other proofs. To start, we define $\boldsymbol{X}_i^{*T} = (X_{ip}, X_{ip}(U_i - u)/\lambda_2, X_{ip}(U_i - u)^2/\lambda_2^2, X_{ip}(U_i - u)^3/\lambda_2^3)$, $\boldsymbol{\theta} = (a_p(u), b_p(u), c_p(u), d_p(u))^T$, $\boldsymbol{\theta}_0 = (a_{0p}(u), b_{0p}(u), c_{0p}(u), d_{0p}(u))^T$, $H = \operatorname{diag}(\underbrace{1, \cdots, 1}_{p}, \underbrace{\lambda_2, \cdots, \lambda_2}_{p}, \underbrace{\lambda_2^2, \cdots, \lambda_2^2}_{p}, \underbrace{\lambda_2^3, \cdots, \lambda_2^3}_{p})$, $\tilde{\boldsymbol{\theta}}_1 = H\boldsymbol{\theta}$, and $\boldsymbol{\theta}_{10} = H\boldsymbol{\theta}_0$.

Let $\delta_n = \lambda_1^2 + \lambda_2^4 + \sqrt{(n\lambda_1^3\lambda_2)^{-1}}$, then it is sufficient to show that for any given η, there exists a large number constant c such that $P\left\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n\left(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}\right) < Q_n\left(\boldsymbol{\theta}_{10}\right)\right\} \geq 1-\eta$, where $\boldsymbol{\theta}_{10}$ is the true value of the parameter. Using Taylor expansion, it follows that

$$
\begin{aligned}
& Q_n\left(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}\right) - Q_n\left(\boldsymbol{\theta}_{10}\right) \\
=& \frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\left[\phi\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\right. \\
& \left. - \frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\phi\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\right] \\
=& \frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\left[-\phi^{(1)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\right. \\
& \quad + \frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)^2 K\left(\frac{U_i-u}{\lambda_2}\right) \\
& \quad \left. - \frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i^*}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)^3 K\left(\frac{U_i-u}{\lambda_2}\right)\right] \\
=& I_{10} + I_{11} + I_{12},
\end{aligned}
$$

where ϵ_i^* is between $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)$ and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*$ in which $\tilde{R}(\boldsymbol{X}_i, U_i) = \sum_{j=1}^{p-1}\alpha_{0j}(U_i)X_{ij} - \sum_{j=1}^{p-1}\tilde{\alpha}_j(U_i)X_{ij} - \boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + X_{ip}\alpha_{0p}(U_i) - \boldsymbol{X}_i^{*T}\boldsymbol{\theta}_{10}$. Based on the result $T_n = \mathbb{E}\left(T_n\right) + O_p(\sqrt{\mathrm{Var}\left(T_n\right)})$, we could consider each part of the above Taylor expansion.

(i) For the first part, which is $I_{10} = \frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^{n}\left(-\phi^{(1)}\left(\frac{\epsilon_i+\tilde{R}(\boldsymbol{X}_i,U_i)}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\right)$, we can re-write it as

$$
\begin{aligned}
\mathbb{E}(I_{10}) &= \frac{-\delta_n}{\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right)\right) \\
&= \frac{-\delta_n}{\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right) + \phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\tilde{R}(\boldsymbol{X}_i,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1^2}K\left(\frac{U_i-u}{\lambda_2}\right)\right. \\
&\qquad \left. + \frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i^{**}}{\lambda_1}\right)\frac{\tilde{R}^2(\boldsymbol{X}_i,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1^3}K\left(\frac{U_i-u}{\lambda_2}\right)\right) \\
&= I_{101} + I_{102} + I_{103},
\end{aligned}
$$

where ϵ_i^{**} is between ϵ_i and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i)$. As the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{103}, we instead use ϵ_i directly. By some calculations for each part, with the conditions $h_3/h_5 \to 0$, $h_4 = o(\lambda_2^2)$, and $h_5 = o(\lambda_2^2)$ held, we can get

$$
I_{101} = \frac{-\delta_n}{\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right)\right) = O_p(\delta_n c\lambda_1^2).
$$

$$
\begin{aligned}
I_{102} =& \frac{-\delta_n}{\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right)\frac{\tilde{R}(\boldsymbol{X}_i,U_i)}{\lambda_1}\right)\\
=& \frac{-\delta_n}{\lambda_1\lambda_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}^*}{\lambda_1}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{\lambda_2}\right)\frac{X_{ip}\alpha_{0p}(U_i)-\boldsymbol{X}_i^{*T}\boldsymbol{\theta}_{10}}{\lambda_1}f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
&+\frac{-\delta_n}{\lambda_1\lambda_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}^*}{\lambda_1}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K\left(\frac{U-u}{\lambda_2}\right)\\
&\frac{\sum_{j=1}^{p-1}\tilde{\alpha}_{0j}(U_i)X_{ij}-\sum_{j=1}^{p-1}\tilde{\alpha}_j(U_i)X_{ij}-\boldsymbol{Z}_i^T(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}{\lambda_1}f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=& O_p(\delta_n c\lambda_2^4)+O_p(h_4^2)+O_p(h_5^2)=O_p(\delta_n c\lambda_2^4).
\end{aligned}
$$

$$
I_{103} \approx \frac{-\delta_n}{\lambda_1\lambda_2}\mathbb{E}\left(\frac{1}{2}\phi^{(3)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\tilde{R}^2(\boldsymbol{X}_i,U_i)\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1^3}K\left(\frac{U_i-u}{\lambda_2}\right)\right)=o_p(\delta_n c\lambda_2^4).
$$

Meanwhile, combining the results from Theorem 2.1, with the condition $\lambda_2^4/h_6 \to 0$ held, we can obtain

$$
\frac{\delta_n^2}{\lambda_1^2\lambda_2^2}\mathbb{E}\left(\phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right)\right)^2=O_p(\delta_n^2c^2(\lambda_1^3\lambda_2)^{-1}).
$$

$$
\begin{aligned}
&\frac{\delta_n^2}{\lambda_1^2\lambda_2^2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}K\left(\frac{U_i-u}{\lambda_2}\right)\frac{\tilde{R}(\boldsymbol{X}_i,U_i)}{\lambda_1}\right)^2\\
=&\frac{\delta_n^2}{\lambda_1^2\lambda_2^2}\iiint \phi^{(2)2}\left(\frac{\epsilon}{\lambda_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}^*)^2}{\lambda_1^2}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{\lambda_2}\right)\frac{\tilde{R}^2(\boldsymbol{X},U)}{\lambda_1^2}f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})\\
=&o_p(\delta_n^2c^2(\lambda_1^3\lambda_2)^{-1}).
\end{aligned}
$$

These indicate $I_{10}=O_p(\delta_n c(\lambda_1^2+\lambda_2^4))+O_p(\sqrt{\delta_n^2c^2(n\lambda_1^3\lambda_2)^{-1}})=O_p(\delta_n^2c)$.

(ii) For the second part, $I_{11}=\frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^n\left(\frac{1}{2}\phi^{(2)}\left(\frac{\epsilon_i+\tilde{R}(\boldsymbol{X}_i,U_i)}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)^2K\left(\frac{U_i-u}{\lambda_2}\right)\right)$, we can re-write it as

$$
\begin{aligned}
\mathbb{E}(I_{11}) =& \frac{\delta_n^2}{2\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i+\tilde{R}(\boldsymbol{X}_i,U_i)}{\lambda_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^2}K\left(\frac{U_i-u}{\lambda_2}\right)\right)\\
=&\frac{\delta_n^2}{2\lambda_1\lambda_2}\mathbb{E}\Bigg(\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^2}K\left(\frac{U_i-u}{\lambda_2}\right)+\phi^{(3)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\tilde{R}(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^3}K\left(\frac{U_i-u}{\lambda_2}\right)\\
&+\frac{1}{2}\phi^{(4)}\left(\frac{\epsilon_i^{**}}{\lambda_1}\right)\frac{\tilde{R}^2(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^4}K\left(\frac{U_i-u}{\lambda_2}\right)\Bigg)\\
=&I_{111}+I_{112}+I_{113}.
\end{aligned}
$$

As the order of ϵ_i^{**} is the same as that of ϵ_i, when we do the calculations associated with I_{113},

we instead use ϵ_i directly. Combining the results obtained from Proof of Theorem 2.1 and assumption that $h_3/h_5 \to 0$, we can get

$$I_{111} = \frac{\delta_n^2}{2\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^2}K\left(\frac{U_i-u}{\lambda_2}\right)\right) = O_p((\delta_n c)^2).$$

$$I_{112} = \frac{\delta_n^2}{2\lambda_1\lambda_2}\mathbb{E}\left(\phi^{(3)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{\tilde{R}(\boldsymbol{X}_i,U_i)(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^3}K\left(\frac{U_i-u}{\lambda_2}\right)\right) = o_p((\delta_n c)^2).$$

Meanwhile, we can prove that $I_{113} = o_p((\delta_n c)^2)$ and obtain the following result

$$\frac{\delta_n^4}{4\lambda_1^2\lambda_2^2}\mathbb{E}\left(\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)\frac{(\boldsymbol{\mu}^T\boldsymbol{X}_i^*)^2}{\lambda_1^2}K\left(\frac{U_i-u}{\lambda_2}\right)\right)^2 = O_p((\delta_n c)^4(\lambda_2\lambda_1^5)^{-1}).$$

These 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 ϵ_i^* is between $\epsilon_i + \tilde{R}(\boldsymbol{X}_i,U_i)$ and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i,U_i) - \delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*$, the order of ϵ_i^* is the same as the order of ϵ_i, indicating that we can obtain

$$I_{12} \approx \frac{1}{n\lambda_1\lambda_2}\sum_{i=1}^n\left(-\frac{1}{6}\phi^{(3)}\left(\frac{\epsilon_i}{\lambda_1}\right)\left(\frac{\delta_n\boldsymbol{\mu}^T\boldsymbol{X}_i^*}{\lambda_1}\right)^3 K\left(\frac{U_i-u}{\lambda_2}\right)\right).$$

Combining the results obtained from the Proof of Theorem 2.1, it can be seen that the second part dominates the third part.

Based on these, we can choose c bigger enough such that the second term dominates the other two terms with probability $1-\eta$. Because the second term is negative, $P\{\sup_{\|\boldsymbol{\mu}\|=c} Q_n(\boldsymbol{\theta}_{10} + \delta_n\boldsymbol{\mu}) < Q_n(\boldsymbol{\theta}_{10})\} \geq 1-\eta$ holds. This completes the proof.

□

S5-14: Proof of Theorem S2

Following the same steps as proving Theorem S1, $\tilde{\boldsymbol{\theta}}_1$ must satisfy the following equation

$$-\frac{1}{n\lambda_1^2\lambda_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i + R(\boldsymbol{X}_i,U_i,\boldsymbol{Z}_i)}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^* = 0,$$

where $R(\boldsymbol{X}_i,U_i,\boldsymbol{Z}_i) = \tilde{R}(\boldsymbol{X}_i,U_i) - \boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$. By taking Taylor expansion, we can obtain

$$-\frac{1}{n\lambda_1^2\lambda_2}\sum_{i=1}^n\phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^* + \frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^n\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*(\tilde{R}(\boldsymbol{X}_i,U_i)$$
$$-\boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})) - \frac{1}{n\lambda_1^4\lambda_2}\sum_{i=1}^n\phi^{(3)}\left(\frac{\tilde{\epsilon}_i^*}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\left(\tilde{R}(\boldsymbol{X}_i,U_i) - \boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})\right)^2 = 0,$$

where $\tilde{\epsilon}_i^*$ is between ϵ_i and $\epsilon_i + \tilde{R}(\boldsymbol{X}_i, U_i) - \boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$. From Theorem S1, we know $\|\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\| = O_p(\delta_n)$, which indicates that

$$\begin{aligned}\sup_{i:|U_i-u|/\lambda_2\le 1} |R(\boldsymbol{X}_i, U_i, \boldsymbol{Z}_i)| &\le \sup_{i:|U_i-u|/\lambda_2\le 1} \{|\tilde{R}(\boldsymbol{X}_i, U_i)| + |\boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})|\} \\ &= O_p(\|\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})\|) = O_p(\delta_n).\end{aligned}$$

Thus, the third part which is associated with $(\tilde{R}(\boldsymbol{X}_i, U_i) - \boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}))^2$ is dominated by the second part which is associated with $\tilde{R}(\boldsymbol{X}_i, U_i) - \boldsymbol{X}_i^{*T}(\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10})$. We then mainly focus on the first two parts of the above equation.

Considering $-\frac{1}{n\lambda_1^2\lambda_2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^* + \frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\tilde{R}(\boldsymbol{X}_i, U_i)$, with the conditions $h_3/h_5 \to 0$, $h_4 = o(\lambda_2^2)$, and $h_5 = o(\lambda_2^2)$ held, we can obtain

$$\begin{aligned}&\mathbb{E}\left(-\frac{1}{n\lambda_1^2\lambda_2}\sum_{i=1}^n \phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\right. \\ &\left. + \frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\tilde{R}(\boldsymbol{X}_i, U_i)\right) \\ =&-\frac{1}{\lambda_1^2\lambda_2}\iiint \phi^{(1)}\left(\frac{\epsilon}{\lambda_1}\right)\boldsymbol{X}^* f_\epsilon(\epsilon|\hat{\boldsymbol{X}}) K\left(\frac{U-u}{\lambda_2}\right) f_U(U) dU d\epsilon dF(\hat{\boldsymbol{X}}) \\ &+\frac{1}{\lambda_1^3\lambda_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{\lambda_1}\right)\boldsymbol{X}^* f_\epsilon(\epsilon|\hat{\boldsymbol{X}}) K\left(\frac{U-u}{\lambda_2}\right)\tilde{R}(\boldsymbol{X}, U) f_U(U) dU d\epsilon dF(\hat{\boldsymbol{X}}) \\ =&\frac{\lambda_1^2}{2} f_U(u)\begin{bmatrix}\mathbb{E}(X f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u) \\ 0 \\ \mu_2\mathbb{E}(X f_\epsilon^{(3)}(0|\hat{\boldsymbol{X}})|u) \\ 0\end{bmatrix} - \left(\frac{\lambda_2^4\alpha_{0p}^{(4)}(u)}{24} f_U(u)\begin{bmatrix}\mu_4\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) \\ 0 \\ \mu_6\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) \\ 0\end{bmatrix}\right)\{1+o_p(1)\}.\end{aligned}$$

Considering $\frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T}$, by directly calculating, we have

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^n \phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right) K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\boldsymbol{X}_i^{*T}\right) \\ &= \frac{1}{\lambda_1^3\lambda_2}\iiint \phi^{(2)}\left(\frac{\epsilon}{\lambda_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T} f_\epsilon(\epsilon|\hat{\boldsymbol{X}}) K\left(\frac{U-u}{\lambda_2}\right) f_U(U) dU d\epsilon dF(\hat{\boldsymbol{X}}) = f_U(u) \\ &\begin{bmatrix}\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 & \mu_2\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 \\ 0 & \mu_2\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 & \mu_4\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) \\ \mu_2\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 & \mu_4\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 \\ 0 & \mu_4\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u) & 0 & \mu_6\mathbb{E}(XX^T f_\epsilon^{(2)}(0|\hat{\boldsymbol{X}})|u)\end{bmatrix}.\end{aligned}$$

Meanwhile, with the condition $\lambda_2^4/\lambda_1 \to 0$ held, we can obtain

$$
\begin{aligned}
&\mathrm{Var}\left(-\frac{1}{n\lambda_1^2\lambda_2}\sum_{i=1}^{n}\phi^{(1)}\left(\frac{\epsilon_i}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^* + \frac{1}{n\lambda_1^3\lambda_2}\sum_{i=1}^{n}\phi^{(2)}\left(\frac{\epsilon_i}{\lambda_1}\right)K\left(\frac{U_i-u}{\lambda_2}\right)\boldsymbol{X}_i^*\tilde{R}(\boldsymbol{X}_i,U_i)\right)\\
&=\frac{1}{n\lambda_1^4\lambda_2^2}\iiint\phi^{(1)2}\left(\frac{\epsilon}{\lambda_1}\right)\boldsymbol{X}^*\boldsymbol{X}^{*T}f_\epsilon(\epsilon|\hat{\boldsymbol{X}})K^2\left(\frac{U-u}{\lambda_2}\right)f_U(U)dUd\epsilon dF(\hat{\boldsymbol{X}})(1+o_p(1))\\
&=\frac{\int\tau^2\phi^2(\tau)d\tau}{n\lambda_1^3\lambda_2}f_U(u)\\
&\begin{bmatrix}
v_0\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0 & v_2\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0\\
0 & v_2\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0 & v_4\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)\\
v_2\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0 & v_4\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0\\
0 & v_4\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u) & 0 & v_6\mathbb{E}(XX^Tf_\epsilon(0|\hat{\boldsymbol{X}})|u)
\end{bmatrix}
\end{aligned}
$$

$(1+o_p(1))$.

For the remaining part, we can follow the same idea in the Proof of Theorem 2.2 to easily obtain the result.

□