Kernel Mode-Based Regression under Random Truncation

Tao Wang[a] Weixin Yao[b]

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

Supplementary Material

S1 Distribution of Data with Truncation

When truncation occurs to the left of the mode, regardless of the heaviness of truncation or the skewness of the data, the mode remains unaffected due to the mode's inherent focus on the most frequent or dense point in the distribution, rather than the tails or extremes. This robustness of the mode stems from its reliance on local maxima, which remain intact as long as the truncation does not eliminate the region around the mode. However, when truncation increases beyond the mode, the data available to estimate the mode shrink considerably, causing the mode to shift or disappear entirely. In such scenarios, the mode ceases to be a reliable measure for reconstructing the relationship between variables, as the available data no longer reflect the underlying distribution accurately (Figure S1).

Different from fixed (and therefore nonrandom) truncation, random

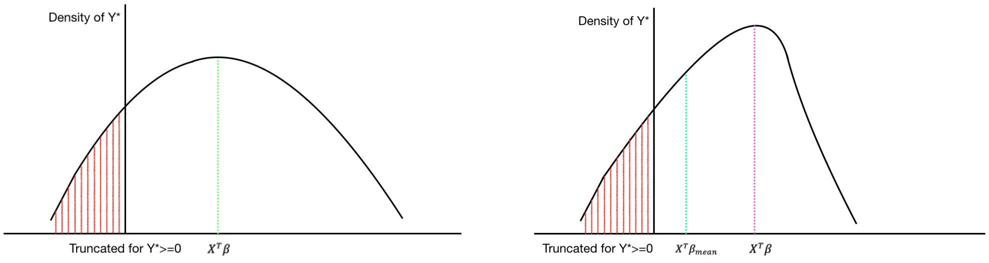

Figure S1: Distribution of Data with Truncation (Fixed Truncation)

truncation represents a biased sampling scheme, introducing a fundamental challenge in the estimation of regression models. In the context of random truncation, observations of the variables $(Y, \mathbf{X})$ are left-truncated by an independent random variable T, which acts as a censoring threshold. Specifically, it indicates that all three quantities of Y, $\mathbf{X}$, and T are observable only when $Y \geq T$. This truncation alters the distributional structure of the observed data, and consequently, the standard regression model $\mathcal{F}[Y \mid \mathbf{X}]$ that we might estimate from untruncated data is no longer directly applicable. The only regression function that can be estimated from the truncated sample is $\mathcal{F}[Y \mid \mathbf{X}, Y \geq T]$, where $\mathcal{F}$ represents a functional operator such as the mean, quantile, or mode. The key issue with random truncation is that the marginal distribution functions of both Y and T are inherently modified in the sample, compared to the original population data. As a result, the mode of the truncated sample $\mathcal{F}[Y \mid \mathbf{X}, Y \geq T]$ may no longer coincide with the mode of the original distribution $\mathcal{F}[Y \mid \mathbf{X}]$.

S2 Mechanism of Kernel Mode-Based Function

The kernel mode-based objective function serves as a generalization of the sample mode to the regression context, just as least squares generalizes the sample mean to the linear model. In the context of modal regression, the objective function in (2.2) is specifically designed to estimate the mode of the conditional distribution of the response variable, rather than the mean. To understand the underlying mechanism of (2.2), which reveals the "most likely" or mode effect of the predictor on the response, we define $g(\varepsilon)$ as the continuous density function of the error term ε, with the kernel function $K(w)$ satisfying $\int |w| K(w) dw < \infty$. The term $M(\varepsilon, K)$ represents the kernel-smoothed approximation of $g(\varepsilon)$, defined as

$$M(\varepsilon, K) = \int \frac{1}{h_0} K\left(\frac{\varepsilon - t}{h_0}\right) g(t) dt = \int K(w) g(\varepsilon + w h_0) dw,$$

where h_0 is the bandwidth parameter that controls the smoothness of the kernel estimator. We can establish the uniform convergence as follows

$$\begin{aligned}
&\sup_{\varepsilon \in \mathbb{R}} |g(\varepsilon) - M(\varepsilon, K)| = \sup_{\varepsilon \in \mathbb{R}} |g(\varepsilon) - \int K(w) g(\varepsilon + w h_0) dw| \\
\leq &\sup_{\varepsilon \in \mathbb{R}} \int |g(\varepsilon) - g(\varepsilon + w h_0)| K(w) dw \\
\leq &\sup_{\varepsilon \in \mathbb{R}} \int |g^{(1)}(\varepsilon) w h_0| K(w) dw = \sup_{\varepsilon \in \mathbb{R}} |g^{(1)}(\varepsilon)| h_0 \int |w| K(w) dw \to 0
\end{aligned}$$

with $h_0 \to 0$, where $g^{(1)}(\varepsilon)$ represents the first derivative of $g(\varepsilon)$. Consequently, $M(\varepsilon, K)$ can converge uniformly to $g(\varepsilon)$ as $h_0 \to 0$, implying that

$\arg\max_{\varepsilon} M(\varepsilon, K) \rightarrow \arg\max_{\varepsilon} g(\varepsilon)$. This indicates that the kernel objective function in (2.2) can be utilized to capture the modal estimator.

In the paper, we demonstrate that to achieve a robust and efficient estimator with $\sqrt{n}$-consistency (parametric convergence rate), the bandwidth h_0 is treated as a tuning parameter or constant that is independent of sample size. This is because, in the modal regression framework, the mode captures local characteristics of the distribution, and oversmoothing (too large h_0) can mask these local features, while undersmoothing (too small h_0) can lead to excessive variance in the estimator. By treating h_0 as fixed, we ensure that the estimator remains stable and achieves optimal convergence properties, making the kernel mode-based objective function in (2.2) both efficient and computationally tractable.

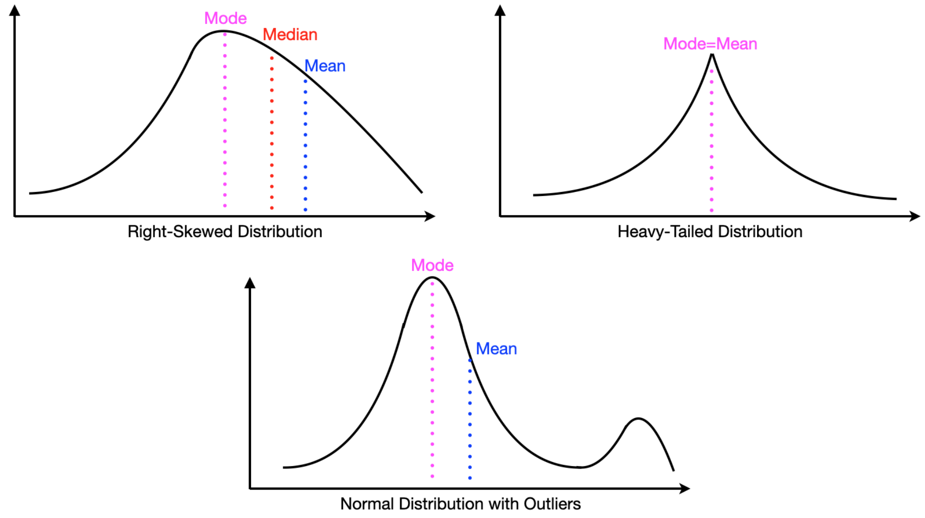
Mode
Median
Mean
Right-Skewed Distribution
Mode=Mean
Heavy-Tailed Distribution
Mode
Mean
Normal Distribution with Outliers

Figure S2: Mode in Different Cases

In addition, it has been discovered that when the data are asymmetrically distributed, the mode can provide an important complement to the mean by allowing the bandwidth $h_0 \to 0$ as sample size $n \to \infty$ (Figure S2-the first plot). In such cases, the target of regression—representing the "most likely" or mode value—diverges from the target of mean estimation, which seeks to estimate the "average" value. The distinction is particularly crucial in skewed distributions, where the mean can be heavily influenced by extreme values, while the mode remains a more stable measure of central tendency. For symmetrically distributed data, the mode and mean coincide, resulting in $Mode(Y) = \mathbb{E}(Y)$ (Figure S2-the second and third plots). In these cases, treating the bandwidth h_0 as a constant or tuning parameter in the kernel objective function allows for the accurate recovery of both the mean and the mode. Theoretically, this reflects the fact that in symmetric distributions, the most likely value is equal to the average value, and thus modal regression and mean regression converge to the same estimate. In our research, we term the resulting estimators in the asymmetric case as "modal estimators" due to their focus on estimating the mode, as demonstrated in Yao and Li (2014) and Ullah et al. (2023). For other cases, where the primary goal is to derive robust and efficient estimators by focusing on the mode, we refer to this class of techniques as "mode-based" estimation; see Yao et al. (2012) and Wang (2024). Similar model and method have been developed in Wang and Li (2021). This framework is particularly useful when the underlying data distribution may not adhere to strong parametric assumptions, and robustness is a key concern.

In summary, modal regression typically involves nonparametric methods where the bandwidth parameter shrinks to zero as the sample size increases. It aims to estimate the conditional mode of the response variable given the predictors, often using techniques such as kernel density estimation. It offers a useful alternative to mean or quantile regression when the most frequent or likely value is of primary interest. This method is sensitive to the choice of bandwidth and is particularly suited for capturing complex, nonlinear relationships in the data. In contrast, mode-based regression, as used in our study, refers to robust estimation methods that employ the mode as a measure of central tendency to mitigate the influence of outliers and heavy-tailed distributions. While modal regression specifically seeks to estimate the mode of the conditional distribution, mode-based regression is more concerned with ensuring robust and efficient estimators by focusing on the mode, rather than the mean or median, as the primary target of estimation. It uses a constant bandwidth (instead of a shrinkage bandwidth) parameter to enhance both robustness and efficiency.

S3 Simulation Example in Section 2

We generate data from the model

$$Y = \beta_1 + \beta_2 X + \varepsilon X = 1 + 2X + \varepsilon X$$

until there are $n \in \{200, 400, 600,\ 1000\}$ observations for which $Y \geq 0$, where $X \sim U[-1, 1]$ and $\varepsilon \sim \mathcal{N}(0, 1)$. This truncation scheme ensures that only the positive Y-values are retained, effectively introducing a left-truncation mechanism in the dataset. We then use the truncated dataset to perform two types of estimation: mode-based estimation and mean estimation (least squares). For the mode-based estimation, we employ the Gaussian kernel alongside the MEM Algorithm 1, where the bandwidth parameter h_0 is chosen by the cross-validation procedure specified in (3.8). The results from both the mode-based and mean estimations are presented in Table S1 and Figure S3, where the average estimate, standard error (SE), and mean squared error (MSE) are reported based on 400 simulations.

From the results, it is evident that kernel mode-based regression is particularly well-suited for truncated data, as it consistently captures the true parameter values in scenarios of fixed truncation. This is in stark contrast to mean regression with least squares estimation, which tends to yield biased estimates when applied to the observed truncated data. The SE and MSE of the mode-based estimator decrease when the sample size n increases, which is expected from the asymptotic property listed in Section 3. Compared to existing truncated estimators, such as those discussed in Lai and Ying (1992), the kernel mode-based estimator offers significant practical advantages. Most notably, it provides numerical simplicity by avoiding the need for explicit bias corrections, which are often required in traditional truncated regression techniques. Furthermore, the mode-based estimator is flexible in terms of the error distribution. Unlike traditional estimators that often assume homoskedasticity or impose specific distributional forms, the kernel mode-based approach does not require such assumptions.

Table S1: Results of Estimation (Fixed Truncation)

	Mode-Based				Mean			
Sample Size	$\beta_{1,mode}$ (SE)	MSE($\beta_{1,mode}$)	$\beta_{2,mode}$ (SE)	MSE($\beta_{2,mode}$)	$\beta_{1,mean}$ (SE)	MSE($\beta_{1,mean}$)	$\beta_{2,mean}$ (SE)	MSE($\beta_{2,mean}$)
n=200	1.0062 (0.0189)	0.0004	1.8180 (0.2220)	0.0822	1.1588 (0.0355)	0.0265	1.6562 (0.1001)	0.1281
n=400	1.0056 (0.0106)	0.0002	1.8380 (0.1896)	0.0620	1.1596 (0.0259)	0.0261	1.6615 (0.0734)	0.1200
n=600	1.0060 (0.0094)	0.0001	1.8242 (0.1614)	0.0568	1.1573 (0.0207)	0.0252	1.6622 (0.0627)	0.1180
n=1000	1.0047 (0.0078)	0.00008	1.8351 (0.1356)	0.0455	1.1576 (0.0147)	0.0250	1.6631 (0.0460)	0.1156
true	β_1=1		$\beta_2 = 2$		β_1=1		$\beta_2 = 2$	

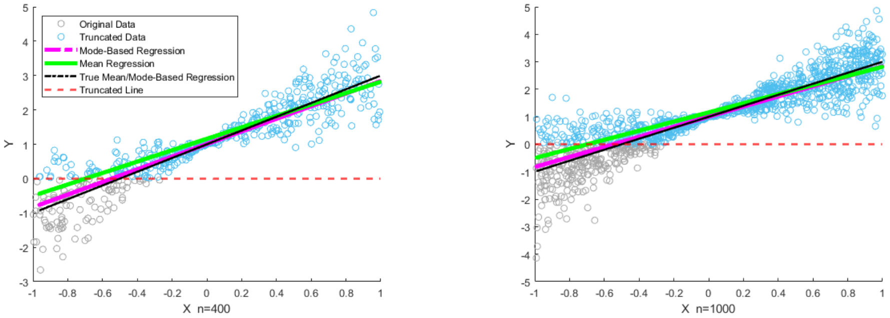

Figure S3: Results of Estimation for n =400 and 1000

S4 Comments for Conditions in Subsection 3.2

C1 assumes that the parameter space is compact, which has been commonly adopted in the literature. It should be noted that since all mean estimators exhibit bias at extreme boundary points, computing mode-based estimator becomes more advantageous when the estimators operate within a bounded and closed parameter space. C2 is utilized to guarantee that the observed data have no ties with probability one and that mode-based coefficients are identifiable. The condition $a_G \leq a_F$ is necessary for identifiability; see He and Yang (2003). Without this condition, we could have two different regression models generating the same randomly truncated observations. The bounded support condition in C3 can be released. As argued by a large number of research (Ullah et al., 2021, 2022, 2023), it is not indispensable for the kernel function to have bounded support as long as its tails are thin, for example, Gaussian kernel is permissible, which is the default kernel we use for numerical calculations. C4 is employed to make sure that associated higher order terms can be asymptotically ignored when employing Taylor

Consider the regression $Y_i = \beta_1 + \beta_2 X_{2i} + \varepsilon_i$, where $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. sequences of $\mathcal{N}(0, \sigma^2)$ random variables. In addition, for some positive constant M, suppose that we have $|\beta_1 + \beta_2 X_{2i}| < M$. If $a_F < a_G$, set $A = a_G - M$. Let $v_i = \varepsilon_i I(\varepsilon_i \geq A) + A I(\varepsilon_i < A)$, where $I(\cdot)$ is an indicator function. By construction, we have $Y_i' = (\beta_1 + \mathbb{E}(v_i)) + \beta_2 X_{2i} + (v_i - \mathbb{E}(v_i))$. Suppose that Y_i and Y_i' are subject to random truncation by an i.i.d. sequence of variable T_i. Then, both regression models will generate the same randomly truncated observations. Therefore, we cannot identify the coefficients.

expansion. C5 is expressed based on conditional expectation, indicating that the distribution of ε can depend on covariates. This is weaker than the corresponding assumptions in Chen et al. (1996) and Stute (1993), where the error term ε is assumed to be independent of the covariates. Notice that the condition $\mathbb{E}[K_h^{(1)}(\varepsilon) \mid \mathbf{X}] = 0$ guarantees the consistency of estimator, which can be satisfied if the error density function is symmetric and has a global unique mode. However, it is not necessary to require the error distribution to be symmetric. If the assumption $\mathbb{E}[K_h^{(1)}(\varepsilon) \mid \mathbf{X}] = 0$ does not hold, the suggested estimate is actually estimating the function $\hat{\mathcal{M}}(\mathbf{x}) = \arg\max_{\mathcal{M}} \mathbb{E}[K_h(Y - \mathcal{M}) \mid \mathbf{X} = \mathbf{x}]$; see Yao et al. (2012). In practice, we also allow the existence of local modes for numerical estimation, since the proposed MEM algorithm can be efficiently employed to capture the global mode estimate. C6 encapsulates the primary characteristic of the suggested mode-based estimation, where the bandwidth parameter h serves to adjust both robustness and efficiency. This assumption of a constant bandwidth enables us to establish a parametric convergence rate, facilitating comparison between the resulting mode-based estimator and traditional mean or median estimator. C7 is a common assumption to ensure the existence of both consistency and asymptotic normality.

Following a reviewer's suggestion, we also provide comments on these conditions for both the simulation studies and real data analysis. As argued by Yao and Li (2014) and Ullah et al. (2021, 2022, 2023), the choice of kernel function does not significantly impact mode estimation. This implies that various kernel functions, including the Gaussian kernel, could be used without substantial differences in the estimation results. Consequently, we opt for the Gaussian kernel for $K(\cdot)$ in this paper to conduct numerical analysis. Thus, conditions C3-C5, related to properties of the kernel function and its derivatives, can be readily satisfied by utilizing a Gaussian kernel. Condition C1 pertains to the compactness of the parameter space. In simulation studies, where both true parameter values and data are known, this condition can be satisfied because we have complete knowledge of the parameter space. Additionally, since simulation studies involve generating data from known distributions, conditions C2 and C7, related to distribution and matrix properties, can also be fulfilled. In empirical analysis, the true parameter values are unknown. However, the finite nature of the dataset and the use of linear regression to model the relationship between variables provide implicit constraints on the parameter space. Therefore, the resulting estimates can indirectly support the existence of a compact parameter space, as required by condition C1. Without meeting this condition, obtaining meaningful estimates from empirical data would be challenging or impossible. Despite the unknown true model, conditions C2 and C7 can still be satisfied in empirical analysis due to the finite sample size. Throughout numerical analysis, the bandwidth h is determined via cross-validation procedure, ensuring the fulfilment of condition C6.

S5 Convergence of the Penalized MEM Algorithm

We discuss the convergence (a sufficiently small change in the parameters) of the proposed penalized MEM algorithm. We follow the classical EM algorithm to define a stationary point of the function $Q_n^p(\boldsymbol{\beta})$ as any point of $\boldsymbol{\beta}$ where the gradient vector is zero (Wu, 1983). Following Lim and Oh (2014), we let $M(\boldsymbol{\beta})$ be the point-to-set map (a function from points to subsets) implicitly defined by the algorithm, which transitions from $\hat{\boldsymbol{\beta}}^{p(m)}$ to $\hat{\boldsymbol{\beta}}^{p(m+1)}$ for any point $\hat{\boldsymbol{\beta}}^{p(m)}$. Subsequently, we provide the theorem listed below to characterize the limit points of the set $\{\hat{\boldsymbol{\beta}}^{p(m)} : m = 0, 1, 2, \cdots\}$.

Theorem S1. *With an initial value $\hat{\boldsymbol{\beta}}^{p(0)}$, let $\hat{\boldsymbol{\beta}}^{p(m)} = M^m(\hat{\boldsymbol{\beta}}^{p(0)})$ denote the corresponding mapping. If $Q_n^p(\boldsymbol{\beta}) = Q_n^p(M(\boldsymbol{\beta}))$ holds only for stationary points $\boldsymbol{\beta}$ of Q_n^p and if $\hat{\boldsymbol{\beta}}^*$ is a limit point of the sequence $\{\hat{\boldsymbol{\beta}}^{p(m)}\}$ such that $M(\boldsymbol{\beta})$ is continuous at $\hat{\boldsymbol{\beta}}^*$, then $\hat{\boldsymbol{\beta}}^*$ is a stationary point of $Q_n^p(\boldsymbol{\beta})$.*

Theorem S1 provides a necessary condition for a point to be a limit point of the suggested algorithm. The existence of a limit point is then considered under a sufficient condition. Given $\hat{\boldsymbol{\beta}}^{p(0)}$, the set $\mathcal{B} = \{\hat{\boldsymbol{\beta}}^{p} \mid Q_n^p(\hat{\boldsymbol{\beta}}^{p}) \geq Q_n^p(\hat{\boldsymbol{\beta}}^{p(0)})\}$ is compact and contains the entire sequence $\{\hat{\boldsymbol{\beta}}^{p(m)}\}_{m=0}^{\infty}$ since $Q_n^p(\cdot^{(m+1)}) \geq Q_n^p(\cdot^{(m)})$ with the nondecreasing SCAD penalty. This ensures that the sequence has at least one limit point, which must be a stationary point of $Q_n^p(\cdot)$ according to the above theorem. If in addition, there is only one stationary point, such as when $Q_n^p(\cdot)$ is strictly concave, we may conclude that the algorithm must converge to the unique stationary point.

As pointed out by a reviewer, the suggested mode-based estimation utilizing the MEM algorithm might not offer computational advantages compared to existing estimations, primarily due to its iterative nature. For instance, with a sample size of 1000 in DGP 1, the computation time for mode-based regression is 113 seconds, whereas it is 47 seconds for Huber regression. Note that all programs are written in R and the computer has a 2.10GHz to 4.90GHz Pentium processor and 32GB memory. However, while this computational disadvantage exists, it is not a significant concern, particularly given the reasonable computation time and the robustness and efficiency achieved by the proposed method. Additionally, due to the favorable convergence property of the suggested MEM algorithm, when initial points are chosen appropriately, the MEM algorithm converges rapidly toward the neighborhood of a stationary point, ensuring efficient computation.

S6 Multimodal Case

As mentioned in the paper, the global unique mode assumption can be released without affecting the estimation procedure. Multimodal datasets are common in economics. For example, when examining a country's income distribution, it is evident that there are often two modes corresponding to developing and developed countries, reflecting a dichotomous world composed of nations with varying income levels. The proposed mode-based regression can be employed to capture these two distinct situations simultaneously by using the suggested MEM algorithm. To demonstrate that the developed estimation method can effectively handle multimodal case, we conduct a Monte Carlo simulation as described below.

We generate random samples from the following model

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + X_{1i}\varepsilon_i,\ i = 1, \cdots, n,$$

where we set the parameters to be $\boldsymbol{\beta} = (1, 2)^T$. The covariate vector $\mathbf{X}_i$ is normally distributed with mean 0, variance $I_{2\times 2}$, and correlation $0.2^{|k-j|}$, where $k, j = 1, 2$. To create a multimodal case, 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 have variances equal to 1 (Figure S4).

The generalized errors $\{\varepsilon_i\}_{i=1}^n$ indicate that $\mathbb{E}(\varepsilon_i) = 2$ and $Mode(\varepsilon_i) = 0$ or 4. In this scenario, mean regression may produce misleading results

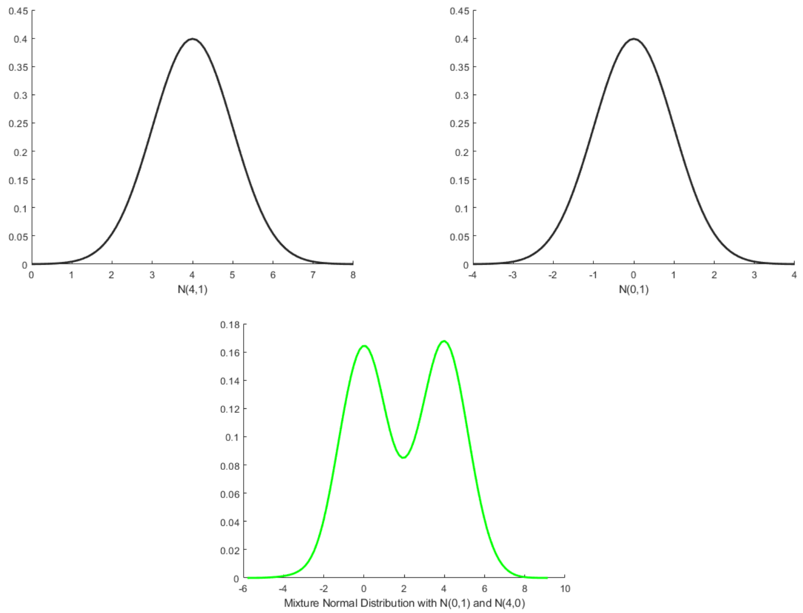

Figure S4: Mixture Normal Distribution with Two Modes

by disregarding data heterogeneity. We then have the following equations showing two different mode-based regression lines

$$\begin{cases} \text{Mean Regression: } \mathbb{E}(Y_i \mid \mathbf{X}_i) = 3X_{1i} + 2X_{2i}, \\ \text{Mode-Based Regression Line 1: } Mode(Y_i \mid \mathbf{X}_i) = X_{1i} + 2X_{2i}, \\ \text{Mode-Based Regression Line 2: } Mode(Y_i \mid \mathbf{X}_i) = 5X_{1i} + 2X_{2i}. \end{cases}$$

We consider data samples of size $n \in \{200, 400, 600\}$ with 400 replications to assess finite sample performance. The average estimate, standard error (SE), and mean squared error (MSE) for each estimator are calculated for evaluation. Table S2 displays the simulation results, demonstrating the effectiveness of the proposed estimation method in handling multimodal case with finite samples. This efficacy stems from the approximate initialization of estimates, which facilitates capturing different mode-based regression lines for data exhibiting multiple modes using the suggested MEM algorithm. For both mode-based regression lines, the estimates of unknown parameters converge closer to the true values and both the SE and MSE of each estimator decrease with increasing sample size n.

Table S2: Results of Estimation for Multimodal Case

Method	n	β_1(SE)	MSE(β_1)	β_2(SE)	MSE(β_2)
Mode-Based 1	200	1.0017 (0.0583)	0.0035	2.0063 (0.0497)	0.0028
	400	0.9963 (0.0378)	0.0015	2.0048 (0.0388)	0.0015
	600	0.9982 (0.0306)	0.0010	2.0012 (0.0295)	0.0009
Mode-Based 2	200	4.9862 (0.1267)	0.0163	2.0072 (0.1091)	0.0119
	400	4.9472 (0.0983)	0.0126	1.9938 (0.0836)	0.0068
	600	4.9603 (0.0822)	0.0092	1.9926 (0.0721)	0.0056
Mean	200	2.9634 (0.3976)	0.1589	2.0023 (0.4028)	0.1625
	400	2.9813 (0.2882)	0.0838	1.9954 (0.2967)	0.0903
	600	2.9920 (0.2357)	0.0560	2.0006 (0.2302)	0.0544

S7 Numerical Examples in Section 5

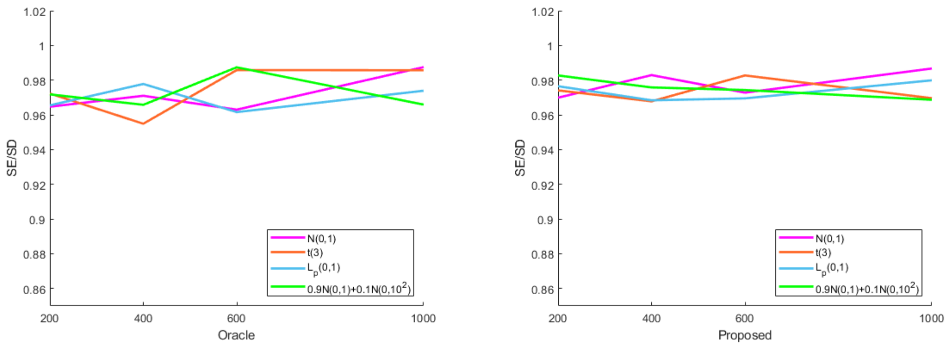

Figure S5: Ratio of SE/SD of DGP 1

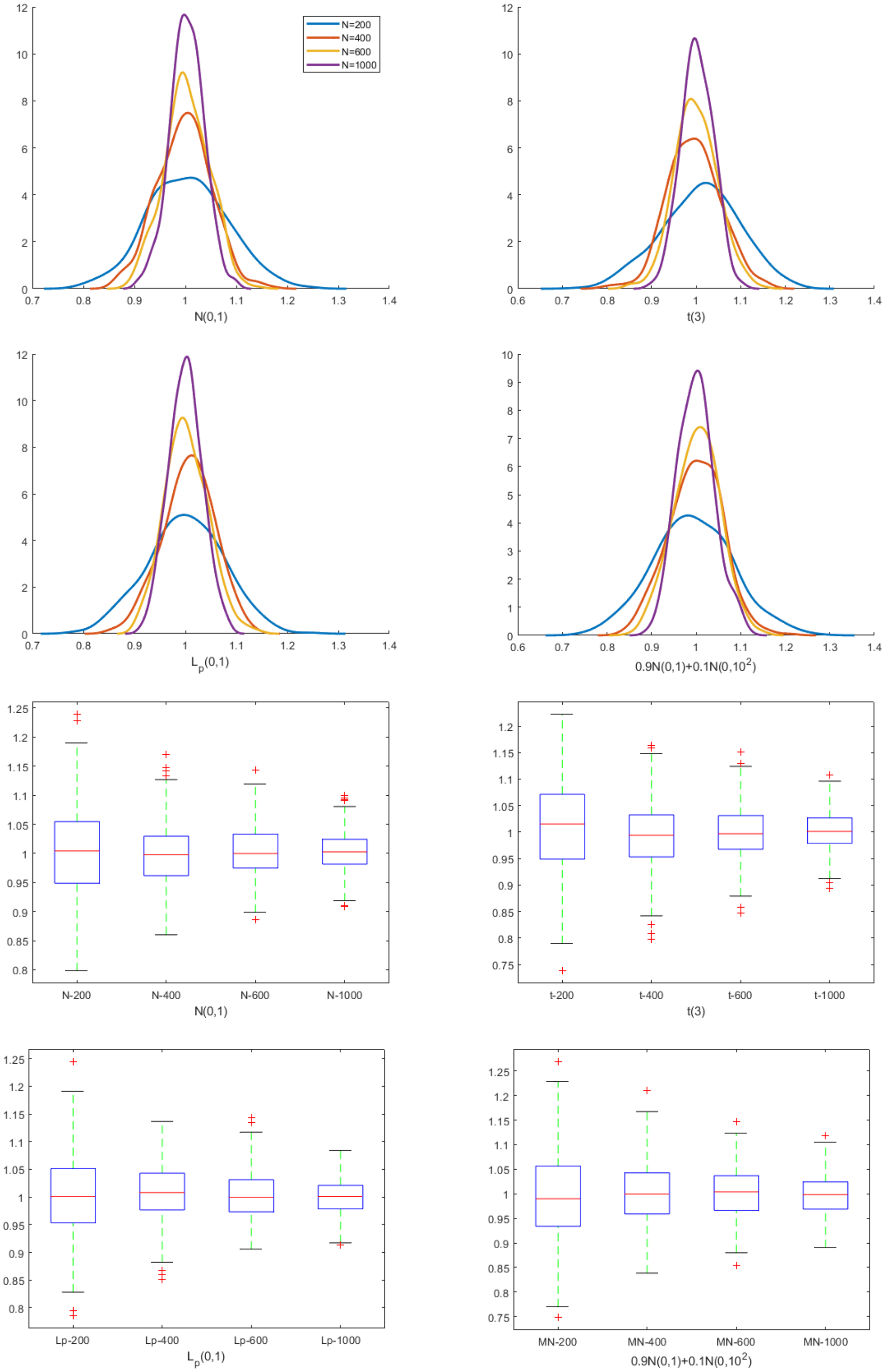

Figure S6: Distributions and Boxplots of Proposed Estimators

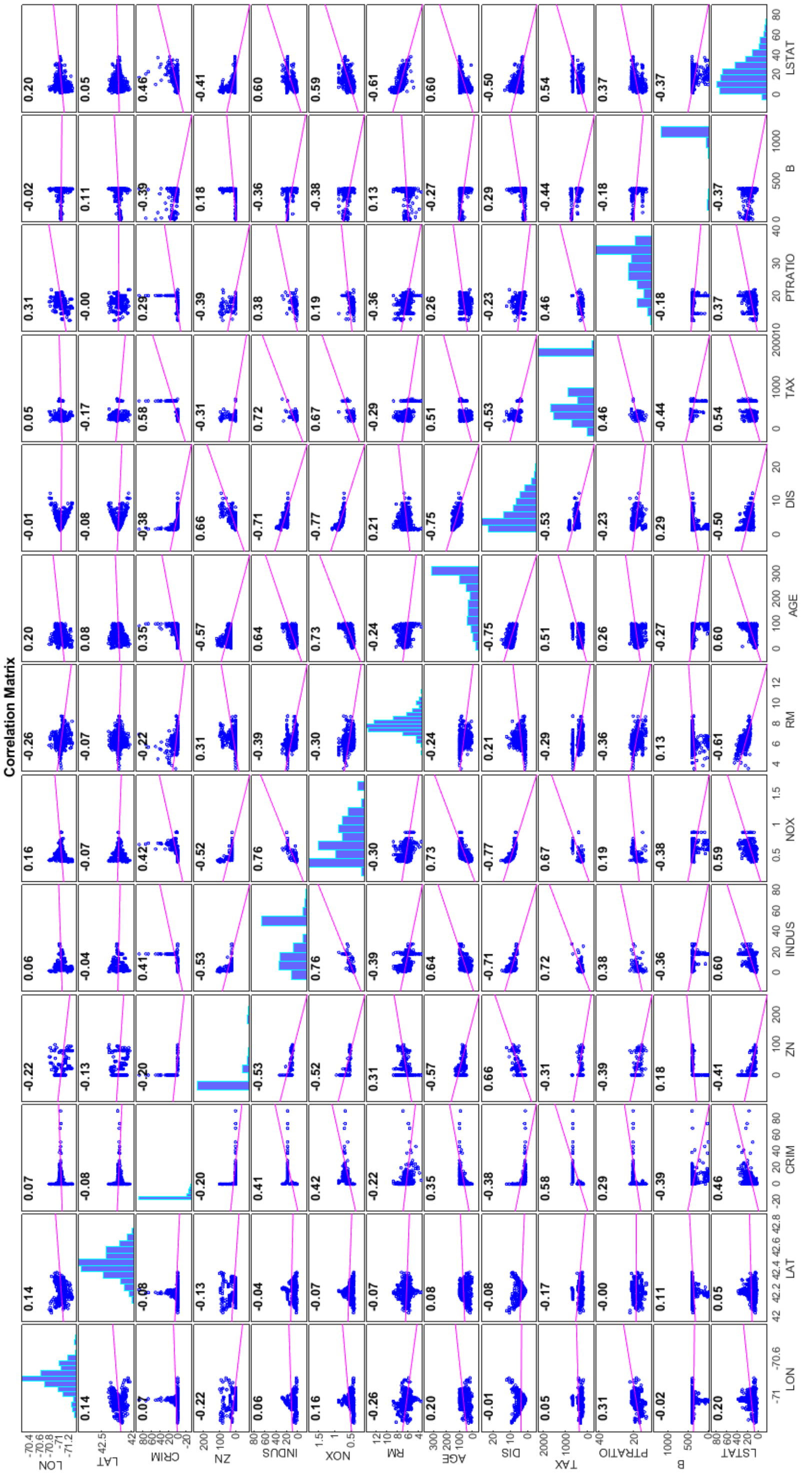

Figure S8: Correlation Matrix of Covariates

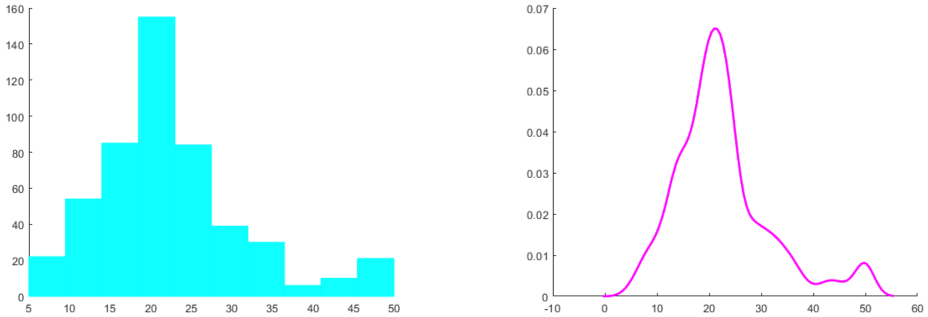

Figure S7: Histogram and Empirical Distribution of Dependent Variable

S8 Extension to Left-Truncated and Right-Censored Data

We in this section demonstrate that the proposed kernel mode-based estimation procedure can be extended to analyze left-truncated and right-censored data, which are commonly encountered in survival analysis and economic studies; see Gross and Lai (1996), Zhou and Yip (1999), Su and Wang (2012), among others. Left-truncated data arise when individuals or units enter the study after a certain event has already occurred, leading to biased sampling if not properly accounted for. In such cases, the risk set is conditioned on having already survived until the truncation point, which can result in overestimating survival probabilities if the truncation mechanism is not adequately modeled. On the other hand, right-censored data occur when the event of interest has not been observed for some individuals by the end of the study period, which often leads to incomplete information.

Suppose that we are interested in studying a certain event for individuals during the time period (τ_0, τ) with $\tau_0 < \tau$. The sampling strategy involves recruiting all individuals who have experienced a first event between τ_0 and τ and who have not experienced a second event by the time τ for a prospective follow-up study. This study will be terminated at time τ^* with $\tau^* > \tau$, representing the final endpoint of the observation period. Let T_s denote the initial time of the first event. The variable T is used to denote the time from T_s to the occurrence of the second event. The variable V is defined as the time from T_s to τ, representing the recruitment time at τ. To account for the censoring that occurs in the follow-up study, we introduce the variable C, which represents the time from T_s to the censoring event. In this framework, the censoring time C is determined as $C = \min(C_1, C_2)$ and $P(C \geq V) = 1$, where $C_1 = V + \tau^* - \tau$ represents the time from the first event (beginning of the study) to the end of the study at τ^*, and C_2 denotes the time from the first event to the individual's dropout from the study, which may occur before the study's termination at τ^*. The condition $P(C \geq V) = 1$ ensures that no individual can be censored before recruitment at τ, meaning that all individuals recruited have experienced the first event within the time window (τ_0, τ) and have not yet experienced a second event. This assumption holds due to the design of the study, where recruitment only occurs after the first event and before the censoring time.

Furthermore, the analysis considers a set of covariates, denoted by $\mathbf{X}$, which may influence the event times and censoring mechanisms. We further assume that the variables T, V, and C are continuous. We make the assumption that $(T, \mathbf{X})$ and (V, C) are independent of each other, but V and C are dependent with the condition $P(C \geq V) = 1$. For left-truncated and right-censored data, we observe nothing if $T < V$, and observe $(Z, V, \delta, \mathbf{X})$ with $\delta = I(T < C)$ and $Z = \min(T, C)$ if $T \geq V$. Note that $\delta = I(T < C)$ is an indicator variable that equals 1 if the second event occurs before censoring and 0 if the observation is censored, and $Z = \min(T, C)$ represents the observed event time, either the time of the second event or the censoring time. Given this structure, we consider the following semiparametric linear regression model for the event time

$$T_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \; i = 1, 2, \cdots, n,$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients to be estimated and ε is the random error term. The presence of left truncation (where $T < V$) and right censoring (where $T > C$) introduces additional complexities into the estimation of $\boldsymbol{\beta}$. To estimate $\boldsymbol{\beta}$, we implement the developed techniques in this paper in combination with kernel smoothing methods to account for the censoring and truncation mechanisms.

Let $F(t) = P(T \leq t)$ denote the cumulative distribution function of the event time T. Additionally, let $Q(t) = P(C \leq t)$ and $G(t) = P(V \leq t)$ denote the cumulative distribution functions of C and V, respectively. Define $F(t, \mathbf{x}) = P(T \leq t, \mathbf{X} \leq \mathbf{x})$. Let $\alpha = P(V \leq T)$ denote the probability of untruncation, that is, the probability that we observe the random variable T and the individual is not truncated out of the study. Let a_F and b_F denote the left and right endpoints of the support of F, and similarly, define (a_G, b_G) and (a_Q, b_Q) as the left and right endpoints of V and C, respectively. To ensure identifiability of $F(t)$, we impose the following assumptions

$$a_G = a_F = a_Q = 0,\ b_G \leq \min(b_F, b_Q),\ \text{and } b_F \leq b_Q,$$

where $b_G \leq \min(b_F, b_Q)$ ensures that truncation occurs before or at the end of the event time and censoring period, and $b_F \leq b_Q$ indicates that the second event either occurs before or at the censoring time.

Given the observed left-truncated and right-censored sample $\{Z_i, V_i, \delta_i, \mathbf{X}_i\}_{i=1}^n$, we can express $\tilde{F}(z, \mathbf{x})$ as

$$\begin{aligned}\tilde{F}(z, \mathbf{x}) &= P(Z_i \leq z, \delta_i = 1, \mathbf{X}_i \leq \mathbf{x}) = \alpha^{-1} P(V \leq T \leq C, T \leq z, \mathbf{X} \leq \mathbf{x}) \\ &= \alpha^{-1} \int_{u \leq z} \int_{\mathbf{w} \leq \mathbf{x}} P(V \leq u \leq C) F(du, d\mathbf{w}).\end{aligned}$$

From this, we can obtain

$$F(z, \mathbf{x}) = \alpha \int_{u \leq z} \int_{\mathbf{w} \leq \mathbf{x}} \frac{1}{P(V \leq u \leq C)} \tilde{F}(z, \mathbf{x}).$$

For left-truncated and right-censored data, the product-limit estimator of $F(t)$, also known as the Kaplan-Meier estimator, is given by

$$\hat{F}_n(z) = 1 - \prod_{u \le z} \left[1 - \frac{N(du)}{nR_n(u)}\right],$$

where $R_n(u) = n^{-1} \sum_{i=1}^{n} I(V_i \le u \le Z_i)$ is the risk set at time u, $N(u) = \sum_{i=1}^{n} I(Z_i \le u, \delta_i = 1)$ is the counting process for the observed events, and $N(du) = N(u) - N(u-)$ represents the increments in the counting process.

Following Shen (2005), we can consider two estimators of the untruncation probability α, i.e.,

$$\alpha_n = \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 - \hat{F}_n(V_i)}\right]^{-1},$$

which is a sample-based estimator that adjusts for truncation by weighting the contributions of the observed individuals based on their truncation times, and

$$\hat{\alpha}_n(z) = \frac{[\hat{G}_n(z) - \hat{Q}_n(z-)][1 - \hat{F}_n(z-)]}{R_n(z)},$$

where $\hat{G}_n(z)$ and $\hat{Q}_n(z)$ are the inverse-probability-weighted estimators such that

$$\hat{G}_n(z) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_n(V_i)}\right]^{-1} \sum_{i=1}^{n} \frac{I(V_i \le z)}{1 - \hat{F}_n(V_i)},$$

and

$$\hat{Q}_n(z) = \left[\sum_{i=1}^{n} \frac{1}{1 - \hat{F}_n(V_i)}\right]^{-1} \sum_{i=1}^{n} \frac{I(Z_i \le z, \delta_i = 0)}{1 - \hat{F}_n(Z_i)},$$

respectively. Shen (2005) showed that $\hat{\alpha}_n(z)$ does not depend on z and is equivalent to α_n. This implies that the untruncation probability α_n, estimated using the sample data, can be treated as a constant and does not vary with z.

Given that $P(V \leq z \leq C)$ can be consistently estimated by $\hat{G}_n(z) - \hat{Q}_n(z)$, we can now construct a nonparametric estimate of $F(z, \mathbf{x})$ for left-truncated and right-censored data, which is

$$\hat{F}_n(z, \mathbf{x}) = \int_{u \leq z} \int_{\mathbf{w} \leq \mathbf{x}} \frac{1}{\alpha_n^{-1}(\hat{G}_n(u) - \hat{Q}_n(u))} \tilde{F}_n(z, \mathbf{x}),$$

where $\tilde{F}_n(z, \mathbf{x}) = n^{-1} \sum_{i=1}^n I(Z_i \leq z, \delta_i = 1, \mathbf{X}_i \leq \mathbf{x})$ is the empirical distribution function of $\tilde{F}(z, \mathbf{x})$ based on the observed data. Note that $\alpha_n^{-1}(\hat{G}_n(u) - \hat{Q}_n(u))$ plays a key role in the nonparametric estimation and adjusts for both truncation and censoring, which is computed as

$$\alpha_n^{-1}(\hat{G}_n(u) - \hat{Q}_n(u)) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 - \hat{F}_n(V_i)} - \frac{1}{n} \sum_{i=1}^n \frac{I(V_i \leq z)}{1 - \hat{F}_n(V_i)}.$$

When there is no censoring or truncation, let $\{T_i, \mathbf{X}_i\}_{i=1}^n$ denote the observed sample. The kernel mode-based regression estimate is obtained by finding the value of $\boldsymbol{\beta}$ that maximizes the objective function $\sum_{i=1}^n K_h(T_i - X_i^T \boldsymbol{\beta})$, which can be expressed in integral form as

$$\frac{1}{h} \int_{-\infty}^{\infty} \int_0^{\infty} K\left(\frac{T - \mathbf{X}^T \boldsymbol{\beta}}{h}\right) \hat{F}_{T\mathbf{X}}(dt, d\mathbf{x}),$$

where $\hat{F}_{T\mathbf{X}}(t, \mathbf{x})$ is the empirical distribution function of $(T_i, \mathbf{X}_i)$. For left-truncated and right-censored data, the empirical joint distribution function $\hat{F}_{T\mathbf{X}}(t, \mathbf{x})$ is replaced by the corresponding estimator $\hat{F}_n(z, \mathbf{x})$. Thus, the kernel mode-based objective function for censored and truncated data becomes

$$\begin{aligned}
&\frac{1}{h}\int_{-\infty}^{\infty}\int_{0}^{\infty} K\left(\frac{Z - \mathbf{X}^T\boldsymbol{\beta}}{h}\right)\hat{F}_n(dz, d\mathbf{x}) \\
=&\frac{1}{h}\int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\alpha_n^{-1}(\hat{G}_n(z) - \hat{Q}_n(z))} K\left(\frac{Z - \mathbf{X}^T\boldsymbol{\beta}}{h}\right)\tilde{F}_n(dz, d\mathbf{x}).
\end{aligned}$$

The kernel mode-based regression estimate $\hat{\boldsymbol{\beta}}$ for left-truncated and right-censored data is then defined as the solution that maximize the following objective function

$$Q_n(\boldsymbol{\beta}) = \frac{1}{nh}\sum_{i=1}^{n} \frac{I(\delta_i = 1)}{\alpha_n^{-1}(\hat{G}_n(Z_i) - \hat{Q}_n(Z_i))} K\left(\frac{Z_i - \mathbf{X}_i^T\boldsymbol{\beta}}{h}\right),$$

which generalizes the kernel mode-based regression framework to handle left-truncated and right-censored data. The detailed asymptotic properties of the resulting estimator, such as its consistency and asymptotic normality, can be derived using empirical process theory and techniques similar to those outlined in Section S9. Additionally, empirical likelihood estimation developed in this paper could be used to improve inference about $\boldsymbol{\beta}$. These topics are left for future research to further refine and extend the proposed methodology for truncated and censored data.

S9 Technical Proofs

Proof of Theorem 1

By Corollary 2.5 of He and Yang (1998), we know that $\alpha_n \to \alpha_0 = \int G_0(x)\, dF(x) = \alpha/G(b_F)$ almost surely as $n \to \infty$. Clearly, $\alpha/G(x) = \alpha_0/G_0(x)$ for all $x \in (a_G, b_F]$. To establish identification of the model, by using the law of total expectation, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{\alpha}{h}\frac{1}{G\left(U\right)}K\left(\frac{U-\mathbf{W}^T\boldsymbol{\beta}}{h}\right)\Big|\mathbf{X}\right]\\
=&\mathbb{E}\left\{\mathbb{E}\left[\frac{\alpha}{h}\frac{1}{G\left(U\right)}K\left(\frac{U-\mathbf{W}^T\boldsymbol{\beta}}{h}\right)\Big|Y,\mathbf{X}\right]\Big|\mathbf{X}\right\}\\
=&\mathbb{E}\left\{\mathbb{E}\left[\frac{\alpha}{h}\frac{1}{G\left(Y\right)}K\left(\frac{Y-\mathbf{X}^T\boldsymbol{\beta}}{h}\right)\Big|Y,\mathbf{X}\right]\Big|\mathbf{X}\right\}=\mathbb{E}\left[\frac{1}{h}K\left(\frac{Y-\mathbf{X}^T\boldsymbol{\beta}}{h}\right)\Big|\mathbf{X}\right]
\end{aligned}$$

because of $F_n(y,\mathbf{x}) = \alpha_n \int_{u\le y}\int_{\mathbf{w}\le\mathbf{x}} \frac{1}{G_n(u)} F_n^*(du, d\mathbf{w})$ and $\int \frac{1}{h}K\left(\frac{y-\mathbf{x}^T\boldsymbol{\beta}}{h}\right) dF_n(y,\mathbf{x}) = \int_{u\le y}\int_{\mathbf{w}\le\mathbf{x}} \frac{\alpha_n}{G_n(u)}\frac{1}{h}K\left(\frac{u-\mathbf{w}^T\boldsymbol{\beta}}{h}\right) dF_n^*(u,\mathbf{w})$ shown in the paper. The identification follows from the fact that this expression is uniquely determined for a given $\boldsymbol{\beta}$ under the regularity conditions C1-C7. Based on this, we can prove $|Q_n(\boldsymbol{\beta}) - Q_N(\boldsymbol{\beta})| = o_p(1)$, where

$$\begin{aligned}
&|Q_n(\boldsymbol{\beta}) - Q_N(\boldsymbol{\beta})| = |Q_n(\boldsymbol{\beta}) - \mathbb{E}(Q_n(\boldsymbol{\beta})) + \mathbb{E}(Q_n(\boldsymbol{\beta})) - \mathbb{E}(Q_N(\boldsymbol{\beta}))\\
+&\mathbb{E}(Q_N(\boldsymbol{\beta})) - Q_N(\boldsymbol{\beta})| = |Q_n(\boldsymbol{\beta}) - \mathbb{E}(Q_n(\boldsymbol{\beta})) + \mathbb{E}(Q_N(\boldsymbol{\beta})) - Q_N(\boldsymbol{\beta})|\\
\le&|Q_n(\boldsymbol{\beta}) - \mathbb{E}(Q_n(\boldsymbol{\beta}))| + |\mathbb{E}(Q_N(\boldsymbol{\beta})) - Q_N(\boldsymbol{\beta})| = o_p(1)
\end{aligned}$$

according to the triangle inequality and the Law of Large Numbers.

$\square$

Lemma 1. *Let $\psi_n(\boldsymbol{\beta})$ be a sequence of random functions on $\mathbb{R}^p$, which is concave in $\boldsymbol{\beta}$. Let $\psi(\boldsymbol{\beta})$ be a random function such that for each fixed $\boldsymbol{\beta}$, $\psi_n(\boldsymbol{\beta}) \to \psi(\boldsymbol{\beta})$ with probability one and in probability. Then, for any compact set $\Omega \subset \mathbb{R}^p$, we can obtain*

$$\sup_{\boldsymbol{\beta}\in\Omega}|\psi_n(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \to 0$$

with probability one and in probability, respectively.

Proof. Consider any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \Omega$ and $\lambda \in [0,1]$. For all n, we have

$$\psi_n(\lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2) \geq \lambda\psi_n(\boldsymbol{\beta}_1) + (1-\lambda)\psi_n(\boldsymbol{\beta}_2).$$

Due to the continuity and almost sure convergence of the sequence $\{\psi_n\}$, by taking the limit as $n \to \infty$ (almost surely) on both sides, we obtain

$$\psi(\lambda\boldsymbol{\beta}_1 + (1-\lambda)\boldsymbol{\beta}_2) \geq \lambda\psi(\boldsymbol{\beta}_1) + (1-\lambda)\psi(\boldsymbol{\beta}_2),$$

which confirms that $\psi(\boldsymbol{\beta})$ is concave. Since $\psi(\boldsymbol{\beta})$ is concave and finite on the compact set Ω, it is continuous on Ω. Specifically, for any $\boldsymbol{\beta}_0 \in \Omega$ and any sequence $\{\boldsymbol{\beta}_k\} \subset \Omega$ such that $\boldsymbol{\beta}_k \to \boldsymbol{\beta}_0$, we have $\lim_{k\to\infty}\psi(\boldsymbol{\beta}_k) = \psi(\boldsymbol{\beta}_0)$.

We then prove by contradiction. Suppose that the convergence is not uniform on Ω. Thereupon, there exists an $\epsilon > 0$ and a subsequence $\{n_k\}$ such that $\sup_{\boldsymbol{\beta}\in\Omega}|\psi_{n_k}(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \geq \epsilon$ for all k. This means that for each k, there exists $\boldsymbol{\beta}_{n_k} \in \Omega$ such that $|\psi_{n_k}(\boldsymbol{\beta}_{n_k}) - \psi(\boldsymbol{\beta}_{n_k})| \geq \epsilon$. Since Ω is compact, the sequence $\{\boldsymbol{\beta}_{n_k}\}$ has a convergent subsequence. Without loss of generality, assume that $\boldsymbol{\beta}_{n_k} \to \boldsymbol{\beta}_0 \in \Omega$ as $k \to \infty$. Since $\psi_n(\boldsymbol{\beta}) \to \psi(\boldsymbol{\beta})$

almost surely for each fixed $\boldsymbol{\beta}$, we have $\lim_{k\to\infty} \psi_{n_k}(\boldsymbol{\beta}_{n_k}) = \psi(\boldsymbol{\beta}_0)$. Also, because $\psi(\boldsymbol{\beta})$ is continuous on Ω, we have $\lim_{k\to\infty} \psi(\boldsymbol{\beta}_{n_k}) = \psi(\boldsymbol{\beta}_0)$. We then get $\lim_{k\to\infty}|\psi_{n_k}(\boldsymbol{\beta}_{n_k}) - \psi(\boldsymbol{\beta}_{n_k})| = 0$, which contradicts with $|\psi_{n_k}(\boldsymbol{\beta}_{n_k}) - \psi(\boldsymbol{\beta}_{n_k})| \geq \epsilon$ for all k. Therefore, $\sup_{\boldsymbol{\beta}\in\Omega}|\psi_n(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \to 0$ almost surely.

To show convergence in probability, for any $\epsilon > 0$ and $\delta > 0$, we have

$$P\left(\sup_{\boldsymbol{\beta}\in\Omega}|\psi_n(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \geq \epsilon\right) \leq \delta$$

for sufficiently large n. This follows because the pointwise convergence in probability implies that for each $\boldsymbol{\beta}$, $P\left(|\psi_n(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \geq \epsilon\right) \to 0$ as $n \to \infty$. Since Ω is compact, it can be covered by a finite number of open balls due to its total boundedness. Let $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m\}$ be a finite $\epsilon/3$-net for Ω, meaning that for every $\boldsymbol{\beta} \in \Omega$, there exists some $\boldsymbol{\beta}_i$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_i\| < (\epsilon/3)L$, where L is a Lipschitz constant for ψ_n and ψ (which exists due to concavity and compactness). Using the union bound, we obtain

$$P\left(\sup_{\boldsymbol{\beta}\in\Omega}|\psi_n(\boldsymbol{\beta}) - \psi(\boldsymbol{\beta})| \geq \epsilon\right) \leq \sum_{i=1}^{m} P\left(|\psi_n(\boldsymbol{\beta}_i) - \psi(\boldsymbol{\beta}_i)| \geq \epsilon/3\right).$$

Since each term on the right-hand side converges to zero as $n \to \infty$, the sum also converges to zero, ensuring convergence in probability. We then prove the lemma. Note that there are many other versions for the proof of this Convexity Lemma. Interested readers are referred to Pollard (1991).

□

Lemma 2. *Define* $K_h(\mathbf{u}) = h^{-1}K(\mathbf{u}/h)$*. For* $\mathbf{u} \in \mathbb{R}^p$*, define* $R_n(\mathbf{x}, y, \mathbf{u}) = K_h(y - \mathbf{x}^T(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n})) - K_h(y - \mathbf{x}^T\boldsymbol{\beta}_0) + K_h^{(1)}(y - \mathbf{x}^T\boldsymbol{\beta}_0)\mathbf{x}^T\mathbf{u}/\sqrt{n}$ *and* $\mathcal{Q}(\mathbf{u}) = \alpha_n \sum_{i=1}^n [G_n(U_i)]^{-1}\ R_n(\mathbf{W}_i, U_i, \mathbf{u}) = \int nR_n(\mathbf{x}, y, \mathbf{u})F_n(dy, d\mathbf{x})$*. Under the conditions C1-C7, for any fixed* $C > 0$*, we have*

$$\sup_{\|\mathbf{u}\|<C} \left|\mathcal{Q}(\mathbf{u}) - \frac{1}{2}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}\right| \to 0 \textit{ in probability.}$$

Proof. We aim to show that $\mathcal{Q}(\mathbf{u})$ converges to its limiting value uniformly over $\mathbf{u} \in \mathbb{R}^p$. Since $\mathcal{Q}(\mathbf{u})$ is concave in $\mathbf{u}$, according to Lemma 1, the result will follow if we can show $\mathcal{Q}(\mathbf{u}) \to \frac{1}{2}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}$ in probability. At first, using a second-order Taylor expansion around $\boldsymbol{\beta}_0$, we obtain

$$\begin{aligned}
&K_h(y - \mathbf{x}^T(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n}))\\
=&K_h(y - \mathbf{x}^T\boldsymbol{\beta}_0) + K_h^{(1)}(y - \mathbf{x}^T\boldsymbol{\beta}_0)\frac{\mathbf{x}^T\mathbf{u}}{\sqrt{n}} + \frac{1}{2}K_h^{(2)}(y - \mathbf{x}^T\boldsymbol{\beta}_0)\frac{(\mathbf{x}^T\mathbf{u})^2}{n} + o_p\left(\frac{1}{n}\right).
\end{aligned}$$

Substituting this into the expression for $R_n(\mathbf{x}, y, \mathbf{u})$, we find that

$$R_n(\mathbf{x}, y, \mathbf{u}) = \frac{1}{2}K_h^{(2)}(y - \mathbf{x}^T\boldsymbol{\beta}_0)\frac{(\mathbf{x}^T\mathbf{u})^2}{n} + o_p\left(\frac{1}{n}\right).$$

Thus, by the Law of Large Numbers, we have

$$\begin{aligned}
&\int nR_n(\mathbf{x}, y, \mathbf{u})F(dy, d\mathbf{x}) = \mathbb{E}[nR_n(\mathbf{x}, y, \mathbf{u})]\\
=&\frac{\mathbf{u}^T}{2}\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u} + o_p(1),
\end{aligned}$$

which provides the pointwise convergence of $\mathcal{Q}(\mathbf{u})$. To extend to uniform convergence, we use the fact that $R_n(\mathbf{x}, y, \mathbf{u})$ is a smooth function of $\mathbf{u}$. The second-order Taylor expansion we applied is valid uniformly over $\mathbf{u} \in \Omega$ because of the compactness of Ω and the smoothness of $K(\cdot)$. By standard results in empirical process theory, the remainder terms $o_p(1)$ in the Taylor expansion are uniformly small over compact sets. This gives

$$\sup_{\|\mathbf{u}\|<C} \big|\mathcal{Q}(\mathbf{u}) - \frac{1}{2}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}\big| = o_p(1).$$

Thus, the convergence is uniform over $\mathbf{u} \in \Omega$.

We now need to control the difference between the empirical measure F_n and the true distribution F. By some calculations, we can get

$$\begin{aligned}
&\Big|\int nR_n(\mathbf{x}, y, \mathbf{u})F_n(dy, d\mathbf{x}) - \int nR_n(\mathbf{x}, y, \mathbf{u})F(dy, d\mathbf{x})\Big| \\
=&\Big|\alpha_n \int nR_n(\mathbf{x}, y, \mathbf{u})\frac{F_n^*(dy, d\mathbf{x})}{G_n(y)} - \alpha \int nR_n(\mathbf{x}, y, \mathbf{u})\frac{F^*(dy, d\mathbf{x})}{G(y)}\Big| \\
\leq&\Big|\alpha_n \int nR_n(\mathbf{x}, y, \mathbf{u})\Big(\frac{1}{G_n(y)} - \frac{1}{G(y)}\Big)F_n^*(dy, d\mathbf{x})\Big| \\
&+ \Big|\alpha_n \int nR_n(\mathbf{x}, y, \mathbf{u})\frac{F_n^*(dy, d\mathbf{x})}{G(y)} - \alpha \int nR_n(\mathbf{x}, y, \mathbf{u})\frac{F^*(dy, d\mathbf{x})}{G(y)}\Big| \\
\leq&\Big|\int nR_n(\mathbf{x}, y, \mathbf{u})\Big(\frac{G(y) - G_n(y)}{G(y)}\Big)F_n(dy, d\mathbf{x})\Big| + \Big|(\alpha_n - \alpha)\int nR_n(\mathbf{x}, y, \mathbf{u}) \\
&\frac{F_n^*(dy, d\mathbf{x})}{G(y)} + \alpha \int nR_n(\mathbf{x}, y, \mathbf{u})\frac{1}{G(y)}(F_n^*(dy, d\mathbf{x}) - F^*(dy, d\mathbf{x}))\Big| \\
\leq&\sup \sqrt{n}|G(y) - G_n(y)| \int |\sqrt{n}R_n(\mathbf{x}, y, \mathbf{u})|\frac{F_n(dy, d\mathbf{x})}{G(y)} \\
&+ \sqrt{n}|\alpha_n - \alpha| \int |\sqrt{n}R_n(\mathbf{x}, y, \mathbf{u})|\frac{F_n^*(dy, d\mathbf{x})}{G(y)} \\
&+ \alpha\Big|\int nR_n(\mathbf{x}, y, \mathbf{u})\frac{1}{G(y)}(F_n^*(dy, d\mathbf{x}) - F^*(dy, d\mathbf{x}))\Big| = o_p(1).
\end{aligned}$$

The first two terms are $o_p(1)$ due to the fact that $\sqrt{n}R_n(\mathbf{x}, y, \mathbf{u}) \to 0$, $\sup \sqrt{n}|G_n(y) - G(y)| = O_p(1)$, and $\sqrt{n}(\alpha_n - \alpha) = O_p(1)$; see He and Yang (2003). The last term is also $o_p(1)$ according to van der Vaart (1998), where empirical processes converge uniformly over compact sets to their population counterparts. Combining all the above results, we conclude that $\mathcal{Q}(\mathbf{u}) - \int nR_n(\mathbf{x}, y, \mathbf{u})F(dy, d\mathbf{x}) \xrightarrow{p} 0$. This completes the proof.

□

Lemma 3. *Define $K_h(\mathbf{u}) = h^{-1}K(\mathbf{u}/h)$. Under the conditions C1-C7, we can obtain that*

$$\arg\max S_n(\mathbf{u}) \to \arg\max V(\mathbf{u}) = (\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\})^{-1}\mathcal{W}$$

in distribution, where $\mathbf{u} \in \mathbb{R}^p$, $S_n(\mathbf{u}) = \sum_{i=1}^n \alpha_n[G_n(U_i)]^{-1}K_h(y - \mathbf{x}^T(\boldsymbol{\beta}_0 + \mathbf{u}/\sqrt{n})) - K_h(y - \mathbf{x}^T\boldsymbol{\beta}_0)$, $V(\mathbf{u}) = 2^{-1}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u} - \mathcal{W}^T\mathbf{u}$, and $\mathcal{W} \sim \mathcal{N}(0, \Sigma)$.

Proof. Due to the use of concave kernel function $K(\cdot)$, it is natural to argue that $S_n(\mathbf{u})$ inherits this concavity and its maximizer is well-defined. Similarly, the quadratic form in $V(\mathbf{u})$ (i.e., the term $2^{-1}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}$) ensures that $V(\mathbf{u})$ has a unique maximizer (due to the positive definite Hessian matrix), which is given by the solution to $\partial V(\mathbf{u})/\partial \mathbf{u} = 0$. This leads to $\mathbf{u} = (\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\})^{-1}\mathcal{W}$.

The Taylor expansion of $S_n(\mathbf{u})$ around $\boldsymbol{\beta}_0$ is given by $S_n(\mathbf{u}) \approx S_n(\mathbf{0}) -$

$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_n[G_n(U_i)]^{-1}K_h^{(1)}(\varepsilon_i)\mathbf{x}_i^T\mathbf{u}+\frac{1}{2n}\sum_{i=1}^{n}\alpha_n[G_n(U_i)]^{-1}K_h^{(2)}(\varepsilon_i)(\mathbf{x}_i^T\mathbf{u})^2$, where $\varepsilon_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0$. We focus on $S_{n,1} = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_n[G_n(U_i)]^{-1}K_h^{(1)}(\varepsilon_i)$ $\mathbf{x}_i^T\mathbf{u}$. Our aim is to show that $S_{n,1}$ converges in distribution to $-\mathcal{W}^T\mathbf{u}$. We then mainly need to show that $n^{-1}\sum_{i=1}^{n}G_n^{-1}(U_i)K_h^{(1)}(\varepsilon_i)\xrightarrow{p}\alpha^{-1}\mathbb{E}[K_h^{(1)}(\varepsilon_i)]$ because this term represents the gradient (or score function) in the estimation process, and the maximization of $S_n(\mathbf{u})$ is driven by the behavior of its gradient. We decompose the sum into two parts

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G_n(U_i)}K_h^{(1)}(\varepsilon_i) =&\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{G_n(U_i)}-\frac{1}{G(U_i)}\right)K_h^{(1)}(\varepsilon_i)\\ &+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i) = D_{n1}+D_{n2}.\end{aligned}$$

The conditional joint density function of $(\mathbf{X}, Y)$ is

$$F^*(\mathbf{x},y) = P(\mathbf{X}\le\mathbf{x}, Y\le y|Y\ge T) = \alpha^{-1}\int_{\mu\le\mathbf{x}}\int_{a_G\le\omega\le y}G(\omega)F(d\mu,d\omega),$$

which yields that

$$f^*(\mathbf{x},y) = F^*(d\mathbf{x},dy) = \alpha^{-1}G(y)F(d\mathbf{x},dy) = \alpha^{-1}G(y)f(\mathbf{x},y).$$

In addition, for each i, since $K_h^{(1)}(\varepsilon_i)\le|K_h^{(1)}(\varepsilon_i)|$, we have $\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i)\le\frac{1}{G(U_i)}|K_h^{(1)}(\varepsilon_i)|$. Therefore, summing over i, we obtain $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i)\le\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}|K_h^{(1)}(\varepsilon_i)|$. By the Law of Large Numbers and under conditions C3-C5, we can derive the following upper bound

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i)\le\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}|K_h^{(1)}(\varepsilon_i)|$$

$$\begin{aligned}
&\to \int\int \frac{1}{G(y)}|K_h^{(1)}(\varepsilon)|f^*(\mathbf{x},y)d\mathbf{x}dy = \int\int \frac{1}{\alpha}|K_h^{(1)}(\varepsilon)|f(\mathbf{x},y)d\mathbf{x}dy \\
&= \frac{1}{\alpha}\int \mathbb{E}(|K_h^{(1)}(\varepsilon)| \mid \mathbf{x})d\mathbf{x} \le M
\end{aligned}$$

for some constant M coming from controlling $K_h^{(1)}(\varepsilon)$ by its upper bound. Furthermore, according to the argument in the paper, we have

$$\begin{aligned}
|G(U_i) - G_n(U_i)| &\le \sup_{U\ge a_F} |G_n(U_i) - G(U_i)|, \\
G_n(U_i)G(U_i) &\ge (G(a_F) - \sup_{U\ge a_F} |G_n(U_i) - G(U_i)|)G(U_i) \\
&\ge (G(a_F) - \sup_{U\ge a_F} |G_n(U_i) - G(U_i)|)G(a_F),
\end{aligned}$$

where $G(a_F)$ is the lower bound for both $G(U_i)$ and $G_n(U_i)$, and the supremum is the least upper bound of the set $\{|G_n(U) - G(U)| : U \ge a_F\}$. By combining these with the consistency result $\sup_{U\ge a_F} |G_n(U_i) - G(U_i)| = O_p(n^{-1/2})$ from Liang et al. (2011), as $n \to \infty$, we obtain

$$\begin{aligned}
|D_{n1}| &= \Big|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{G_n(U_i)} - \frac{1}{G(U_i)}\right)K_h^{(1)}(\varepsilon_i)\Big| = \Big|\frac{1}{n}\sum_{i=1}^{n}\frac{G(U_i) - G_n(U_i)}{G_n(U_i)G(U_i)}K_h^{(1)}(\varepsilon_i)\Big| \\
&\le \frac{\sup_{U\ge a_F}|G_n(U_i) - G(U_i)|}{G(a_F) - \sup_{U\ge a_F}|G_n(U_i) - G(U_i)|}\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}|K_h^{(1)}(\varepsilon_i)| \\
&\le \frac{\sup_{U\ge a_F}|G_n(U_i) - G(U_i)|}{G(a_F) - \sup_{U\ge a_F}|G_n(U_i) - G(U_i)|}M = o_p(1).
\end{aligned}$$

Regarding D_{n2} associated with the true function $G(U_i)$, by the Law of Large Numbers and under conditions C3-C5, we know that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i) \to \mathbb{E}\left(\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i)\right).$$

By combining this with the joint conditional density $f^*(\mathbf{x},y) = \alpha^{-1}G(y)f(\mathbf{x},y)$, we get

$$\mathbb{E}\left(\frac{1}{G(U_i)}K_h^{(1)}(\varepsilon_i)\right)=\frac{1}{\alpha}\int\mathbb{E}\left(K_h^{(1)}(\varepsilon)\mid\mathbf{x}\right)d\mathbf{x}.$$

Hence, we conclude that $D_{n2}\to\mathbb{E}(K_h^{(1)}(\varepsilon))$.

After that, we apply the central limit theorem to obtain the asymptotic normality of the gradient term. By Theorem 4.4 of He and Yang (2003), we know that for any measurable function $g(\mathbf{x},y)$, if the following finite moment conditions hold: $\int_{-\infty}^{b_F}\frac{dG}{1-F}<\infty$, $\int\frac{dF}{G^2}<\infty$, and $\frac{g^2(\mathbf{x},y)}{G(\mathbf{x})}F(d\mathbf{x},dy)<\infty$, as $n\to\infty$, we can obtain

$$\sqrt{n}\left\{\int g(\mathbf{x},y)F_n(d\mathbf{x},dy)-\int g(\mathbf{x},y)F(d\mathbf{x},dy)\right\}\xrightarrow{d}\mathcal{N}(0,\sigma^2),$$

where $\sigma^2=\alpha\int\frac{(g(\mathbf{x},y)-\mu)^2}{G(\mathbf{x})}F(d\mathbf{x},dy)+\alpha\frac{\tilde{g}^2}{(1-F)G^2}dG$, $\mu=\int g(\mathbf{x},y)F(d\mathbf{x},dy)$, and $\tilde{g}(s)=\int_{\mathbf{x}\le s}(g(\mathbf{x},y)-\mu)F(d\mathbf{x},dy)$. Following that, we can apply the above result to obtain the nominal convergence of $\alpha_n\sum_{i=1}^n\frac{1}{G_n(U_i)}K_h^{(1)}(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i^Tu/\sqrt{n}$. For this purpose, let $(g(\mathbf{x},y)-\mu)^2=K_h^{(1)}(\varepsilon)X_jK_h^{(1)}(\varepsilon)X_k$ and $\tilde{g}^2=\int_{Y\le s}K_h^{(1)}(\varepsilon)\ X_jF(dY,dX)\int_{Y\le s}K_h^{(1)}(\varepsilon)X_kF(dY,dX)$. Thereafter, we have the following asymptotic normal distribution with $\Sigma=(\sigma)_{jk}$

$$\mathcal{W}=\alpha_n\sum_{i=1}^n\frac{1}{G_n(U_i)}K_h^{(1)}(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i^Tu/\sqrt{n}\xrightarrow{d}\mathcal{N}(0,\Sigma),$$

where

$$\begin{aligned}\sigma_{jk}=&\alpha\bigg\{\int\frac{K_h^{(1)}(\varepsilon)X_jK_h^{(1)}(\varepsilon)X_k}{G(Y)}F(dY,dX)\\&+\int\frac{\int_{Y\le s}K_h^{(1)}(\varepsilon)X_jF(dY,dX)\int_{Y\le s}K_h^{(1)}(\varepsilon)X_kF(dY,dX)}{(1-F(s))G^2(s)}G(ds)\bigg\}.\end{aligned}$$

For $S_{n,2} = \frac{1}{2n}\sum_{i=1}^{n}\alpha_n[G_n(U_i)]^{-1}K_h^{(2)}(\varepsilon_i)(\mathbf{x}_i^T\mathbf{u})^2$, based on the result from Lemma 2, we can conclude that $S_{n,2} \xrightarrow{p} \frac{1}{2}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}$. Combining the first-order and second-order terms, we have

$$S_n(\mathbf{u}) = S_{n,1} + S_{n,2} \xrightarrow{d} -\mathcal{W}^T\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\}\mathbf{u}.$$

By the theory of epi-convergence (Geyer, 1994), convergence of functions implies convergence of their maximizers when the limit function is strictly concave with a unique maximizer. Since $S_n(\mathbf{u}) \xrightarrow{d} V(\mathbf{u})$ and both functions are concave with unique maximizers, we have $\arg\max_{\mathbf{u}} S_n(\mathbf{u}) \xrightarrow{d} \arg\max_{\mathbf{u}} V(\mathbf{u})$. Because $V(\mathbf{u})$ has a unique maximizer, by taking $\partial V(\mathbf{u})/\partial\mathbf{u} = 0$, we can obtain the unique maximizer $\mathbf{u} = (\mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{x}]\mathbf{x}\mathbf{x}^T\})^{-1}\mathcal{W}$. We then complete the proof.

□

Proof of Theorem 2

Let $M_n(\boldsymbol{\beta}) = \int K_h(y-\mathbf{x}^T\boldsymbol{\beta})F_n(dy, d\mathbf{x}) - \int K_h(y-\mathbf{x}^T\boldsymbol{\beta}_0)F_n(dy, d\mathbf{x})$. Then, the maximizers of M_n are identical to those of $Q_n(\boldsymbol{\beta})$ defined in the paper, since M_n is a shift of $Q_n(\boldsymbol{\beta})$ by a constant term independent of $\boldsymbol{\beta}$. By Theorem 3.2 of He and Yang (2003), we can obtain

$$\int \phi(y,\mathbf{x})F_n(dy, d\mathbf{x}) \to \int \phi(y,\mathbf{x})F(dy, d\mathbf{x}) \text{ almost surely,}$$

for any measurement function $\phi(\cdot)$. Note that this is a uniform convergence result for empirical processes based on Glivenko-Cantelli-type theorems, which ensures that the empirical distribution F_n converges to the true distribution F uniformly over measurable functions.

With the above result, by defining $\phi(y, \mathbf{x}) = K_h(y - \mathbf{x}^T\boldsymbol{\beta}) - K_h(y - \mathbf{x}^T\boldsymbol{\beta}_0)$, for any fixed $\boldsymbol{\beta} \in \mathbb{R}^p$, we have $M_n(\boldsymbol{\beta}) \to M(\boldsymbol{\beta})$ almost surely, where $M(\boldsymbol{\beta}) = \mathbb{E}[K_h(y - \mathbf{x}^T\boldsymbol{\beta}) - K_h(y - \mathbf{x}^T\boldsymbol{\beta}_0)]$ is the population counterpart. To demonstrate the consistency of the estimator, we analyze the behavior of the population function $M(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$. Under condition C5, we can get

$$\frac{dM(\boldsymbol{\beta})}{d\boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \mathbb{E}[\mathbf{x}K_h^{(1)}(y - \mathbf{x}^T\boldsymbol{\beta})]|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = 0,$$

$$\frac{d^2M(\boldsymbol{\beta})}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T} = \mathbb{E}[\mathbf{x}K_h^{(2)}(y - \mathbf{x}^T\boldsymbol{\beta})\mathbf{x}^T] \geq 0.$$

The strict inequality holds for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ in a neighborhood of $\boldsymbol{\beta}_0$. Thus, $\boldsymbol{\beta}_0$ is the unique maximizer for $M(\boldsymbol{\beta})$.

To establish the consistency, let Ω be any compact set such that for all $\boldsymbol{\beta}$ in Ω, $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \gamma < \infty$. Because of the concave of $M_n(\boldsymbol{\beta})$ in $\boldsymbol{\beta}$, by the concavity and the results from Lemma 1, we have

$$\sup_{\boldsymbol{\beta}\in\Omega}|M_n(\boldsymbol{\beta}) - M(\boldsymbol{\beta})| \to 0 \text{ almost surely},$$

which implies that $M_n(\boldsymbol{\beta})$ uniformly converges to $M(\boldsymbol{\beta})$ over Ω. Since $M(\boldsymbol{\beta}_0) > M(\boldsymbol{\beta})$ for $\boldsymbol{\beta} \in \Omega$, $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, and $M(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$, Lemma 1 implies with probability one that for large enough n, $M_n(\boldsymbol{\beta}_0)$ is greater than $M_n(\boldsymbol{\beta})$ for any value on the boundary of Ω. Therefore, with probability one, M_n contains a local maximum in Ω for sufficiently large n. By the definition of $\hat{\boldsymbol{\beta}}$, this implies that with probability one, $\hat{\boldsymbol{\beta}}$ is eventually in Ω. Since Ω can be chosen to be arbitrarily small, it follows that $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$ almost surely. By following the results of Lemma 2 and Lemma 3, we can straightforwardly demonstrate the asymptomatic normality.

□

Proof of Theorem 3

Denote $\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0$ with $\mathbf{v}_0$ being a unit vector. To prove Theorem 3, following Owen (1990), we first need to prove that

$$\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| \le \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \max_{1\le i\le n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = O_p(n^{-1/2})o_p(n^{1/2}) = o_p(1)$$

to provide support for the following Taylor expansion. It means that we need to demonstrate $\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| = O_p(n^{-1/2})$ and $\max_{1\le i\le n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = o_p(n^{1/2})$.

Let $\Xi_i(\boldsymbol{\beta}_0) = G^{-1}(U_i)K_h^{(1)}(\varepsilon_i)\mathbf{W}_i$. We first prove $\max_{1\le i\le n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = o(n^{1/2})$ almost surely. At first, according to the construction, we know that $|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| \ge 0$ are i.i.d. random variables due to the independence of the data. Also, based on kernel conditions listed in C5, it is evident that $\mathbb{E}(|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|^2) < \infty$ because $G^{-1}(U_i)$ and $\mathbf{W}_i$ are also well-behaved in terms of moments. Since $\mathbb{E}(|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|^2) < \infty$, the probability that

$|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)|^2$ exceeds n decays sufficiently fast. Specifically, we have $\sum_{i=1}^{\infty} P(|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)|^2 > n) < \infty$, which implies that $|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)|^2 > n$ occurs only finitely often with probability one. Now, consider $\sum P(|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| > n^{1/2})$. By applying the same reasoning, the sum of these probabilities is finite $\sum P(|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| > n^{1/2}) < \infty$. Therefore, by the Borel-Cantelli Lemma (used to control the growth of sums of probabilities for large n), we can conclude that $|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| > n^{1/2}$ occurs only finitely often with probability one. If $|\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| > n^{1/2}$ happens only finitely often, it implies that for sufficiently large n, the maximum $\max_{1 \le i \le n} |\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)|$ cannot exceed $n^{1/2}$ for large n. The above argument holds for any $A > 0$ such that $\max_{1 \le i \le n} |\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| > A n^{1/2}$ occurs finitely often. Therefore, we achieve

$$\limsup \max_{1 \le i \le n} |\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| / n^{1/2} \le A$$

with probability one. The above inequality holds simultaneously with probability one for any countable set of values for A, so $\max_{1 \le i \le n} |\mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)| = o(n^{1/2})$ with probability one.

We then prove $\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| = O_p(n^{-1/2})$. Since $\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}$ is the solution of $\frac{1}{n} \sum_{i=1}^{n} \frac{\Xi_i(\boldsymbol{\beta}_0)}{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T \Xi_i(\boldsymbol{\beta}_0)} = 0$, we can have the following equations

$$\begin{aligned} 0 &= \Big\| \frac{1}{n} \sum_{i=1}^{n} \frac{\Xi_i(\boldsymbol{\beta}_0)}{1 + \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)} \Big\| \ge \Big| \frac{\mathbf{v}_0^T}{n} \sum_{i=1}^{n} \frac{\Xi_i(\boldsymbol{\beta}_0)}{1 + \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)} \Big| \\ &= \Big| \frac{\mathbf{v}_0^T}{n} \sum_{i=1}^{n} \Xi_i(\boldsymbol{\beta}_0) \left(1 - \frac{\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)}{1 + \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \mathbf{v}_0^T \Xi_i(\boldsymbol{\beta}_0)} \right) \Big| = \Big| \frac{\mathbf{v}_0^T}{n} \sum_{i=1}^{n} \Xi_i(\boldsymbol{\beta}_0) \end{aligned}$$

$$+ \frac{\mathbf{v}_0^T}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\frac{G(U_i)-G_n(U_i)}{G_n(U_i)} - \frac{\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0^T}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\mathbf{v}_0}{1+\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)}\Bigg[1+$$

$$2\frac{G(U_i)-G_n(U_i)}{G_n(U_i)} + \left(\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}\right)^2\Bigg]\Bigg| \geq \frac{\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0^T\Phi_n(\boldsymbol{\beta}_0)\mathbf{v}_0}{1+\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\max_{1\leq i\leq n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|}$$

$$-\left|\mathbf{v}_0^T\frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right| - \frac{\sup_{U\geq a_F}|G_n(U_i)-G(U_i)|}{G(a_F)-\sup_{U\geq a_F}|G_n(U_i)-G(U_i)|}$$

$$\left(\frac{1}{n}\sum_{i=1}^n|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| + \frac{2\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0^T\Phi_n(\boldsymbol{\beta}_0)\mathbf{v}_0}{1+\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\max_{1\leq i\leq n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|}\right)$$

$$-\left(\frac{\sup_{U\geq a_F}|G_n(U_i)-G(U_i)|}{G(a_F)-\sup_{U\geq a_F}|G_n(U_i)-G(U_i)|}\right)^2\frac{\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\mathbf{v}_0^T\Phi_n(\boldsymbol{\beta})\mathbf{v}_0}{1+\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\max_{1\leq i\leq n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|},$$

where $\Phi_n(\boldsymbol{\beta}_0) = \sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)$.

According to Lemma 3 and Theorem 4.4 of He and Yang (2003), we can get $\alpha_n\sum_{i=1}^n \frac{1}{G_n(U_i)}K_h^{(1)}(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0)\mathbf{x}_i^T\mathbf{v_0}/\sqrt{n} \xrightarrow{d} \mathcal{N}(0,\Sigma)$. Combining with the Law of Large Numbers, we know that $\Phi_n(\boldsymbol{\beta}_0)$ converges to a positive definite matrix as $n\to\infty$. These imply that there exists a constant $c>0$ such that $P(\mathbf{v}_0^T\Phi_n(\boldsymbol{\beta}_0)\mathbf{v}_0) > c \to 1$ as $n\to\infty$. In view of (3.5) and condition C5, by using $\mathbb{E}[K_h^{(1)}(\varepsilon)\mid\mathbf{X}] = 0$, we have $\mathbb{E}(n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)) = \mathbb{E}(K_h^{(1)}(\varepsilon_i)\mathbf{W}_iG^{-1}(U_i)) = 0$ and $\mathrm{Var}(n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)) = O(n^{-1})$. Therefore, we can achieve $n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) = O_p(n^{-1/2})$. Also, according to the results in Liang et al. (2011), we know that $\sup_{U\geq a_F}|G_n(U)-G(U)| = O_p(n^{-1/2})$. Then, we can obtain that

$$\frac{\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|}{1+\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\max_{1\leq i\leq n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)|} = O_p(n^{-1/2}).$$

We already prove that $\max_{1\leq i\leq n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = o(n^{1/2})$ with probability one.

We therefore can conclude that $\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| = O_p(n^{-1/2})$. Based on the above procedures, we complete the proof of $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| \le \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \max_{1\le i\le n}$ $|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = O_p(n^{-1/2})o_p(n^{1/2}) = o_p(1)$.

The preceding result implies that the upcoming Taylor expansion is valid. Following Owen (1990), applying the second-order Taylor expansion on $(1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^{-1}$ for i from 1 to n and approximating the high-order terms by factoring them into the $\frac{(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}$ form, we can obtain from $\frac{1}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)} = 0$ that

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)} = \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\left(1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0) + \frac{(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}\right)\\
&= \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) - \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} + \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\frac{(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}.
\end{aligned}$$

The above expansion gives

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) + \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1} r_n(\boldsymbol{\beta}_0),$$

where $r_n(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)(1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^{-1}\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\}^2$. It is obvious that $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| = o_p(1)$ based on the previous proofs. Thus,

$$\begin{aligned}
|r_n(\boldsymbol{\beta}_0)| &\le \max_{1\le i\le n}\|\Xi_i(\boldsymbol{\beta}_0)\|(1-\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|)^{-1}\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\\
&= o_p(n^{1/2})O_p(n^{-1}) = o_p(n^{-1/2}).
\end{aligned}$$

Therefore, we have

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) + o_p(n^{-1/2}).$$

The above equation can also be proved by the following way. From $\frac{1}{n}\sum_{i=1}^n$

$\frac{\Xi_i(\boldsymbol{\beta}_0)}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)} = 0$, we write

$$0 = \frac{1}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)} = \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\left(1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)+\frac{(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}\right)$$
$$= \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) - \Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} + \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}$$
$$-\frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\left[2\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}+\left(\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}\right)^2\right]$$
$$+\frac{1}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}\left(1+\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}\right)^3.$$

At the same time, we know that

$$\left\|\frac{1}{n}\sum_{i=1}^n \frac{\Xi_i(\boldsymbol{\beta}_0)(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2}{1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)}\right\| \le \frac{1}{n}\sum_{i=1}^n \|\Xi_i(\boldsymbol{\beta}_0)\|^3\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|^2|1-\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|^{-1}$$
$$= o_p(n^{-1/2}).$$

Thus, from $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| \le \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \max_{1\le i\le n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = O_p(n^{-1/2})$ $o_p(n^{1/2}) = o_p(1)$ and the consistency result $\sup_{U\ge a_F}|G_n(U)-G(U)| = O_p(n^{-1/2})$ from Liang et al. (2011), we can achieve the result.

Similarly, by the third order Taylor expansion on $\log(1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))$ for all i, we can approximate $\mathcal{L}(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\}$ as

$$\mathcal{L}(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n\left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\right\}^2 + \frac{2}{3}\sum_{i=1}^n\left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\right\}^3(1+\eta_i^*)^{-3},$$

where η_i^* lies between 0 and $\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)$. Since η_i^* lies in the interval $[0, \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)]$, we know that $0 \le \eta_i^* \le \max_{1\le i\le n}|\eta_i^*|$. Thus, $(1+\eta_i^*)^{-3}$ is bounded from above by $(1+\eta_i^*)^{-3} \le (1-\max_{1\le i\le n}|\eta_i^*|)^{-3}$. By the previous result $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| \le \|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\| \max_{1\le i\le n}|\mathbf{v}_0^T\Xi_i(\boldsymbol{\beta}_0)| = O_p(n^{-1/2})o_p(n^{1/2})$ $= o_p(1)$, this upper bound remains well-behaved as a constant independent of n. Also, using matrix notation, we can rewrite the quadratic sum as

$$\sum_{i=1}^n(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2 = n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\left(n^{-1}\sum_{i=1}^n\Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\right)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0},$$

where $\Phi_n(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n\Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)$. Then, we can calculate that

$$\Big|\sum_{i=1}^n\left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\right\}^3(1+\eta_i^*)^{-3}\Big| \le \sum_{i=1}^n|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|^3(1-\max_{1\le i\le n}|\eta_i^*|)^{-3}$$
$$\le \max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|(1-\max_{1\le i\le n}|\eta_i^*|)^{-3}\sum_{i=1}^n|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|^2$$
$$= \max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|(1-\max_{1\le i\le n}|\eta_i^*|)^{-3}n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = o_p(1)O_p(1) = o_p(1),$$

where according to the previous calculations, $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| = o_p(1)$, $(1-\max_{1\le i\le n}|\eta_i^*|)^{-3}$ is bounded as a constant, and $n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \sum_{i=1}^n(\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0))^2 = O_p(1)$ is controlled by $\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)| = o_p(1)$ and the convergence of $\Phi_n(\boldsymbol{\beta}_0)$. The above equation can also be proved by the following way. By Taylor's expansion, we have

$$\mathcal{L}(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n\left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\right\}^2$$
$$+ 2\sum_{i=1}^n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}$$
$$- \sum_{i=1}^n\left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)\right\}^2\left[2\frac{G(U_i)-G_n(U_i)}{G_n(U_i)} + \left(\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}\right)^2\right]$$
$$+ O\left(\|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}\|\max_{1\le i\le n}|\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T\Xi_i(\boldsymbol{\beta}_0)|\sum_{i=1}^n\Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\left(1+\frac{G(U_i)-G_n(U_i)}{G_n(U_i)}\right)^2\right).$$

Following the previous arguments, by incorporating $\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1} \frac{1}{n} \sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) + o_p(n^{-1/2})$, we obtain $\mathcal{L}(\boldsymbol{\beta}_0) = 2\sum_{i=1}^n \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n \left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0)\right\}^2 + o_p(1)$. Then, the rest task is to prove chi-square distribution for $\mathcal{L}(\boldsymbol{\beta}_0)$.

Define $\overline{\Xi} = n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)$. Rewriting $\mathcal{L}(\boldsymbol{\beta}_0)$ in a matrix format,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}_0) &= 2\sum_{i=1}^n \log\left\{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0)\right\} = 2\sum_{i=1}^n \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n \left\{\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0)\right\}^2 \\ &+ o_p(1) = 2\sum_{i=1}^n \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0) - \sum_{i=1}^n \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} + o_p(1) \\ &= 2n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \overline{\Xi} - n\boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Phi_n(\boldsymbol{\beta}_0)\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} + o_p(1) = 2n\overline{\Xi}^T\{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\overline{\Xi} - \\ &\quad n\overline{\Xi}^T\{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\Phi_n(\boldsymbol{\beta}_0)\{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\overline{\Xi} + o_p(1) = n\overline{\Xi}^T\{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\overline{\Xi} \\ &+ o_p(1) = \left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\}^T \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\} + o_p(1),\end{aligned}$$

where $\Phi_n(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)$, $\boldsymbol{\lambda}_{\boldsymbol{\beta}_0} = \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1} \frac{1}{n}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0) + o_p(n^{-1/2}) = \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\overline{\Xi} + o_p(n^{-1/2})$, and $\Xi_i(\boldsymbol{\beta}_0) = G_n^{-1}(U_i)K_h^{(1)}(\varepsilon_i)\mathbf{W}_i$. According to Lemma 3 and Theorem 4.4 of He and Yang (2003), we have the asymptotic normality $\sum_{i=1}^n \frac{1}{G_n(U_i)}K_h^{(1)}(\varepsilon_i)\mathbf{W}_i/\sqrt{n} = \sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)/\ \sqrt{n} \xrightarrow{d} \mathcal{N}(0, \Sigma^*)$, where the covariance matrix $\Sigma^* = \lim_{n\to\infty} \mathrm{Var}(n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0))$.

According to the theory of quadratic forms, if $\mathbf{Z} \sim \mathcal{N}(0, \Sigma^*)$, the quadratic form $\mathbf{Z}^T\Sigma^{*-1}\mathbf{Z} \sim \chi_p^2$, where p is the dimension of $\mathbf{Z}$. Based on these arguments, let $\mathbf{Z} = \sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)/\sqrt{n}$, we can have

$$\left(\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)/\sqrt{n}\right)^T \left(\lim_{n\to\infty} \mathrm{Var}(n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0))\right)^{-1} \left(\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)/\sqrt{n}\right) \sim \chi_p^2.$$

In addition, we know that $\Xi_i(\boldsymbol{\beta}_0)$ are independent random variables with $\mathbb{E}(\Xi_i(\boldsymbol{\beta}_0)) = 0$ and covariance matrix $\Phi_n(\boldsymbol{\beta}_0) = n^{-1}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\Xi_i^T(\boldsymbol{\beta}_0)$. As $n \to \infty$, the sample covariance matrix $\Phi_n(\boldsymbol{\beta}_0)$ converges in probability to the true asymptotic covariance matrix Σ^* with the regularity conditions C1-C7. Therefore, according to the above equation, as $n \to \infty$, we obtain

$$\left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\}^T \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\} \xrightarrow{d} \chi_p^2.$$

This completes the proof with

$$\begin{aligned}\boldsymbol{\mathcal{L}}(\boldsymbol{\beta}_0) &= 2\sum_{i=1}^n \log\left\{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}_0}^T \Xi_i(\boldsymbol{\beta}_0)\right\}\\ &= \left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\}^T \{\Phi_n(\boldsymbol{\beta}_0)\}^{-1}\left\{n^{-1/2}\sum_{i=1}^n \Xi_i(\boldsymbol{\beta}_0)\right\} + o_p(1) \xrightarrow{d} \chi_p^2.\end{aligned}$$

□

Proof of Theorem 4

Recall that $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{1,0}^T, \boldsymbol{\beta}_{2,0}^T)^T$ is the true parameter vector, where $\boldsymbol{\beta}_{1,0}$ corresponds to the significant variables, and $\boldsymbol{\beta}_{2,0} = \mathbf{0}_{(p-s)\times 1}$ corresponds to the insignificant variables. Denote $\boldsymbol{\xi} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$, $\hat{\boldsymbol{\xi}} = \sqrt{n}(\hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta}_0)$, and $\hat{\boldsymbol{\xi}}_1 = \sqrt{n}(\hat{\boldsymbol{\beta}}_1^p - \boldsymbol{\beta}_{1,0})$. Then, $\hat{\boldsymbol{\beta}}^p$ is the maximizer of the following penalized function

$$\alpha_n \sum_{i=1}^n \frac{1}{G_n(U_i)} K_h(U_i - \mathbf{W}_i^T\boldsymbol{\beta}) + n\sum_{j=1}^p p_\lambda^{(1)}\left(|\beta_j^{(0)}|\right)\operatorname{sgn}(\beta_j^{(0)})(\beta_j - \beta_{0j}).$$

Notice that $U_i - \mathbf{W}_i^T\boldsymbol{\beta} = U_i - \mathbf{W}_i^T\boldsymbol{\beta}_0 - \mathbf{W}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = \varepsilon_i - \frac{1}{\sqrt{n}}\mathbf{W}_i^T\boldsymbol{\xi}$, where $\varepsilon_i = U_i - \mathbf{W}_i^T\boldsymbol{\beta}_0$. Thus, the objective function becomes

$$\begin{aligned} Q_n(\boldsymbol{\xi}) =&\alpha_n \sum_{i=1}^{n} \frac{1}{G_n\left(U_i\right)} \{K_h(\varepsilon_i - \mathbf{W}_i^T \boldsymbol{\xi}/\sqrt{n}) - K_h(\varepsilon_i)\} \\ &+ n \sum_{j=1}^{p} p_\lambda^{(1)} \left(|\beta_j^{(0)}|\right) \operatorname{sgn}(\beta_j^{(0)})(\beta_j - \beta_{0j}). \end{aligned}$$

Note that the second term (the penalty term) converges to zero when the true parameters are within the correct model. Specifically, we have

$$n \sum_{j=1}^{p} p_\lambda^{(1)} \left(|\beta_j^{(0)}|\right) \operatorname{sgn}(\beta_j^{(0)})(\beta_j - \beta_{0j}) \xrightarrow{p} \begin{cases} 0 & \text{if } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{2,0}, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, we obtain $\hat{\boldsymbol{\beta}}_2^p \xrightarrow{p} \mathbf{0}$, meaning that the penalty drives the irrelevant variables to zero.

To further analyze the asymptotic properties, we define $A_n = \frac{\alpha_n}{n} \sum_{i=1}^n \frac{1}{G_n(U_i)} K_h^{(2)}(\varepsilon_i) \mathbf{W}_i \mathbf{W}_i^T$ and $B_n = \frac{\alpha_n}{\sqrt{n}} \sum_{i=1}^n \frac{1}{G_n(U_i)} K_h^{(1)}(\varepsilon_i)\ \mathbf{W}_i$. Then, the penalized objective function can be approximated as $Q_n(\boldsymbol{\xi}) \approx \dfrac{1}{2} \boldsymbol{\xi}^T A \boldsymbol{\xi} - B_n^T \boldsymbol{\xi} + n \sum_{j=1}^p p_\lambda^{(1)}(|\beta_j^{(0)}|) \operatorname{sgn}(\beta_j^{(0)})(\beta_j - \beta_{0j}) + o_p(1)$. We also have $A = \mathbb{E}(A_n) = \mathbb{E}\{[K_h^{(2)}(\varepsilon) \mid \mathbf{X}] \mathbf{X}\mathbf{X}^T\}$. Denote $B_{n,11}$ be the upper-left $s \times s$ submatrix of B_n. Note that $\hat{\boldsymbol{\xi}}$ is the maximizer of the $Q_n(\boldsymbol{\xi})$, which can be written asymptotically as

$$\begin{aligned} &Q_n((\boldsymbol{\xi}_1^T, 0^T)^T) = \frac{1}{2} (\boldsymbol{\xi}_1^T, 0^T) A (\boldsymbol{\xi}_1^T, 0^T)^T - B_n^T (\boldsymbol{\xi}_1^T, 0^T)^T \\ &+ n \sum_{j=1}^{p} p_\lambda^{(1)} \left(|\beta_j^{(0)}|\right) \operatorname{sgn}(\beta_j^{(0)})(\beta_j - \beta_{0j}) + o_p(1) \to Q(\boldsymbol{\xi}) = \frac{1}{2} \boldsymbol{\xi}^T \Sigma_1 \boldsymbol{\xi}_1 - B_{n,11}^T \boldsymbol{\xi}_1. \end{aligned}$$

Since $L_n(\boldsymbol{\xi})$ is a concave function of $\boldsymbol{\xi}$ and $L(\boldsymbol{\xi}_1)$ has a unique maximizer, the epi-convergence theory of Geyer (1994) implies that

$$\arg\max Q_n(\boldsymbol{\xi}) = \sqrt{n}(\hat{\boldsymbol{\beta}}^p - \boldsymbol{\beta}_0) \xrightarrow{d} \arg\max Q(\boldsymbol{\xi}_1),$$

which establishes the asymptotic normality part.

To prove the consistency of model selection, we need to show $\hat{\boldsymbol{\beta}}_2^p = \mathbf{0}_{(p-s)\times 1}$ with probability tending to one. It is equivalent to prove that for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{1,0}\| = O_p(n^{-1/2})$ and any constant C, we have

$$Q_n^p\{(\boldsymbol{\beta}_1, \mathbf{0})^T\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q_n^p\{(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T\}.$$

According to Fan and Li (2001), for $\beta_j \neq 0$ and $j = s+1, \cdots, d$, we have

$$\frac{dL_p(\boldsymbol{\beta})}{d\beta_j} = -n\lambda \Big\{ \lambda^{-1} p_\lambda^{(1)}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p\Big(\frac{1}{\sqrt{n}\lambda}\Big) \Big\}.$$

Since $\liminf_{n\to\infty} \liminf_{t\to 0^+} p_\lambda^{(1)}(t)/\lambda > 0$ and $n^{1/2}\lambda \to \infty$, the sign of the derivative for $\beta_j \in (-Cn^{-1/2}, Cn^{1/2})$ is completely determined by that of β_j. This implies that the maximum of $Q_n^p(\boldsymbol{\beta})$ occurs at $\beta_j = 0$ for $j = s+1, \ldots, p$. Therefore, we conclude that $\hat{\boldsymbol{\beta}}_2^p = \mathbf{0}_{(p-s)\times 1}$ with probability tending to one, proving the consistency of model selection.

□

Proof of Theorem 5

To demonstrate the consistency of the extended BIC selection, that is, the probability of the selected model being equal to the true model asymptotically approaches one, we can follow Wang et al. (2007) to study the BIC corresponding to estimators that fail to select all of the significant variables and estimators that select too many variables. We outline the procedure here.

Suppose that S_T denotes the true model, i.e., the set of indices corresponding to the true significant covariates, S_λ indicates the set of the indices of the covariates selection by the penalized kernel mode-based regression with tuning parameter λ, $\Omega_- = \{\lambda : S_\lambda \not\supset S_T\}$ denotes the under-fitted models, i.e., models that fail to include all the significant variables, and $\Omega_+ = \{\lambda : S_\lambda \not\supseteq S_T\}$ represents the over-fitted models, i.e., models that include additional insignificant variables. We construct a sequence of reference tuning parameters $\lambda_n = \log(n)/\sqrt{n}$ (i.e., $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$). Then, one can verify that $P\left(\inf_{\lambda \in \Omega_- \cup \Omega_+} BIC_\lambda > BIC_{\lambda_n}\right) \to 1$ under mild conditions. Particularly, for under-fitted models, these models are misspecified and will have a larger bias. Also, the lack of significant variables leads to a poorer fit, increasing the residual sum of squares. Therefore, we have $P(\inf_{\lambda \in \Omega_-} BIC_\lambda > BIC_{\lambda_n}) \to 1$. For over-fitted models, these models may fit the data slightly better due to additional parameters, but they suffer from overfitting. The BIC includes a penalty term for model complexity proportional to $\log(n)$ times the number of parameters. The additional insignificant variables increase the penalty without substantially improving the fit. Therefore, $P(\inf_{\lambda \in \Omega_+} BIC_\lambda > BIC_{\lambda_n}) \to 1$. This means that we cannot asymptotically choose a λ that identifies an over-fitted or under-fitted model.

Because the penalized mode-based estimator $\hat{\boldsymbol{\beta}}^p_{\lambda_n}$ with $\lambda_n = \log(n)/\sqrt{n}$ is exactly the same as the oracle estimator, it follows immediately that $P(BIC_{\lambda_n} = BIC_{S_T}) \to 1$. As a result, we have $P(S_{\lambda_{n,opt}} = S_T) \to 1$, indicating that if the true model is contained within the set of candidate models, it can be guaranteed to be selected by the proposed BIC method.

□

Proof of Theorem S1

Let $\Theta_{\boldsymbol{\beta}}$ denote the set of limit points of the sequence $\{\hat{\boldsymbol{\beta}}^{p(m)}\}$. That is,

$$\Theta_{\boldsymbol{\beta}} = \left\{\hat{\boldsymbol{\beta}}^* : \hat{\boldsymbol{\beta}}^{p(g_m)} \to \hat{\boldsymbol{\beta}}^* \text{ for some subsequence } \{\hat{\boldsymbol{\beta}}^{p(g_m)}\} \subseteq \{\hat{\boldsymbol{\beta}}^{p(m)}\}\right\}.$$

Let $\hat{\boldsymbol{\beta}}^* \in \Theta_{\boldsymbol{\beta}}$. Then, there exists a subsequence $\{\hat{\boldsymbol{\beta}}^{p(g_m)}\}$ such that $\hat{\boldsymbol{\beta}}^{p(g_m)} \to \hat{\boldsymbol{\beta}}^*$ as $m \to \infty$. Since the algorithm is a hill-climbing algorithm, it produces a sequence where the penalized objective function values are non-decreasing $Q^p_n(\hat{\boldsymbol{\beta}}^{p(g_{m+1})}) \geq Q^p_n(\hat{\boldsymbol{\beta}}^{p(g_m)})$. This implies that $\{Q^p_n(\hat{\boldsymbol{\beta}}^{p(g_m)})\}$ is a non-decreasing sequence bounded above. Therefore, the sequence converges $\lim_{m\to\infty} Q^p_n(\hat{\boldsymbol{\beta}}^{p(g_m)}) = Q^p_n(\hat{\boldsymbol{\beta}}^*)$.

By the nature of the algorithm, each new estimate is obtained by applying the mapping M to the previous estimate, i.e., $\hat{\boldsymbol{\beta}}^{p(g_{m+1})} = M(\hat{\boldsymbol{\beta}}^{p(g_m)})$. Using the monotonicity and the mapping M, we have

$$Q_n^p(\hat{\boldsymbol{\beta}}^{p(g_{m+1})}) = Q_n^p(M(\hat{\boldsymbol{\beta}}^{p(g_m)})) \geq Q_n^p(\hat{\boldsymbol{\beta}}^{p(g_m)}).$$

Taking limits as $m \to \infty$, we get

$$\lim_{m\to\infty} Q_n^p(\hat{\boldsymbol{\beta}}^{p(g_{m+1})}) = \lim_{m\to\infty} Q_n^p(M(\hat{\boldsymbol{\beta}}^{p(g_m)})) = Q_n^p(\hat{\boldsymbol{\beta}}^*).$$

Assuming that $M(\boldsymbol{\beta})$ is continuous at $\hat{\boldsymbol{\beta}}^*$, we have

$$\lim_{m\to\infty} M(\hat{\boldsymbol{\beta}}^{p(g_m)}) = M\left(\lim_{m\to\infty} \hat{\boldsymbol{\beta}}^{p(g_m)}\right) = M(\hat{\boldsymbol{\beta}}^*).$$

Therefore, we have

$$Q_n^p(\hat{\boldsymbol{\beta}}^*) = Q_n^p\left(\lim_{m\to\infty} M(\hat{\boldsymbol{\beta}}^{p(g_m)})\right) = Q_n^p(M(\hat{\boldsymbol{\beta}}^*)),$$

which completes the proof.

□

Bibliography

Chen, C. H., Tsai, W. Y., and Chao, W. H. (1996). The Product-Limit Correlation Coefficient and Linear Regression for Truncated Data. *Journal of the American Statistical Association*, 91 (435), 1181-1186.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96 (456), 1348-1360.

He, S. and Yang, G. L. (1998). Estimation of the Truncation Probability in the Random Truncation Model. *The Annals of Statistics*, 26 (3), 1011-27.

He, S. and Yang, G. L. (2003). Estimation of Regression Parameters with Left Truncated Data. *Journal of Statistical Planning and Inference*, 117 (1), 99-122.

Geyer, C. J. (1994). On the Asymptotics of Constrained M-Estimation. *The Annals of Statistics*, 22 (4), 1993-2010.

Gross, S. T. and Lai, T. L. (1996). Nonparametric Estimation and Regression Analysis with Left-Truncated and Right-Censored Data. *Journal of the American Statistical Association*, 91 (435), 1166-1180.

Lai, T. L. and Ying, Z. (1992). Asymptotic Theory of A Bias-Corrected Least Squares Estimator in Truncated Regression. *Statistica Sinica*, 2 (2), 519-539.

Liang, H.-Y., Una-Aivarez, J., and Iglesias-Perez, M. C. (2011). Local Polynomial Estimation of A Conditional Mean Function with Dependent Truncated Data. *TEST*, 20, 653-677.

Lim, Y. and Oh, H. (2014). Variable Selection in Quantile Regression When the Models Have Autoregressive Errors. *Journal of the Korean Statistical Society*, 43 (4), 513-530.

Owen, A. B. (1990). Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18 (1), 90-120.

Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression Estimators. *Econometric Theory*, 7 (2), 186-199.

Shen, P.-S. (2005). Estimation of the Truncation Probability with Left-Truncated and Right-Censored Data. *Journal of Nonparametric Statistics*, 17 (8), 957-969.

Stute, W. (1993). Almost Sure Representations of the Product-Limit Estimator for Truncated Data. *The Annals of Statistics*, 21 (1), 146-156.

Su, Y.-R. and Wang, J.-L. (2012). Modeling Left-Truncated and Right-Censored Survival Data with Longitudinal Covariates. *The Annals of Statistics*, 40 (3), 1465-1488.

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*, 185 (3), 1424-1453.

Ullah, A., Wang, T., and Yao, W. (2023). Semiparametric Partially Linear Varying Coefficient Modal Regression. *Journal of Econometrics*, 235 (2), 1001-1026.

van der Vaart, A. W. (1998). Asymptotic Statistics. *Cambridge University Press, Cambridge, U.K.*

Wang, K. and Li, S. (2021). Robust Distributed Modal Regression for Massive Data. *Computational Statistics* & *Data Analysis*, 160, 107225.

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

Wang, T. (2024). Nonlinear Kernel Mode-Based Regression for Dependent Data. *Journal of Time Series Analysis*, 45 (2), 189-213.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11 (1), 95-103.

Yao, W. and Li, L. (2014). A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, 41 (3), 656-671.

Yao, W., Lindsay, B. G., and Li, R. (2012). Local Modal Regression. *Journal of Nonparametric Statistics*, 24 (3), 647-663.

Zhou, Y. and Yip, P. S. (1999). A Strong Representation of the Product-Limit Estimator for Left Truncated and Right Censored Data. *Journal of Multivariate Analysis*, 69 (2), 261-280.