NONPARAMETRIC REGRESSION UNDER CLUSTER SAMPLING

YUYA SHIMIZU

Abstract. This paper develops a general asymptotic theory for nonparametric kernel regression in the presence of cluster dependence. We examine nonparametric density estimation, Nadaraya-Watson kernel regression, and local linear estimation. Our theory accommodates growing and heterogeneous cluster sizes. We derive asymptotic conditional bias and variance, establish uniform consistency, and prove asymptotic normality. Our findings reveal that under heterogeneous cluster sizes, the asymptotic variance includes a new term reflecting within-cluster dependence, which is overlooked when cluster sizes are presumed to be bounded. We propose valid approaches for bandwidth selection and inference, introduce estimators of the asymptotic variance, and demonstrate their consistency. In simulations, we verify the effectiveness of the cluster-robust bandwidth selection and show that the derived cluster-robust confidence interval improves the coverage ratio. We illustrate the application of these methods using a policy-targeting dataset in development economics.

1. Introduction

Nonparametric regression is widely used in economics for its flexibility. Typically, data are assumed to be independently and identically distributed; however, in reality, observations may exhibit dependence within a group structure called a cluster. Examples of clusters are classrooms, schools, families, hospitals, firms, industries, villages, regions, and so on. The cluster sampling framework assumes independence between observations from different clusters but allows dependence within each cluster.

The previous literature on nonparametric regression under cluster sampling assumes a bounded and homogeneous number of observations per cluster. This assumption may not hold in real data due to heterogeneous cluster sizes. To fill this gap, this paper studies nonparametric kernel regressions that accommodate heterogeneous cluster sizes, including those that grow to infinity asymptotically. Our approach is general, allowing for both bounded and growing clusters simultaneously, and includes cluster-level regressors.

We develop a comprehensive asymptotic theory for nonparametric density estimation, Nadaraya-Watson kernel regression, and local linear estimation. Our results on asymptotic conditional bias and variance, uniform consistency, and asymptotic normality enable us to propose valid methods for bandwidth selection and inference.

This version: June 24, 2024.
I am grateful to Bruce Hansen for his invaluable advice and encouragement. I thank Naoki Aizawa, Harold Chiang, Junho Choi, Jack Collison, Kenta Fukuda, Woosik Gong, Michael Jansson, J.C. Lazzaro, Taisuke Otsu, Jack Porter, Xiaoxia Shi, Gonzalo Vazquez-Bare, and Kohei Yata for helpful comments and suggestions. This paper won the Kanematsu Prize 2023 from the Institute of Economic Management at Kobe University, Japan. I would like to thank the anonymous reviewers of the Kanematsu Prize for their valuable comments. This work is supported by the summer research fellowship from the University of Wisconsin-Madison.

For clusters of growing sizes, the asymptotic variance contains a novel term for within-cluster dependence, which does not appear under the assumption of bounded cluster sizes. This term becomes significant due to the potential for a cluster to contain a growing number of observations within a local neighborhood, making cluster dependence non-negligible asymptotically. We propose consistent estimators of the asymptotic variance that account for cluster dependence and validate its importance through simulation. Our cluster-robust confidence interval achieves improved coverage ratios, while conventional confidence intervals could suffer from under-coverage in our simulated datasets.

Nonparametric regression, while significant on its own, also serves as an intermediate tool for other estimators, such as regression discontinuity design, nonparametric auction estimation, and semiparametric models under cluster sampling. Our results could extend to these areas as well.

Related literature. There is a substantial body of literature on cluster sampling in econometrics. C. Hansen (2007) provides an asymptotic theory for parametric regression with homogeneous cluster sizes. Djogbenou, MacKinnon and Nielsen (2019) and B. Hansen and Lee (2019) extend this theory to heterogeneous cluster sizes. Bugni, Canay, Shaikh and Tabord-Meehan (2022) considers heterogeneous and random cluster sizes for cluster-level randomized experiments. For further literature on parametric models under cluster sampling, the reader can refer to Cameron and Miller (2015) and MacKinnon, Nielsen and Webb (2022).

Conversely, the theory on nonparametric regression under cluster dependence, even with homogeneous cluster sizes, is limited. Lin and Carroll (2000) and Wang (2003) examine local polynomial and local linear regressions, assuming fixed and homogeneous cluster sizes and focusing primarily on asymptotic efficiency. Bhattacharya (2005) offers an asymptotic theory for local constant estimators under multi-stage samples, analogous to cluster sampling. When the number of first-stage strata is set to one, his setup becomes a standard cluster sampling with fixed and homogeneous cluster sizes. He puts a similar structure on error terms as this paper, but the fixed cluster sizes render the term reflecting within-cluster dependence asymptotically negligible. For the regression discontinuity literature, Bartalotti and Brummet (2017) has derived asymptotic theories for local polynomial regression under bounded and homogeneous cluster sizes.

Menzel (2024) proposes a method for estimating nonparametric regressions in the presence of cluster dependence, aiming to extrapolate treatment effects across clusters. He considers *independent* but not identical observations between clusters, with a fixed number of clusters exhibiting *uniformly growing* size. Our approach differs by incorporating general dependence within a cluster and allowing for both bounded and growing cluster sizes simultaneously, leading to distinct asymptotic results and theories.

To the best of our knowledge, there is no literature on nonparametric models with *growing and heterogeneous size* clusters. Our paper adopts the same cluster size framework as Djogbenou *et al.* (2019) and Hansen and Lee (2019). The presence of clusters with growing sizes complicates the proofs for asymptotic theories, as cluster dependence becomes non-negligible. Consequently, this paper introduces new technical results for nonparametric regressions under cluster sampling, notably developing Bernstein's inequality for cluster sampling to demonstrate uniform consistency. These novel contributions are believed to offer valuable theoretical tools for future research.

This research also sheds new light on the literature regarding nonparametric regressions with dependence. Following the foundational work on i.i.d. datasets (e.g., Stone, 1982, Fan, 1992, Ruppert and Wand, 1994), the results have been extended to time series (Robinson, 1983, Hansen, 2008, Kristensen, 2009, Vogt, 2012, Vogt and Linton, 2020) and spatial datasets (Robinson, 2011, Lee and Robinson, 2016), as well as to the cluster dependence framework discussed above.

The remainder of this paper is organized as follows: Section 2 introduces the cluster sampling framework under consideration. Sections 3-5 discuss asymptotic theories for nonparametric density estimators, Nadaraya-Watson estimators, and local linear estimators, respectively. Section 6 demonstrates uniform convergence of these estimators. Section 7 provides guidelines for selecting bandwidth in nonparametric regressions. Section 8 addresses cluster-robust inference. Section 9 presents Monte Carlo simulations for bandwidth selections and inference. Section 10 illustrates our methods with an application in development economics using a dataset by Alatas, Banerjee, Hanna, Olken and Tobias (2012). The paper concludes with Section 11. All proofs, technical lemmas, technical discussions, and additional simulation results are included in the Appendix.

2. Cluster sampling

The researcher observes $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^d$ for $i = 1, \ldots, n$, with cluster sizes given by $n_g \in \{1, 2, \cdots\}$ for $g = 1, \ldots, G$. Here, Y_i represents a dependent variable, and regressors X_i are continuous random variables with the Lebesgue density $f(x)$. Assume that each observation can be grouped into one cluster.[1] Thus, the total number of observations is $n = \sum_{g=1}^{G} n_g$. To explicitly represent the cluster structure, we also use the notation (Y_{gj}, X_{gj}) for $g = 1, \ldots, G$ and $j = 1, \ldots, n_g$. We treat cluster size n_g as nonrandom and possibly heterogeneous across clusters. We assume that observations belonging to different clusters are mutually independent but permit general dependence within the same cluster. We decompose X_{gj} into $X_{gj} = \left(X_{gj}^{(\mathrm{ind})\top}, X_g^{(\mathrm{cls})\top}\right)^\top \in \mathbb{R}^d$ where $X_{gj}^{(\mathrm{ind})} \in \mathbb{R}^{d_{\mathrm{ind}}}$ represents individual-level regressors and $X_g^{(\mathrm{cls})} \in \mathbb{R}^{d_{\mathrm{cls}}}$ represents cluster-level regressors. We assume that the regressors contain at least one individual-level regressors, $d_{\mathrm{ind}} \geq 1$. By construction, $d = d_{\mathrm{ind}} + d_{\mathrm{cls}}$ holds.

We denote $\mathbf{X}_g = \left(X_{g1}, \ldots, X_{gn_g}\right)$ and aim to estimate the nonparametric regression model:

$$Y_{gj} = m\left(X_{gj}\right) + e_{gj}, \tag{1}$$

$$\mathbb{E}\left[e_{gj} \mid \mathbf{X}_g\right] = \mathbb{E}\left[e_{gj} \mid X_{gj}\right] = 0. \tag{2}$$

We also assume

$$\mathbb{E}\left[e_{gj}^2 \mid \mathbf{X}_g\right] = \mathbb{E}\left[e_{gj}^2 \mid X_{gj}\right] = \sigma^2\left(X_{gj}\right), \tag{3}$$

$$\begin{aligned} \mathbb{E}\left[e_{gj}e_{g\ell} \mid \mathbf{X}_g\right] &= \mathbb{E}\left[e_{gj}e_{g\ell} \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right] \\ &= \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right) \text{ for } j \neq \ell. \end{aligned} \tag{4}$$

[1] Formally, we assume that for any i, we know a function $g(i) \in \{1, \cdots, G\}$.

The model specified through (1)-(4) exhibits greater flexibility than initially apparent. The constraint imposed by (3) is that the conditional variance of the error term for an individual is dependent only on the individual's own regressors, both at the individual and cluster levels. Additionally, (4) states that the conditional covariance of the error terms between any two individuals within the same cluster is a function only of their individual-level regressors and shared cluster-level regressors. This framework accommodates the inclusion of cluster random effects in e_{gj} and allows for the dependence of regressors within clusters.

Assumption 1. *We assume the following data-generating process:*

(i) *The pairs (Y_{gj}, X_{gj}) and $\left(Y_{g'\ell}, X_{g'\ell}\right)$ are mutually independent for any $g \neq g'$, $j = 1, \cdots, n_g$, and $\ell = 1, \cdots, n_{g'}$.*

(ii) *The data is generated according to the model described through (1)-(4).*

(iii) *The variables X_{gj} are identically distributed across all g and j, possessing a common marginal density $f(x)$. For any $\underline{n}_g \in \{2,3,4\}$, and for any cluster g* with *$n_g \geq \underline{n}_g$, the random vector $\left(X_{gj_1}^{(\mathrm{ind})}, \cdots, X_{gj_{\underline{n}_g}}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)$ is identically distributed across all g and $j_1, \ldots, j_{\underline{n}_g}$, with a common joint density represented by:*

$$f_{\underline{n}_g}\left(x_1^{(\mathrm{ind})}, \cdots, x_{\underline{n}_g}^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right).$$

Remark 1. The conditions in Assumption 1 (iii) for $f(x)$ and $f_2\left(x_1^{(\mathrm{ind})}, x_2^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ are sufficient for their consistent estimation. On the other hand, since we are not interested in estimating $f_3\left(x_1^{(\mathrm{ind})}, x_2^{(\mathrm{ind})}, x_3^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ and $f_4\left(x_1^{(\mathrm{ind})}, x_2^{(\mathrm{ind})}, x_3^{(\mathrm{ind})}, x_4^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$, the associated conditions in Assumption 1 (iii) could be weakened. For a detailed discussion, refer to Appendix C.

Remark 2. In nonparametric regressions, unobserved cluster heterogeneity is equivalent to a mixture structure. Consider a scenario where the true data-generating process is defined as follows:

$$\begin{aligned} Y_{gj} &= m\left(X_{gj}, U_g\right) + e_{gj}, \\ \mathbb{E}\left[e_{gj} \mid \mathbf{X}_g, U_g\right] &= 0, \end{aligned}$$

where U_g is an unobserved cluster-level variable. The critical condition here is that U_g and e_{gj} are separable, and U_g is exogenous. Under these conditions, the estimand, derived through the law of iterated expectations, is expressed as:

$$\begin{aligned} m\left(X_{gj}\right) = \mathbb{E}\left[Y_{gj} \mid \mathbf{X}_g\right] &= \mathbb{E}\left[\mathbb{E}\left[Y_{gj} \mid \mathbf{X}_g, U_g\right] \mid \mathbf{X}_g\right] \\ &= \mathbb{E}\left[m\left(X_{gj}, U_g\right) \mid \mathbf{X}_g\right] = \int m\left(X_{gj}, U_g\right) f_{U_g \mid \mathbf{X}_g}(U_g \mid \mathbf{X}_g) \mathrm{d}U_g. \end{aligned}$$

This formulation implies that $m\left(X_{gj}\right)$ is essentially a mixture of $m\left(X_{gj}, U_g\right)$, integrated over the unknown conditional density $f_{U_g \mid \mathbf{X}_g}(U_g \mid \mathbf{X}_g)$. Additionally, the condition $\mathbb{E}\left[e_{gj} \mid \mathbf{X}_g, U_g\right] = 0$ ensures $\mathbb{E}\left[e_{gj} \mid \mathbf{X}_g\right] = 0$, allowing us to treat $m\left(X_{gj}\right)$ as homogeneous across clusters without loss of generality.

Similarly, consider a scenario where the true density of X_{gj} exhibits cluster heterogeneity, represented by the marginal density $f_{X,V_g}(X_{gj}, V_g)$, with V_g being an unobserved cluster-level variable.

In this context, our estimand becomes a mixture of $f_{X,V_g}(X_{gj}, V_g)$, which can be formally expressed as:

$$f(X_{gj}) = \int f_{X,V_g}\left(X_{gj}, V_g\right) f(V_g)\mathrm{d}V_g.$$

This integral representation implies that the regressors possess identical marginal distributions across clusters. Analogously to the treatment of marginal densities, cluster heterogeneities within joint densities can be conceptualized as mixture structures.

Remark 3. Although the majority of research on cluster sampling treats cluster sizes as deterministic, as does this paper, Bugni *et al.* (2022) treat cluster sizes as a random variable in a cluster-level randomized experiment setup. Their investigation primarily focuses on estimating treatment effects across clusters of varying sizes and developing inference methods that account for the randomness of cluster sizes. This methodological divergence stems from differing concepts of the data-generating process. Bugni *et al.* (2022) address scenarios where researchers sample *clusters* in an experiment, viewing cluster sizes as one of the attributes. Conversely, we consider cases where researchers sample *individuals* with given cluster sizes. Abadie, Athey, Imbens and Wooldridge (2023) propose an alternate sampling framework wherein clusters are sampled from a larger population of cluster, followed by the sampling of individuals from these selected clusters' subpopulations.

3. Nonparametric density estimation

In this section, we show the consistency of nonparametric density estimators. In this paper, we will use kernel functions satisfying the following definitions.

Definition 1. A *univariate kernel function* $k : \mathbb{R} \to \mathbb{R}$ is defined to satisfy the following criteria:

(i) $0 \le k(u) \le \overline{k} < \infty$.
(ii) $k(u) = k(-u)$.
(iii) $\int_{-\infty}^{\infty} k(u)\mathrm{d}u = 1$.
(iv) $\kappa_2 \equiv \int_{-\infty}^{\infty} u^2 k(u)\mathrm{d}u < \infty$ and $\int_{-\infty}^{\infty} u^4 k(u)\mathrm{d}u < \infty$.

Definition 2. A *multivariate kernel function* $K : \mathbb{R}^d \to \mathbb{R}$ is constructed as the product of univariate kernel functions across dimensions,

$$K\left(X\right) = \prod_{q=1}^{d} k\left(X^{(q)}\right),$$

where $k(\cdot)$ is a univariate kernel function and $X^{(q)}$ is the q-th component of X. The upper bound of the multivariate kernel is $K\left(X\right) \le \overline{k}^d \equiv \overline{K}$.[2]

The kernel density estimator for $f(x)$ is:

$$\widehat{f}(x) = \frac{1}{nh^d} \sum_{g=1}^{G} \sum_{j=1}^{n_g} K\left(\frac{X_{gj} - x}{h}\right), \tag{5}$$

where $h > 0$ is a bandwidth.

[2] Without loss of generality, we assume $\overline{k} \ge 1$.

Remark 4. The kernel density estimator given in (5) can be rewritten as $\widehat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$ as in the i.i.d. case. Thus, at least for the estimation, we can use a standard software package. This also applies to nonparametric regression.

For the sake of simplicity, our discussion will focus on scenarios where a single bandwidth is used for all components of X. However, our theory can be generalized to accommodate multivariate bandwidths by substituting h with a bandwidth matrix, as discussed by Ruppert and Wand (1994).

Assumption 2.

(i) $nh^d \to \infty$.

(ii) $h \to 0$ and $\left(\max_{g \le G} n_g\right) h^{d_{\text{ind}}} = O(1)$.

(iii) There exists some neighborhood $\mathcal{N}$ of $x = \left(x^{(\text{ind})\top}, x^{(\text{cls})\top}\right)^\top$ such that $f(x)$ is twice continuously differentiable and $f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ is continuously differentiable.

Remark 5. Assumption 2 (ii) notably extends the i.i.d. case to cluster-dependent settings, introducing a novel condition for bandwidth in the presence of cluster heterogeneity. This condition necessitates a more cautious selection of bandwidth under cluster sampling, balancing the need for $nh^d \to \infty$ against the constraint of $\left(\max_{g \le G} n_g\right) h^{d_{\text{ind}}} = O(1)$. The condition $\left(\max_{g \le G} n_g\right) h^{d_{\text{ind}}} = O(1)$ requires that the maximum cluster size is not growing faster than the shrinking speed of the h neighborhood for the individual-level regressors.
Furthermore, Assumption 2 (iii) underscores the importance of smoothness in both marginal and joint densities within clusters, emphasizing the need for careful examination of density shapes affecting within-cluster observation relationships.
To be precise, Assumption 2 (iii) means that $f(\widetilde{x})$ is twice continuously differentiable at any $\widetilde{x} \in \mathcal{N}$ and $f_2\left(\widetilde{x}_1^{(\text{ind})}, \widetilde{x}_2^{(\text{ind})}; \widetilde{x}^{(\text{cls})}\right)$ is continuously differentiable at any $\left(\widetilde{x}_1^{(\text{ind})\top}, \widetilde{x}^{(\text{cls})\top}\right)^\top$, $\left(\widetilde{x}_2^{(\text{ind})\top}, \widetilde{x}^{(\text{cls})\top}\right)^\top \in \mathcal{N}$. Assumption 2 (iii) limits our analysis to interior points. Although we focus on interior points x, the results could be extended to boundary points.

Remark 6. $nh^d \to \infty$ and $\left(\max_{g \le G} n_g\right) h^{d_{\text{ind}}} = O(1)$ together imply that $\left(\max_{g \le G} n_g\right)/n \to 0$, which is a key assumption of Hansen and Lee (2019) for parametric models under cluster sampling. Moreover, $\left(\max_{g \le G} n_g\right)/n \to 0$ implies $G \to \infty$. Thus, our theory requires $G \to \infty$ implicitly. If we only have the bounded size of clusters $\max_{g \le G} n_g = O(1)$, then, n has the same asymptotic order as G.

Theorem 1. ***(Pointwise consistency)*** *Suppose that Assumptions 1 and 2 hold. Then,* $\widehat{f}(x) \xrightarrow{p} f(x)$.

Remark 7. Beyond pointwise consistency, it is possible to derive expressions for the asymptotic conditional bias and variance, as well as establish the asymptotic normality of $\widehat{f}(x)$. These derivations, while omitted for brevity, follow directly from analogous proofs for the Nadaraya-Watson estimator discussed subsequently.

4. Nadaraya-Watson estimator

In this section, we derive an asymptotic theory for the Nadaraya-Watson estimator (a.k.a. local constant estimator) for estimating the conditional expectation $\mathbb{E}\left[Y_{gj} \mid X_{gj}=x\right]$. The estimator is:

$$\hat{m}_{\mathrm{nw}}(x)=\frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) Y_{gj}}{\sum_{g=1}^{G} \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)}. \tag{6}$$

Assumption 3.

(i) The density function is strictly positive at x, $f(x)>0$.

(ii) There exists some neighborhood $\mathcal{N}$ of $x=\left(x^{(\mathrm{ind})\top}, x^{(\mathrm{cls})\top}\right)^{\top}$ such that $m(x)$ and $f(x)$ are twice continuously differentiable, $f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ is continuously differentiable, and $f_3\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$, $f_4\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}, x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$, $\sigma^2(x)$, and $\sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ are continuous.

Remark 8. Assumption 3 (i) is standard for the Nadaraya-Watson estimator. Assumption 3 (ii) generalizes the assumption for the i.i.d. case. It requires smoothness for joint densities of observations within the same cluster and the conditional covariance as well as the marginal density and the conditional variance.

Theorem 2. ***(Asymptotic bias)*** *Suppose that Assumptions 1-3 hold. Then,*

$$\mathbb{E}\left[\hat{m}_{\mathrm{nw}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]=m(x)+h^2 B_{\mathrm{nw}}(x)+o_p\left(h^2\right)+O_p\left(\sqrt{\frac{1}{n h^{d-2}}}\right),$$

where

$$B_{\mathrm{nw}}(x)=\kappa_2 \sum_{q=1}^{d}\left(\frac{1}{2} \partial_{qq} m(x)+f(x)^{-1} \partial_q f(x) \partial_q m(x)\right),$$

$\partial_q f(x)=\partial f(x) / \partial x^{(q)}$*,* $\partial_q m(x)=\partial m(x) / \partial x^{(q)}$*, and* $\partial_{qq} m(x)=\partial^2 m(x) / \partial\left(x^{(q)}\right)^2$*.*

We use the following assumption to derive the asymptotic variance.

Assumption 4. $\left(\frac{1}{n} \sum_{g=1}^{G} n_g^2\right) h^{d_{\mathrm{ind}}} \rightarrow \lambda \in[0, \infty)$.

Remark 9. $\left(\frac{1}{n} \sum_{g=1}^{G} n_g^2\right) h^{d_{\mathrm{ind}}}=O(1)$ is implied by $\left(\max_{g \leq G} n_g\right) h^{d_{\mathrm{ind}}}=O(1)$ since $\sum_{g=1}^{G} n_g=n$. Assumption 4 guarantees its convergence.

Theorem 3. ***(Asymptotic variance)*** *Suppose that Assumptions 1-4 hold. Then,*

$$\begin{aligned}&\operatorname{Var}\left[\hat{m}_{\mathrm{nw}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\ &\quad=\frac{R_k^d \sigma^2(x)}{f(x) n h^d}+\frac{\lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2 n h^d}+o_p\left(\frac{1}{n h^d}\right),\end{aligned} \tag{7}$$

where $R_k=\int_{-\infty}^{\infty} k(u)^2 \mathrm{d}u$*. In particular, if* $\lambda=0$*,*

$$\operatorname{Var}\left[\hat{m}_{\mathrm{nw}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]=\frac{R_k^d \sigma^2(x)}{f(x) n h^d}+o_p\left(\frac{1}{n h^d}\right).$$

In the special case of $\lambda = 0$, the asymptotic conditional variance is equivalent to the i.i.d. case. A sufficient condition for $\lambda = 0$ is $\left(\max_{g \leq G} n_g\right) h^{d_{\text{ind}}} = o(1)$. In a finite sample, it is more precise to consider $\lambda > 0$. The sign of the second term of (7) depends on the sign of $\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$. In economic applications, it usually takes a positive value, indicating positive conditional covariance of error terms within clusters. Neglecting this term will lead to under-coverage in empirical applications.

Remark 10. The pivotal condition for this theorem is Assumption 4. Note that we can calculate $\left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) h^{d_{\text{ind}}}$ directly. The part $\left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right)$ can be interpreted as follows. Although we are considering *deterministic* cluster sizes n_g, the value $\frac{1}{n}\sum_{g=1}^{G} n_g^2 = \left(\frac{1}{G}\sum_{g=1}^{G} n_g^2\right) / \left(\frac{n}{G}\right)$ can be interpreted as the second moment of the cluster sizes over the first moment of the cluster sizes "$\mathbb{E}\left[n_g^2\right] / \mathbb{E}\left[n_g\right]$", where expectations are taken over $\{n_g\}_{g=1}^{G}$.

Remark 11. In the following two special cases, the second term of (7) has a simpler form. Firstly, if we assume the conditional independence $f_2\left(x^{(\text{ind})}, x^{(\text{ind})} \mid x^{(\text{cls})}\right) = f\left(x^{(\text{ind})} \mid x^{(\text{cls})}\right)^2$, (7) simplifies to $\lambda R_k^{d_{\text{cls}}} \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) / \left(f\left(x^{(\text{cls})}\right) nh^d\right)$. Secondly, if we assume the independence between individual and cluster-level regressors (or assume that there are no cluster-level regressors, $d_{\text{cls}} = 0$), (7) simplifies to

$$\frac{\lambda R_k^{d_{\text{cls}}} \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) f\left(x^{(\text{ind})} \mid x^{(\text{ind})}\right)}{f\left(x\right) nh^d}$$

(or $\lambda \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}\right) f\left(x^{(\text{ind})} \mid x^{(\text{ind})}\right) / \left(f\left(x^{(\text{ind})}\right) nh^d\right)$, respectively).

Theorem 4. ***(Pointwise consistency)*** *Suppose that Assumptions 1-3 hold. Then,*

$$\widehat{m}_{\text{nw}}\left(x\right) \xrightarrow{p} m\left(x\right). \tag{8}$$

Assumption 5.

(i) *There exists some $r \geq 2$ such that*

(a) *for any $\widetilde{x} = \left(\widetilde{x}^{(\text{ind})\top}, \widetilde{x}^{(\text{cls})\top}\right)^\top \in \mathcal{N}$,*

$$\mathbb{E}\left[|e|^{2r} \mid X = \widetilde{x}\right] \leq \overline{v}^2 < \infty, \tag{9}$$

(b) *for some constant $C > 0$,*

$$\frac{\left(\sum_{g=1}^{G} n_g^r\right)^{1/r}}{n^{1/4}} \leq C < \infty, \tag{10}$$

(c) *and*

$$\frac{1}{n^{r/2} h^{dr-d}} = O(1). \tag{11}$$

(ii) *We also assume*

$$nh^{d+4} = O(1), \tag{12}$$

$$R_k^d f(x)\sigma^2(x) + \lambda R_k^{d_{\text{cls}}} f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) > 0,$$

and

$$\max_{g \leq G} \frac{n_g^4}{n} \to 0 \tag{13}$$

as $n \to \infty$.

Theorem 5. ***(Asymptotic Normality)*** *Suppose that Assumptions 1-5 hold. Then,*

$$\sqrt{nh^d}\left(\widehat{m}_{\mathrm{nw}}(x) - m(x) - h^2 B_{\mathrm{nw}}(x)\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k^d \sigma^2(x)}{f(x)} + \frac{\lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2}\right). \quad (14)$$

The asymptotic distribution has the same bias and the same convergence rate as in the i.i.d. case. The asymptotic variance is a scaled value of the primal terms of asymptotic conditional variance that include the conditional covariance term due to the cluster dependence. Our simulation in Section 9 shows the importance of considering this term in inference.

The asymptotic variance in the previous literature with bounded cluster sizes (e.g., Bhattacharya, 2005) has only the first term of (14). Under bounded cluster sizes, cluster dependence is asymptotically negligible since an observation in the g-th cluster has a negligible number of observations belonging to the same cluster around the local neighborhood. On the other hand, under growing cluster sizes $n_g \to \infty$, the observation could have a non-negligible number of neighboring observations belonging to the same cluster. Thus, the conditional covariance of error terms matters in our general setup.

Remark 12. Conditions (9) and (12) are standard in the kernel regressions. Replacing (12) by $nh^{d+4} = o(1)$ eliminates the asymptotic bias (undersmoothing). Conditions (10) and (13) require smaller cluster sizes than conditions in Hansen and Lee (2019). Indeed, they require $\left(\sum_{g=1}^{G} n_g^r\right)^{1/r} / n^{1/2} \leq C < \infty$ and $\max_{g \leq G} n_g^2/n \to 0$, which are implied by (10) and (13). Condition (11) is not strict if regressors have small dimension d. For example, the AIMSE-optimal bandwidth in Section 7 satisfies nh^{d+4} is bounded away from zero. In this case, (11) is always satisfied if $d \leq 4$ and equivalent to $r \leq 2d/(d-4)$ if $d > 4$.

5. Local linear estimator

In this section, we consider the local linear estimator

$$\hat{m}_{\mathrm{LL}}(x) = \sum_{g=1}^{G} \sum_{j=1}^{n_g} K_{\mathrm{LL}}(X_{gj}, x) Y_{gj}, \quad (15)$$

where

$$K_{\mathrm{LL}}(u, x) = \mathbf{e}_1^\top \left(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \begin{bmatrix} 1 \\ u - x \end{bmatrix} K_h(u - x),$$

$$\underbrace{\mathbf{e}_1}_{(d+1)\times 1} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \underbrace{\mathbf{X}_x}_{n\times(d+1)} = \begin{bmatrix} 1 & (X_1 - x)^\top \\ \vdots & \vdots \\ 1 & (X_n - x)^\top \end{bmatrix}, \quad \underbrace{\mathbf{W}_x}_{n\times n} = \begin{bmatrix} K_h(X_1 - x) & & O \\ & \ddots & \\ O & & K_h(X_n - x) \end{bmatrix},$$

and $K_h(\cdot) = \frac{1}{h^d} K\left(\frac{\cdot}{h}\right)$. We will assume an additional condition for the simplicity of proofs.

Assumption 6. *K has a compact support.*

Remark 13. Assumption 6 is a standard technical assumption for local linear estimators. It can be replaced by a tail decay assumption for K (see e.g., Fan and Gijbels, 1992).

We can establish similar asymptotic theories for local linear estimators as we derived for Nadaraya-Watson estimators. As in the i.i.d. case, the asymptotic bias of a local linear estimator does not include the term of first-order derivatives.

Theorem 6. ***(Asymptotic bias)*** *Suppose that Assumptions 1-3 and 6 hold. Then,*

$$\mathbb{E}\left[\hat{m}_{\mathrm{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] = m(x) + h^2 B_{\mathrm{LL}}(x) + o_p\left(h^2\right),$$

where

$$B_{\mathrm{LL}}(x) = \frac{\kappa_2}{2}\sum_{q=1}^{d} \partial_{qq} m(x).$$

Theorem 7. ***(Asymptotic variance)*** *Suppose that Assumptions 1-4 and 6 hold. Then,*

$$\begin{aligned}&\operatorname{Var}\left[\hat{m}_{\mathrm{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]\\ &= \frac{R_k^d \sigma^2(x)}{f(x) n h^d} + \frac{\lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2 n h^d} + o_p\left(\frac{1}{nh^d}\right).\end{aligned}$$

In particular, if $\lambda = 0$,

$$\operatorname{Var}\left[\hat{m}_{\mathrm{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] = \frac{R_k^d \sigma^2(x)}{f(x) n h^d} + o_p\left(\frac{1}{nh^d}\right).$$

Theorem 8. ***(Pointwise consistency)*** *Suppose that Assumptions 1-3 and 6 hold. Then,*

$$\widehat{m}_{\mathrm{LL}}(x) \xrightarrow{p} m(x). \tag{16}$$

Theorem 9. ***(Asymptotic normality)*** *Suppose that Assumptions 1-6 hold. Then,*

$$\begin{aligned}&\sqrt{nh^d}\left(\widehat{m}_{\mathrm{LL}}(x) - m(x) - h^2 B_{\mathrm{LL}}(x)\right)\\ &\xrightarrow{d} \mathrm{N}\left(0, \frac{R_k^d \sigma^2(x)}{f(x)} + \frac{\lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2}\right).\end{aligned} \tag{17}$$

6. Uniform convergence

If we impose further assumptions, our pointwise consistency result can be strengthened to uniform consistency. Before proving uniform consistency for nonparametric estimators, we will show uniform consistency for the generic function

$$\widehat{\psi}(x) = \frac{1}{nh^d}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj} - x}{h}\right) W_{gj} \tag{18}$$

to its expectation, where $X_{gj} \in \mathbb{R}^d$ and $W_{gj} \in \mathbb{R}$.

We assume the cluster samples $\{W_{gj}, X_{gj}\}$ satisfy the following assumptions.

Assumption 7. *There exists a constant $\overline{V}$ such that*

$$\sup_x \operatorname{Var}\left(\widehat{\psi}(x)\right) \leq \frac{\overline{V}}{nh^d}$$

for sufficiently large n.

Assumption 8. *For every $i = 1, \dots, n$ and for some $s > 2$, we have*

$$\mathbb{E}\left[|W_i|^s\right] < B_1 < \infty \tag{19}$$

and

$$\sup_x \mathbb{E}\left[|W_i|^s \mid X_i = x\right] f(x) < B_2 < \infty. \tag{20}$$

We also assume that

$$\frac{\left(\max_{g\le G} n_g\right)^2 \log n}{n^{1-(2/s)} h^d} = O(1). \tag{21}$$

Remark 14. The conditions are standard to establish uniform convergence except for (21). Equation (21) has an additional component $\left(\max_{g\le G} n_g\right)^2$ in cluster sampling. If we focus on bounded size clusters, (21) can be reduced to the standard assumption for the i.i.d. case.
For some applications, W_i has a bounded support. In this case, Assumption 8 is satisfied with $s = \infty$ after rescaling $W_i \in [-1, 1]$.

We also require a further assumption on the kernel function.

Assumption 9. *For some $0 < L < \infty$, K has a compact support, that is, $K(u) = 0$ for $\|u\| > L$. Furthermore, K is Lipschitz, i.e., for some constant $\Lambda < \infty$ and for all $u, u' \in \mathbb{R}$, $|K(u) - K(u')| \le \Lambda \|u - u'\|$.*

Theorem 10. ***(Uniform consistency for the general estimator)*** *Suppose that $\{W_{gj}, X_{gj}\}$ satisfies Assumption 1 and Assumptions 7, 8, and 9 hold.*

Then, for any

$$c_n = O\left(\left(\max_{g\le G} n_g\right)^{2/d} (\log n)^{1/d}\right) \tag{22}$$

and

$$a_n = \left(\frac{\log n}{nh^d}\right)^{1/2}, \tag{23}$$

$\widehat{\psi}(x)$ converges in probability to $\mathbb{E}\left[\widehat{\psi}(x)\right]$ uniformly on $\|x\| \le c_n$, i.e.,

$$\sup_{\|x\|\le c_n} \left|\widehat{\psi}(x) - \mathbb{E}\left[\widehat{\psi}(x)\right]\right| = O_p(a_n), \tag{24}$$

as $nh^d \to \infty$, $h \to 0$, and $\left(\max_{g\le G} n_g\right) h^{d_{\text{ind}}} = O(1)$.

The proof for Theorem 10 relies on the following cluster sampling version of Bernstein's inequality, which could be of independent interest.

Lemma 1. ***(Bernstein's inequality for cluster sampling)***

For random variables under cluster sampling $\left\{\{Y_{gj}\}_{j=1}^{n_g}\right\}_{g=1}^{G}$ with bounded ranges $[-B, B]$ and zero means,

$$\mathbb{P}\left[\left|\widetilde{\mathbf{Y}}_1 + \cdots + \widetilde{\mathbf{Y}}_G\right| > \varepsilon\right] \le 2\exp\left\{-\frac{1}{2}\frac{\varepsilon^2}{v + \left(\max_{g\le G} n_g\right) B\varepsilon/3}\right\}$$

for every $\varepsilon > 0$ *and* $v \geq \operatorname{Var}\left(\widetilde{\mathbf{Y}}_1 + \cdots + \widetilde{\mathbf{Y}}_G\right)$, *where* $\widetilde{\mathbf{Y}}_g = \sum_{j=1}^{n_g} Y_{gj}$.

Based on Theorem 10, we will show the uniform consistency of the nonparametric density estimator and nonparametric regressions. It requires the following conditions, including uniform smoothness.

Assumption 10.

(i) $nh^d \to \infty$.

(ii) $h \to 0$ and $\left(\max_{g \leq G} n_g\right) h^{d_{\text{ind}}} = O(1)$.

(iii) $m(x)$ and $f(x)$ have uniformly continuous second-order derivatives and they are uniformly bounded up to second-order derivatives, $f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ has uniformly continuous first-order derivative and is uniformly bounded up to first-order derivative, and $f_3\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$, $f_4\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$, $\sigma^2(x)$, and $\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ are uniformly continuous and uniformly bounded.

Theorem 11. ***(Uniform consistency for the nonparametric density estimator)*** *Suppose that Assumptions 1, 9, and 10 hold. We also assume that*

$$\frac{\left(\max_{g \leq G} n_g\right)^2 \log n}{nh^d} = O(1).$$

Then, for any sequence c_n *satisfying the condition* (22),

$$\sup_{\|x\| \leq c_n} \left|\widehat{f}(x) - f(x)\right| = O_p\left(a_n + h^2\right). \tag{25}$$

Theorem 12. ***(Uniform consistency for the Nadaraya-Watson estimator)*** *Suppose that the assumptions for Theorem 11 hold. We also also assume that Assumption 8 holds for the cluster observations* $\{Y_{gj}, X_{gj}\}$. *If* c_n *is a sequence satisfying the condition* (22),

$$\delta_n = \inf_{\|x\| \leq c_n} f(x) > 0, \tag{26}$$

and

$$\delta_n^{-1}\left(a_n + h^2\right) = o(1),$$

then,

$$\sup_{\|x\| \leq c_n} \left|\widehat{m}_*(x) - m(x)\right| = O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right) \tag{27}$$

for $\widehat{m}_*(x) = \widehat{m}_{\text{nw}}(x)$ *or* $\widehat{m}_{\text{LL}}(x)$.

The range $\{x : \|x\| \leq c_n\}$ expands slowly to $\mathbb{R}^d$ since our condition (22) can cover a sequence $\{c_n\}$ such that $c_n \to \infty$ slowly as $n \to \infty$. This expansion is useful to establish asymptotic theories for semiparametric estimation with a nonparametric kernel estimator in the first-stage.

Suppose that $c_n = c$ (constant) and δ_n is far away zero. Then, the uniform convergence rate for kernel regressions is $a_n + h^2 = \left(\log n/(nh^d)\right)^{1/2} + h^2$. By choosing $h = (\log n/n)^{1/(d+4)}$, the optimal rate $(\log n/n)^{2/(d+4)}$ is attained. This convergence rate is equivalent to Stone (1982)'s optimal rate in the i.i.d. case.

7. Bandwidth selection

In this section, we provide guidelines for selecting bandwidth in nonparametric regressions. We suggest three types of methods: the asymptotic integrated mean squared error (AIMSE) optimal bandwidth selection, the cluster-robust rule-of-thumb, and the cluster-robust cross-validation.

7.1. AIMSE-optimal bandwidth.

Let $B_*(x) = B_{\text{nw}}(x)$ or $B_{\text{LL}}(x)$. The asymptotic integrated mean squared error of the estimator $\widehat{m}_*(x)$ is

$$\begin{aligned}
&\int_{\mathbb{R}^d} h^4 B_*(x)^2 f(x) w(x) \mathrm{d}x \\
&+\int_{\mathbb{R}^d}\left\{\frac{R_k^d \sigma^2(x)}{f(x) n h^d}+\frac{\lambda R_k^{d_{\text{cls}}} f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)}{f(x)^2 n h^d}\right\} f(x) w(x) \mathrm{d}x \\
=\quad & h^4 \overline{B}+\frac{R_k^d \overline{\sigma}^2}{n h^d}+\frac{\left(\frac{1}{n} \sum_{g=1}^G n_g^2\right)}{n} R_k^{d_{\text{cls}}} \overline{\sigma}_{\text{cls}}+o_p\left(\frac{1}{n h^d}\right), \qquad (28)
\end{aligned}$$

where $w(x)$ is some integrable weight function which ensures that $\overline{B} \equiv \int_{\mathbb{R}^d} B_*(x)^2 f(x) w(x) \mathrm{d}x$, $\overline{\sigma}^2 \equiv \int_{\mathbb{R}^d} \sigma^2(x) w(x) \mathrm{d}x$, and

$$\overline{\sigma}_{\text{cls}} \equiv \int_{\mathbb{R}^d} \frac{f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) \sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)}{f(x)} w(x) \mathrm{d}x$$

are finite. We define

$$\text{AIMSE} \equiv h^4 \overline{B}+\frac{R_k^d \overline{\sigma}^2}{n h^d} \qquad (29)$$

as an objective function for bandwidth selection since the third term in (28) does not depend on h and the fourth term in (28) is asymptotically negligible.

Theorem 13. *The AIMSE-optimal bandwidth that minimizes the AIMSE (29) is*

$$h_0=\left(\frac{d R_k^d \overline{\sigma}^2}{4 \overline{B}}\right)^{1/(d+4)} n^{-1/(d+4)}. \qquad (30)$$

Our asymptotic theorems rely on the assumption $(\max_{g \leq G} n_g) h^{d_{\text{ind}}} = O(1)$. When $(\max_{g \leq G} n_g) n^{-d_{\text{ind}}/(d+4)} \to \infty$, the AIMSE-optimal h_0 does not satisfy this order. In this case, the AIMSE-optimal bandwidth does not make sense since the AIMSE criterion itself relies on the assumption $(\max_{g \leq G} n_g) h^{d_{\text{ind}}} = O(1)$. Thus, when the largest cluster size is large compared to the sample size n, we recommend using the cross-validation criterion (see Section 7.3).

7.2. Rule-of-thumb.

In practice, it is not easy to compute the AIMSE-optimal bandwidth since (30) contains unknown parameters. As suggested by Fan and Gijbels (1996, Section 4.2) for the i.i.d. case, we provide a cluster-robust Rule-of-Thumb (CR-ROT) bandwidth choice for a one-dimensional individual-level regressor $x \in \mathbb{R}$. This bandwidth could be a crude estimator of the AIMSE-optimal bandwidth, but the primary purpose of it is to give a guess of the bandwidth requiring little computational effort. Let

$$\check{m}_{-g}(x)=\check{\alpha}_{0,-g}+\cdots+\check{\alpha}_{4,-g} x^4 \qquad (31)$$

be a fitted 4th-order *global* polynomial regression leaving out the g-th cluster. Given this parametric model and a user-specified integrable weight function $w(x)$, the CR-ROT bandwidth is calculated by

$$h_{\text{CR-ROT}} = \left(\frac{d R_k^d \check{\sigma}^2}{4 \check{B}} \right)^{1/(d+4)} n^{-1/(d+4)}, \tag{32}$$

where

$$\begin{aligned}
\check{B} &= \frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} \left\{ \frac{1}{2} \check{m}''_{-g} \left(X_{gj} \right) \right\}^2 w \left(X_{gj} \right) \\
&= \frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} \left\{ \check{\alpha}_{2,-g} + 3 \check{\alpha}_{3,-g} X_{gj} + 6 \check{\alpha}_{4,-g} X_{gj}^2 \right\}^2 w \left(X_{gj} \right), \\
\check{\sigma}^2 &= \left(\frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} \check{e}_{gj}^2 \right) \int_{\mathbb{R}^d} w(x) \mathrm{d}x,
\end{aligned}$$

and $\check{e}_{gj} = Y_{gj} - \check{m}_{-g} \left(X_{gj} \right)$. In words, $\check{B}$ and $\check{\sigma}^2$ are computed by the parametric model (31) and the homoskedastic standard error assumption for local linear estimators. For Nadaraya-Watson estimators, we also assume that X has a uniform distribution for simplicity. Then, we have $f'(x) = 0$ and can compute $\bar{B}$ as for local linear estimators by $B_{\text{nw}}(x) = B_{\text{LL}}(x)$. A common choice of $w(x)$ is an indicator function of some interval.

Equation (32) is different from the standard Rule-of-Thumb (ROT) bandwidth choice by Fan and Gijbels (1996) since it uses $\check{m}_{-g}(x)$ instead of $\check{m}(x)$, which is estimated by the full sample. We use $\check{m}_{-g}(x)$ to eliminate dependence between the estimator $\check{m}_{-g}(\cdot)$ and (Y_{gj}, X_{gj}). This modification should provide a better estimation of out-of-sample prediction error.

7.3. **Cross-validation.**

A heuristic cross-validation function for clustered sampling is

$$\mathrm{CV}(h) \equiv \frac{1}{n} \sum_{g=1}^{G} \sum_{j=1}^{n_g} \tilde{e}_{gj} \left(h \right)^2 w \left(X_{gj} \right), \tag{33}$$

where $\widetilde{e}_{gj} = Y_{gj} - \widetilde{m}_{-g} \left(X_{gj}, h \right)$, and $\widetilde{m}_{-g} \left(X_{gj}, h \right)$ is the leave-one-cluster-out nonparametric estimator computed with bandwidth h and without cluster g. For example, Hansen (2022a, p.693-695) suggests this form of cross-validation, but he does not provide any theoretical guarantees. For Nadaraya-Watson estimators, the leave-one-cluster-out nonparametric estimator is defined by

$$\widetilde{m}_{\text{nw},-g} \left(x, h \right) = \frac{\sum_{g' \neq g} \sum_{j=1}^{n_{g'}} K \left(\frac{X_{g'j} - x}{h} \right) Y_{g'j}}{\sum_{g' \neq g} \sum_{j=1}^{n_{g'}} K \left(\frac{X_{g'j} - x}{h} \right)}. \tag{34}$$

Similarly, for local linear estimators, the leave-one-cluster-out nonparametric estimator is defined by

$$\widetilde{m}_{\text{LL},-g} \left(x, h \right) = \sum_{g' \neq g} \sum_{j=1}^{n_{g'}} K_{\text{LL},-g} \left(X_{g'j}, x \right) Y_{g'j}, \tag{35}$$

where

$$K_{\mathrm{LL},-g}\left(u,x\right)=\mathbf{e}_1^\top\left(\mathbf{X}_{x,-g}^\top\mathbf{W}_{x,-g}\mathbf{X}_{x,-g}\right)^{-1}\left[\begin{array}{c}1\\ u-x\end{array}\right]K_h\left(u-x\right),$$

$\mathbf{X}_{x,-g}$ and $\mathbf{W}_{x,-g}$ are defined by the same way as $\mathbf{X}_x$ and $\mathbf{W}_x$, but without using the variables in the g-th cluster. We will show that this cross-validation criterion works appropriately.

Theorem 14. *Let $\overline{\sigma}_w^2=\mathbb{E}\left[e_{gj}^2 w\left(X_{gj}\right)\right]=\mathbb{E}\left[\sigma^2\left(X_{gj}\right)w\left(X_{gj}\right)\right]$ and $w(x)$ be some integrable weight function. Under Assumption 1, we can decompose the expectation of the cross-validation function over $\left\{\mathbf{Y}_g,\mathbf{X}_g\right\}_{g=1}^G$ as*

$$\mathbb{E}\left[\mathrm{CV}(h)\right]=\overline{\sigma}_w^2+\mathrm{IMSE}_{G-1}(h) \tag{36}$$

where

$$\mathrm{IMSE}_{G-1}(h)\equiv\sum_{g=1}^G\frac{n_g}{n}\mathbb{E}_{-g}\left[\int_{\mathbb{R}^d}\left\{m\left(x\right)-\widetilde{m}_{-g}\left(x,h\right)\right\}^2 f\left(x\right)w\left(x\right)\mathrm{d}x\right], \tag{37}$$

and the last expectation is taken over the sample except for the g-th cluster $\left(\mathbf{Y}_{-g},\mathbf{X}_{-g}\right)=\left\{\mathbf{Y}_{g'},\mathbf{X}_{g'}\right\}_{g'\neq g}$.

Since $\overline{\sigma}_w^2$ does not depend on h, minimizing $\mathbb{E}\left[\mathrm{CV}(h)\right]$ on h is equivalent to minimizing $\mathrm{IMSE}_{G-1}(h)$, which is a sum of the expected mean squared errors weighted by cluster sizes. Thus, this theorem justifies the use of the leave-one-cluster-out cross-validation. We can choose the bandwidth by minimizing a cluster-robust cross-validation function $\mathrm{CV}(h)$ over some finite grid points $H=\left[h_1,\cdots,h_J\right]$,

$$h_{\text{CR-CV}}=\underset{h\in H}{\operatorname{argmin}}\ \mathrm{CV}(h). \tag{38}$$

Note that the decomposition theorem holds for finite samples and does not rely on assumptions such as $\left(\max_{g\leq G}n_g\right)h^{d_{\mathrm{ind}}}=O(1)$.

8. A new cluster-robust variance estimation

Since the asymptotic variance of (14) contains the joint density $f\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)$, the conditional variance $\sigma^2(x)$, and the conditional covariance $\sigma\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)$, we need to estimate each of them for inference. Alternatively, Calonico, Cattaneo and Farrell (2019) and Hansen (2022a) propose to use a finite sample conditional variance of $\widehat{m}\left(X_{gj}\right)$ with estimated error terms as an estimator of the asymptotic variance. To the best of our knowledge, there is no theoretical guarantee of their methods, and this paper is the first research providing asymptotic theories of inference for nonparametric regressions under general cluster sizes.

For the joint density estimation, we propose to use

$$\begin{aligned}&\widehat{f}_2\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)\\ =\ &\frac{1}{Nb^{2d_{\mathrm{ind}}+d_{\mathrm{cls}}}}\\ &\times\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g}K\left(\frac{\left(X_{gj}^{(\mathrm{ind})\top},X_{g\ell}^{(\mathrm{ind})\top},X_g^{(\mathrm{cls})\top}\right)^\top-\left(x^{(\mathrm{ind})\top},x^{(\mathrm{ind})\top},x^{(\mathrm{cls})\top}\right)^\top}{b}\right),\end{aligned} \tag{39}$$

where b is a bandwidth and $N = \sum_{g:n_g \geq 2} n_g(n_g - 1)/2$.

The expression (39) can be interpreted as a standard nonparametric density estimator. We estimate the density using $(2d_{\text{ind}} + d_{\text{cls}})$-dimensional regressors $\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_g^{(\text{cls})\top}\right)^\top$, thus we have $b^{2d_{\text{ind}}+d_{\text{cls}}}$ in the denominator in (39). For clusters larger than 2 (i.e., $n_g \geq 2$), there are $\sum_{1\leq j<\ell\leq n_g} 1 = n_g(n_g-1)/2$ possible combinations of $X_{gj}^{(\text{ind})}$ and $X_{g\ell}^{(\text{ind})}$. Each cluster has a $n_g(n_g-1)/2$ effective size observations, and we have the $N = \sum_{g:n_g\geq 2} n_g(n_g-1)/2$ effective size sample in total. In these senses, (39) is a standard nonparametric density estimator for $(2d_{\text{ind}} + d_{\text{cls}})$-dimensional regressors and $n_g(n_g-1)/2$ size clusters.

Remark 15. Note that we use the kernel $K\left(\frac{\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_g^{(\text{cls})\top}\right)^\top - \left(x^{(\text{ind})\top}, x^{(\text{ind})\top}, x^{(\text{cls})\top}\right)^\top}{b}\right)$ instead of $K\left(\frac{X_{gj}-x}{b}\right) K\left(\frac{X_{g\ell}-x}{b}\right)$. The latter is the kernel to estimate $f_{2'}\left(x_{gj}, x_{g\ell}\right)|_{(x_{gj},x_{g\ell})=(x,x)}$, which is not continuous around $(x_{gj}, x_{g\ell}) = (x, x)$. Indeed, this joint density is degenerate in coordinates of cluster-level regressors since we can rewrite $f_{2'}\left(x_{gj}, x_{g\ell}\right) = f_{2'}\left(x_{gj}, x_{g\ell}\right) \mathbf{1}\left\{x_{gj}^{(\text{cls})} = x_{g\ell}^{(\text{cls})}\right\}$, where $x_{gj} = \left(x_{gj}^{(\text{ind})\top}, x_{gj}^{(\text{cls})\top}\right)^\top$, $x_{g\ell} = \left(x_{g\ell}^{(\text{ind})\top}, x_{g\ell}^{(\text{cls})\top}\right)^\top$, and $x_{gj}^{(\text{cls})} = x_{g\ell}^{(\text{cls})} = x_g^{(\text{cls})}$ by the definition.

We make the following assumptions to estimate the joint density consistently.

Assumption 11. *Define* $\varsigma^2\left(X_{gj}\right) \equiv \mathbb{E}\left[e_{gj}^4 \mid \mathbf{X}_g\right] = \mathbb{E}\left[e_{gj}^4 \mid X_{gj}\right]$ *and*

$$\begin{aligned}\varsigma\left(X_{gj}^{(\text{ind})}, X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}; X_g^{(\text{cls})}\right) &\equiv \mathbb{E}\left[e_{gj}e_{g\ell}e_{gt}e_{gs} \mid \mathbf{X}_g\right] \\ &= \mathbb{E}\left[e_{gj}e_{g\ell}e_{gt}e_{gs} \mid X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}, X_{gt}^{(\text{ind})}, X_{gs}^{(\text{ind})}; X_g^{(\text{cls})}\right].\end{aligned}$$

(i) $Nb^{2d_{\text{ind}}+d_{\text{cls}}} \to \infty$.

(ii) $b \to 0$ *and* $\left(\max_{g\leq G} n_g^2\right) b^{2d_{\text{ind}}} = O(1)$.

(iii) $f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) > 0$.

(iv) *There exists some neighborhood* $\mathcal{N}$ *of* $x = \left(x^{(\text{ind})\top}, x^{(\text{ind})\top}\right)^\top$ *such that* $\sigma^2(x)$, $\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$, *and* $f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *are twice continuously differentiable,* $f_3\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *and* $f_4\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *are continuously differentiable, and* $\varsigma^2(x)$ *and* $\varsigma\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$ *are continuous. Also, joint densities of up to 8 individual-level regressors and cluster-level regressors within the same cluster follow common distributions, and these joint densities*

$$f_5\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right), \cdots,$$

$$f_8\left(x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)$$

are continuous on the neighborhood $\mathcal{N}$.

(v) *There exits some sequence* $\{c_n\}$ *satisfying the condition* (22) *such that for any* $g = 1, \cdots, G$ *and for any* $j = 1, \cdots, n_g$, *we have* $\|X_{gj}\| \leq c_n$ *with probability approaching one.*

Remark 16. Assumption 11 (i) and (ii) correspond to $nh^d \to \infty$, $h \to 0$, and $(\max_{g \leq G} n_g)\, h^{d_{\text{ind}}} = O(1)$ in the marginal density estimation. Assumption 11 (iii) and (iv) are stronger than Assumption 3 so that we can cover regressors $\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_{g}^{(\text{cls})\top}\right)^\top$ constructed by two observations X_{gj} and $X_{g\ell}$. A sufficient condition for Assumption 11 (v) is $\mathbb{E}\left\|X_{gj}\right\| < \infty$ since Markov's inequality implies

$$\Pr\left(\left\|X_{gj}\right\| \geq c_n\right) \leq \mathbb{E}\left\|X_{gj}\right\| / c_n \to 0$$

if we choose $c_n \to \infty$.

Theorem 15. ***(Consistency of the joint density estimator)*** *Suppose that Assumption 11 holds. Then,*

$$\widehat{f}_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) \xrightarrow{p} f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right).$$

Next, we will consider conditional variance and covariance estimation. We only provide Nadaraya-Watson type estimators, but they can be easily extended to local linear type ones. Since the goal here is to estimate $\sigma^2(x)$, we can estimate it as we did for $m(x)$. The infeasible Nadaraya-Watson estimator is

$$\widehat{\sigma}_{\text{nw}}^{2*}\left(x\right) = \frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) e_{gj}^2}{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)},$$

This estimator is infeasible because e_{gj} is unknown. We can replace it by $\widehat{e}_{gj} = Y_{gj} - \widehat{m}_*\left(X_{gj}\right)$ with $\widehat{m}_*(x) = \widehat{m}_{\text{nw}}(x)$ or $\widehat{m}_{\text{LL}}(x)$. The feasible variance estimator of the conditional variance is

$$\widehat{\sigma}_{\text{nw}}^{2}\left(x\right) = \frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) \widehat{e}_{gj}^2}{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)}. \tag{40}$$

The following theorem shows $\widehat{\sigma}_{\text{nw}}^{2}\left(x\right)$ is a consistent estimator.

Theorem 16. ***(Consistency of the variance estimator)*** *Let $\widehat{e}_{gj} = Y_{gj} - \widehat{m}_*\left(X_{gj}\right)$ and $\widehat{m}_*(x) = \widehat{m}_{\text{nw}}(x)$ or $\widehat{m}_{\text{LL}}(x)$. Suppose that the assumptions for Theorem 12 and Assumption 11 hold. Then,*

$$\widehat{\sigma}_{\text{nw}}^{2}\left(x\right) \xrightarrow{p} \sigma^2\left(x\right). \tag{41}$$

Similar to the joint density estimator, we can construct a Nadaraya-Watson type estimator for the conditional covariance using $(2d_{\text{ind}} + d_{\text{cls}})$-dimensional regressors $\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_{g}^{(\text{cls})\top}\right)^\top$:

$$\begin{aligned}&\widehat{\sigma}_{\text{nw}}^{*}\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)\\ = &\frac{\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g} K\left(\frac{\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_{g}^{(\text{cls})\top}\right)^\top - \left(x^{(\text{ind})\top}, x^{(\text{ind})\top}, x^{(\text{cls})\top}\right)^\top}{b}\right) e_{gj}e_{g\ell}}{\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g} K\left(\frac{\left(X_{gj}^{(\text{ind})\top}, X_{g\ell}^{(\text{ind})\top}, X_{g}^{(\text{cls})\top}\right)^\top - \left(x^{(\text{ind})\top}, x^{(\text{ind})\top}, x^{(\text{cls})\top}\right)^\top}{b}\right)}.\end{aligned}$$

Because e_{gj} is unknown, it is infeasible as $\widehat{\sigma}^{2*}_{\mathrm{nw}}(x)$. The feasible version of $\widehat{\sigma}^{2*}_{\mathrm{nw}}$ is estimated by replacing e_{gj} with $\widehat{e}_{gj} = Y_{gj} - \widehat{m}_*\left(X_{gj}\right)$,

$$\begin{aligned}
&\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \\
=\; &\frac{\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g} K\left(\frac{\left(X_{gj}^{(\mathrm{ind})\top}, X_{g\ell}^{(\mathrm{ind})\top}, X_{g}^{(\mathrm{cls})\top}\right)^{\top} - \left(x^{(\mathrm{ind})\top}, x^{(\mathrm{ind})\top}, x^{(\mathrm{cls})\top}\right)^{\top}}{b}\right)\widehat{e}_{gj}\widehat{e}_{g\ell}}{\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g} K\left(\frac{\left(X_{gj}^{(\mathrm{ind})\top}, X_{g\ell}^{(\mathrm{ind})\top}, X_{g}^{(\mathrm{cls})\top}\right)^{\top} - \left(x^{(\mathrm{ind})\top}, x^{(\mathrm{ind})\top}, x^{(\mathrm{cls})\top}\right)^{\top}}{b}\right)}. \qquad (42)
\end{aligned}$$

Theorem 17. ***(Consistency of the covariance estimator)*** *Let* $\widehat{e}_{gj} = Y_{gj} - \widehat{m}_*\left(X_{gj}\right)$ *and* $\widehat{m}_*(x) = \widehat{m}_{\mathrm{nw}}(x)$ *or* $\widehat{m}_{\mathrm{LL}}(x)$. *Suppose that the assumptions for Theorem 12 and Assumption 11 hold. Then,*

$$\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \xrightarrow{p} \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right). \qquad (43)$$

Corollary 1. *Let* $\widehat{\lambda} = \left(\frac{1}{n}\sum_{g=1}^{G} n_g^2\right) h^{d_{\mathrm{ind}}}$. *Let* $\widehat{m}_*(x) = \widehat{m}_{\mathrm{nw}}(x)$ *and* $B_*(x) = B_{\mathrm{nw}}(x)$ *(or* $\widehat{m}_*(x) = \widehat{m}_{\mathrm{LL}}(x)$ *and* $B_*(x) = B_{\mathrm{LL}}(x)$*). Suppose that the assumptions for Theorem 5 (or Theorem 9, respectively), Theorem 16, and Theorem 17 hold. Then,*

$$\begin{aligned}
&\left(\frac{R_k^d \widehat{\sigma}^2_{\mathrm{nw}}(x)}{\widehat{f}(x)} + \frac{\widehat{\lambda} R_k^{d_{\mathrm{cls}}} \widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{\left(\widehat{f}(x)\right)^2}\right)^{-1/2} \\
&\times\sqrt{nh^d}\left(\widehat{m}_*(x) - m(x) - h^2 B_*(x)\right) \\
&\quad\xrightarrow{d} \mathrm{N}(0,1). \qquad (44)
\end{aligned}$$

Corollary 1 suggests to use

$$\sqrt{\frac{1}{nh^d}}\sqrt{\frac{R_k^d \widehat{\sigma}^2_{\mathrm{nw}}(x)}{\widehat{f}(x)} + \frac{\widehat{\lambda} R_k^{d_{\mathrm{cls}}} \widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{\left(\widehat{f}(x)\right)^2}} \qquad (45)$$

as a standard error. The estimator $\widehat{\lambda} R_k^{d_{\mathrm{cls}}} \widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) / \left(\widehat{f}(x)\right)^2$ could be too difficult to estimate in practice for the following two main reasons. First, it contains $\widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ and $\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$, which put most kernel weights for observations that $X_{gj}^{(\mathrm{ind})}$ and $X_{g\ell}^{(\mathrm{ind})}$ are both in the neighborhood of $x^{(\mathrm{ind})}$. In a finite sample, such observations could be rarely observed, and these estimators could be imprecise. Second, it contains a density ratio $\widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)/\widehat{f}(x)$, which is difficult to estimate, especially nonparametrically.

To overcome these difficulties, we provide a parametric compromise under additional assumptions. We assume that $f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ follows a multivariate normal distribution, $x^{(\mathrm{ind})}$ and $x^{(\mathrm{cls})}$ are independent or there are no cluster-level regressors (see also Remark 11), and the conditional covariance is homoskedastic. Then, we can simplify

$$\begin{aligned}
&\frac{\widehat{\lambda} R_k^{d_{\mathrm{cls}}} \widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{\left(\hat{f}(x)\right)^2} \\
= \;& \widehat{\lambda} R_k^{d_{\mathrm{cls}}} \left(\frac{1}{N} \sum_{g: n_g \geq 2} \sum_{1 \leq j < \ell \leq n_g} \check{e}_{gj} \check{e}_{g\ell}\right) \frac{p\left(x^{(\mathrm{ind})} \mid x^{(\mathrm{ind})}, \widehat{\mu}, \widehat{\Sigma}\right)}{\widehat{f}(x)}, \qquad (46)
\end{aligned}$$

where $\check{e}_{gj} = Y_{gj} - \check{m}_{-g}\left(X_{gj}\right)$, $\check{m}_{-g}(x)$ is estimated by the *global* polynomial regression as (31), and $p\left(x_1 \mid x_2, \mu, \Sigma\right)$ is a conditional density function of x_1 given x_2 with the joint distribution $\left(x_1^\top, x_2^\top\right)^\top \sim \mathrm{N}(\mu, \Sigma)$. We can estimate $\widehat{\mu} = \left(\widehat{\mu}_1^\top, \widehat{\mu}_1^\top\right)^\top$ and $\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_{11} & \widehat{\Sigma}_{12} \\ \widehat{\Sigma}_{12} & \widehat{\Sigma}_{11} \end{pmatrix}^\top$ easily by using sample moments. Note that the expectation $\widehat{\mu}_1$ and the variance matrix $\widehat{\Sigma}_{11}$ are the same for x_1 and x_2 since we initially assumed identical marginal densities in Assumption 1.

In practice, we can estimate $\sigma^2(x)$ and $\sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ by using clustered-level jackknife estimators. Hansen (2022b) shows that for parametric linear regressions, clustered-level jackknife variance estimators are better than conventional variance estimators with respect to the worst-case bias. Clustered-level jackknife variance estimators $\widetilde{\sigma}^2_{\mathrm{nw}}(x)$ and $\widetilde{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ are estimated by replacing e_{gj} with $\widetilde{e}_{gj} = Y_{gj} - \widetilde{m}_{-g}\left(X_{gj}\right)$, where $\widetilde{m}_{-g}(\cdot)$ is a nonparametric estimator estimated leaving out the g-th cluster observations. In the simulation section, we will compare coverage ratios of confidence intervals constructed by the conventional standard error $\widehat{\sigma}^2_{\mathrm{nw}}(x)$ and the cluster-robust standard error $\widetilde{\sigma}^2_{\mathrm{nw}}(x)$.

Remark 17. The theorems use the same bandwidth h for $\widehat{m}_*(x)$, $\widehat{f}(x)$, and $\widehat{\sigma}^2_{\mathrm{nw}}(x)$, and the same bandwidth b for $\widehat{f}_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ and $\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ for notational simplicity. However, we can easily extend these results to the case where different bandwidths are used (denoted by $h_m, h_f, h_{\sigma^2}, b_f, b_\sigma$) as long as h_m, h_f, h_{σ^2} and b_{f_2}, b_σ have the same asymptotic orders as h and b, respectively.

Remark 18. Because the estimator must be centered by the unknown bias $h^2 B_*(x)$ as well as the true value $m(x)$ in (44), the bias term should be considered in inference. There are three main ways to handle it. The first way is to ignore it. This ignorance could be justified by an undersmoothing assumption $nh^{d+4} = o(1)$. It is the simplest way but not ideal since the bias exists in a finite sample. Second, in the context of RDD, Calonico, Cattaneo and Titiunik (2014) suggest estimating $B_*(x)$ nonparametrically and using a new standard error to take the randomness due to the bias estimation into account. Third, Armstrong and Kolesár (2018) characterize finite sample optimal confidence intervals with the worst-case bias correction for i.i.d. observations. Comparing these procedures in the cluster dependence case is important, though it is outside of the scope of this paper.

9. Monte Carlo simulation

In this section, we will check the validity of bandwidth selections and confidence intervals in simulated datasets under cluster sampling. For both simulation studies, we consider the following setup. We fix the number of clusters $G = 100$ and cluster sizes $n_g = 20$ for $g = 1, \ldots G - 1$. To evaluate the effect of the largest cluster size, we try two cluster sizes $n_G \in \{20, 100\}$ for cluster $g = G$. Thus, we try two scenarios with $(\max_{g \leq G} n_g)/n \approx \{0.02, 0.09\}$, also corresponding to homogeneous or heterogeneous size clusters. We generated 2000 datasets for replication. For the data-generating process, the following two models are considered.

Setup 1 (homoskedastic errors):

$$Y_{gj} = \sin(2X_{gj}) + 2\exp\left(-16X_{gj}^2\right) + 0.5e_{gj},$$

where $X_{gj} = \sqrt{\rho_X}\,(X_1)_g + \sqrt{1-\rho_X}\,(X_2)_{gj}$, $e_{gj} = \sqrt{\rho_e}c_g + \sqrt{1-\rho_e}u_{gj}$, and we generate $(X_1)_g \sim \mathcal{N}(0,1)$, $(X_2)_{gj} \sim \mathcal{N}(0,1)$, $c_g \sim \mathcal{N}(0,1)$, $u_{gj} \sim \mathcal{N}(0,1)$ independently. We set $\rho_X, \rho_e \in \{0.2, 0.5\}$. Note that larger ρ_X and ρ_e imply stronger cluster dependence on the regressor and the error term, respectively.

Setup 2 (heteroskedastic errors):

$$Y_{gj} = X_{gj}\sin(2\pi X_{gj}) + \sigma(X_{gj})\,e_{gj},$$
$$\sigma(X_{gj}) = \frac{2 + \cos(2\pi X_{gj})}{5},$$

and X_{gj} and e_{gj} are generated in the same way as Setup 1.

A key feature is that Setup 1 has homoskedastic errors, and Setup 2 has heteroskedastic errors. We adopted the functional form $m(\cdot)$ for Setup 1 from Fan and Gijbels (1992) and Setup 2 from Kai, Li and Zou (2010). The data-generating process for X_{gj} and e_{gj} are standard in the cluster dependence literature (Cameron, Gelbach and Miller, 2008; Bartalotti and Brummet, 2017). We set the weight function $w(x)$ for cross-validation and IAMSE equals to $w(x) = \mathbb{I}\{\xi_{\mathrm{L}} \leq x \leq \xi_{\mathrm{U}}\}$, where we set $\xi_{\mathrm{L}} = -1.5$ and $\xi_{\mathrm{U}} = 1.5$ for Setup 1, and $\xi_{\mathrm{L}} = 0$ and $\xi_{\mathrm{U}} = 1$ for Setup 2, respectively. For nonparametric regression, we use the Epachenikov kernel and local linear estimators. Results when using Nadaraya-Watson estimators are presented in Appendix D because their values are similar to the ones by local linear estimators.

9.1. Bandwidth selection.

We will compare four methods of bandwidth choice: (i) rule-of-thumb (ROT), (ii) cluster-robust rule-of-thumb (CR-ROT), (iii) cross-validation (CV), and cluster-robust cross-validation (CR-CV). $h_{\text{CR-ROT}}$ (Equation 32) and $h_{\text{CR-CV}}$ (Equation 38) are what we suggested. The ROT bandwidth choice h_{ROT} is proposed by Fan and Gijbels (1996) for i.i.d. observations. Instead of leave-one-cluster-out global fit as (31) for $h_{\text{CR-ROT}}$, it uses the global fit using the entire sample. h_{CV} minimizes the cross-validation function. The difference between h_{CV} and $h_{\text{CR-CV}}$ is that h_{CV} minimizes a criterion based on leave-one-out prediction errors, while $h_{\text{CR-CV}}$ minimizes a criterion based on leave-one-*cluster*-out prediction errors.

In simulation, we first compute h_{ROT} and $h_{\text{CR-ROT}}$. Then, h_{CV} and $h_{\text{CR-CV}}$ are found by the grid search for 50 points over $[h_{\text{CR-ROT}}/3, 3h_{\text{CR-ROT}}]$. The performance of the methods of bandwidth selection is evaluated by the average squared error (ASE):

$$\text{ASE}(h) = \frac{1}{n_{\text{grid}}} \sum_{k=1}^{n_{\text{grid}}} \left\{\widehat{m}_{\text{LL}}(u_k, h) - m(u_k)\right\}^2,$$

Table 1. Mean of ASE and mean of selected bandwidth (m_{LL}, Setup 1)

	$\max n_g = 20$				$\max n_g = 100$			
	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$
(ρ_X,ρ_e)=(0.2,0.2)	0.0054	0.0053	0.0041	0.0041	0.0053	0.0053	0.0041	0.0041
	{0.0297}	{0.0302}	{0.0482}	{0.0483}	{0.0292}	{0.0297}	{0.0477}	{0.0479}
(ρ_X,ρ_e)=(0.2,0.5)	0.0062	0.0061	0.0049	0.0049	0.0063	0.0062	0.0050	0.0050
	{0.0297}	{0.0302}	{0.0482}	{0.0484}	{0.0292}	{0.0297}	{0.0479}	{0.0479}
(ρ_X,ρ_e)=(0.5,0.2)	0.0055	0.0054	0.0042	0.0042	0.0056	0.0055	0.0042	0.0042
	{0.0292}	{0.0300}	{0.0484}	{0.0486}	{0.0288}	{0.0295}	{0.0482}	{0.0484}
(ρ_X,ρ_e)=(0.5,0.5)	0.0066	0.0065	0.0052	0.0052	0.0068	0.0067	0.0054	0.0054
	{0.0292}	{0.0300}	{0.0486}	{0.0486}	{0.0288}	{0.0295}	{0.0482}	{0.0483}

Note: Means of selected bandwidths are shown in curly brackets.

where $\widehat{m}_{\text{LL}}(u_k, h)$ is the local linear estimator with the bandwidth h, and $\{u_1, \ldots, u_{n_{\text{grid}}}\}$ are the grid points to evaluate the performance. We set the number of the grid $n_{\text{grid}} = 50$ and $\{u_1, \ldots, u_{n_{\text{grid}}}\}$ are evenly distributed over $[\xi_{\text{L}}, \xi_{\text{U}}]$.

Tables 1 and 2 show means of the ASE for the local linear estimator and means of selected bandwidths (in curly brackets) across each simulation draw for Setup 1 and 2, respectively. Each table contains four methods of bandwidth choice in several scenarios. We consider combinations of homogeneous or heterogeneous size clusters, high or low cluster dependence on regressors, and high or low cluster dependence on error terms. In Setup 1 (Table 1, homoskedastic errors), h_{ROT} and $h_{\text{CR-ROT}}$ have similar values of the ASE and the selected bandwidth, and h_{CV} and $h_{\text{CR-CV}}$ have the similar values of them, but h_{CV} and $h_{\text{CR-CV}}$ work better than h_{ROT} and $h_{\text{CR-ROT}}$ in terms of the ASE. Within the same method of bandwidth choice, heterogeneous size clusters $n_G = 100$ and high cluster dependence on regressors $\rho_X = 0.5$ give a slightly larger ASE. Compared to them, high cluster dependence on error terms $\rho_e = 0.5$ gives a much larger ASE.

In Setup 2 (Table 2, heteroskedastic errors), h_{ROT} and $h_{\text{CR-ROT}}$ work poorly because they assume homoskedasticity. Different from Setup 1, h_{ROT} has a larger ASE than $h_{\text{CR-ROT}}$. As Setup 1, h_{CV} and $h_{\text{CR-CV}}$ work well and have similar values of the ASE and the selected bandwidth. The good performance of h_{CV} can not be explained by our theoretical results. We probably need an asymptotic analysis of h_{CV} under the cluster dependence, which is outside of the scope of this paper.

To investigate how close the selected bandwidths are to the bandwidth that minimizes the ASE, we plot two figures for a scenario with $n_G = 100$ and $\rho_X = \rho_e = 0.5$. Figures 1 and 2 have values of bandwidth h in the x-axis and means of the function ASE(h) in the y-axis, which are calculated from simulation draws for Setup 1 and 2, respectively. These figures also contain means of selected bandwidths by four selection methods and h_{argmin} minimizing ASE(h). We find that h_{CV} and $h_{\text{CR-CV}}$ are very close to h_{argmin} in both setups.

We recommend $h_{\text{CR-CV}}$ because it has a theoretical guarantee (Theorem 14) and because it performs the best in our simulation, although the difference of ASEs between h_{CV} and $h_{\text{CR-CV}}$ is subtle. In terms of the computational cost, $h_{\text{CR-CV}}$ is also better than h_{CV} since leave-one-cluster-out estimators use smaller sample sizes than leave-one-out estimators do. $h_{\text{CR-ROT}}$ is

TABLE 2. Mean of ASE and mean of selected bandwidth (m_{LL}, Setup 2)

	$\max n_g = 20$				$\max n_g = 100$			
	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$
(ρ_X, ρ_e)=(0.2,0.2)	0.0096	0.0080	0.0028	0.0028	0.0090	0.0076	0.0027	0.0028
	{0.0890}	{0.0865}	{0.0461}	{0.0462}	{0.0876}	{0.0853}	{0.0457}	{0.0458}
(ρ_X, ρ_e)=(0.2,0.5)	0.0104	0.0087	0.0033	0.0033	0.0098	0.0083	0.0034	0.0034
	{0.0893}	{0.0868}	{0.0461}	{0.0462}	{0.0878}	{0.0855}	{0.0457}	{0.0459}
(ρ_X, ρ_e)=(0.5,0.2)	0.0098	0.0084	0.0029	0.0029	0.0096	0.0082	0.0029	0.0029
	{0.0896}	{0.0877}	{0.0465}	{0.0467}	{0.0889}	{0.0869}	{0.0463}	{0.0465}
(ρ_X, ρ_e)=(0.5,0.5)	0.0103	0.0091	0.0036	0.0036	0.0104	0.0090	0.0037	0.0037
	{0.0892}	{0.0874}	{0.0464}	{0.0466}	{0.0886}	{0.0866}	{0.0461}	{0.0463}

Note: Means of selected bandwidths are shown in curly brackets.

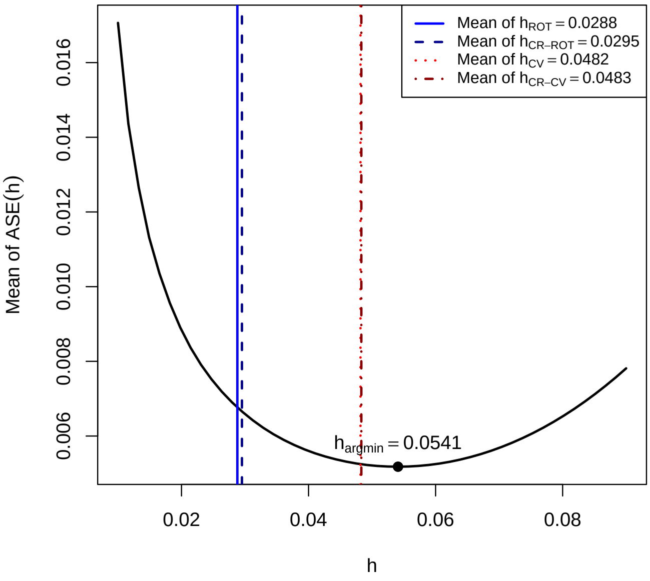

FIGURE 1. Mean of ASE(h) for m_{LL} in Setup 1 with $\max_{g \leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$

useful for a rough estimation and for choosing the range of the grid search in cross-validation. We recommend $h_{\text{CR-ROT}}$ over h_{ROT} for these purposes because it has a smaller ASE.

9.2. **Inference.** We will compare three methods to calculate 95% confidence intervals: (i) using the conventional standard error as for i.i.d. datasets (CI), (ii) using the cluster-robust standard error without the term related to the conditional covariance (CI_{CR}), and (iii) using the cluster-robust standard error with the term related to the conditional covariance (CI_λ). More precisely, we calculate CI with the standard error $\sqrt{R_k^d \widehat{\sigma}^2_{\mathrm{nw}}(x) \,/\, \left(nh^d \widehat{f}(x)\right)}$, CI_{CR} with the standard error

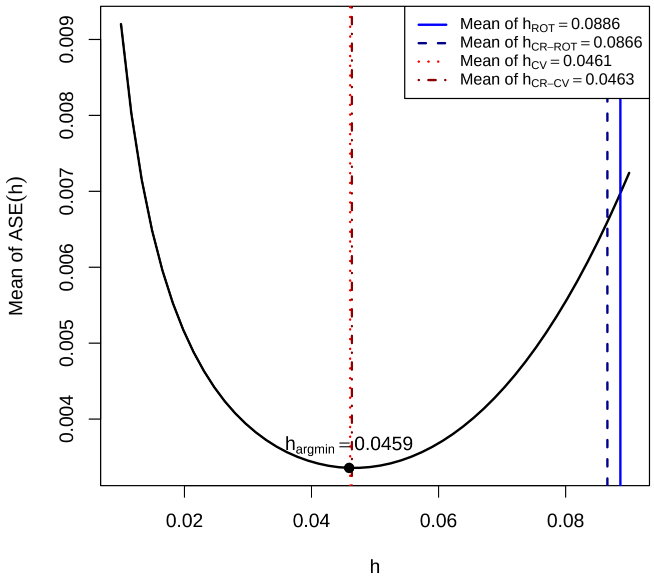

FIGURE 2. Mean of ASE(h) for m_{LL} in Setup 2 with $\max_{g\leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$

$\sqrt{R_k^d \widetilde{\sigma}^2_{\mathrm{nw}}(x) / \left(nh^d \widehat{f}(x)\right)}$, and CI_λ with the standard error

$$\sqrt{\frac{1}{nh^d}}\sqrt{\frac{R_k^d \widetilde{\sigma}^2_{\mathrm{nw}}(x)}{\widehat{f}(x)} + \frac{\widehat{\lambda}\widehat{f}_2(x,x)\,\widehat{\sigma}_{\mathrm{nw}}(x,x)}{\left(\widehat{f}(x)\right)^2}},$$

where $\widehat{\sigma}^2_{\mathrm{nw}}(x)$ and $\widetilde{\sigma}^2_{\mathrm{nw}}(x)$ are nonparametrically estimated with $\widehat{e}_{gj} = Y_{gj} - \widehat{m}_{\mathrm{LL}}(X_{gj})$ and $\widetilde{e}_{gj} = Y_{gj} - \widetilde{m}_{\mathrm{LL},-g}(X_{gj})$, and $\widehat{\lambda f}(x,x)\,\widehat{\sigma}_{\mathrm{nw}}(x,x)$ is calculated parametrically as (46). Note that in our data-generating processes, we have no cluster-level regressor $x^{(\mathrm{cls})}$. In nonparametric regressions, bandwidths are selected as follows. The bandwidth h_m for $\hat{m}_{\mathrm{LL}}(x)$ is calculated by the CR-CV method in the same way as in Section 9.1, the bandwidth h_f for $\widehat{f}(x)$ is calculated by the reference bandwidth of the Epanechnikov kernel $h_f \approx 1.049 \cdot S_X \cdot n^{-1/5}$ where S_X is a standard deviation of X (e.g., see Li and Racine, 2007, Section 1.2). The bandwidth h_{σ^2} for $\widehat{\sigma}^2_{\mathrm{nw}}(x)$ and $\widetilde{\sigma}^2_{\mathrm{nw}}(x)$ is set to h_f. Choosing $h_{\sigma^2} = h_f$ is a conventional choice, for example, used by Imbens and Kalyanaraman (2012). In this simulation, $\widehat{\lambda} = 20 \cdot h_m$ for $n_G = 20$ and $\widehat{\lambda} \approx 23.846 \cdot h_m$ for $n_G = 100$.

To focus on comparisons of inference, we de-bias estimators by the true bias derived analytically. Appendix D contains results without this infeasible bias correction. The CIs are constructed at $x = 0.75$ for Setup 1 and at $x = 0.8$ and 0.4 for Setup 2. Performances of confidence intervals are measured by the coverage ratio across each simulation draw.

Tables 3-5 show the coverage ratio for local linear estimators and means of the length of confidence intervals (in curly brackets) across each simulation draw for Setup 1, Setup 2 with $x = 0.8$, and Setup 2 with $x = 0.4$, respectively. Each table contains results for three types of

Table 3. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 1)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.923	0.926	0.953	0.914	0.916	0.950
	{0.190}	{0.193}	{0.215}	{0.187}	{0.189}	{0.215}
(ρ_X, ρ_e)=(0.2,0.5)	0.875	0.880	0.959	0.859	0.864	0.953
	{0.189}	{0.192}	{0.244}	{0.186}	{0.189}	{0.248}
(ρ_X, ρ_e)=(0.5,0.2)	0.915	0.921	0.956	0.906	0.909	0.953
	{0.189}	{0.192}	{0.225}	{0.186}	{0.189}	{0.226}
(ρ_X, ρ_e)=(0.5,0.5)	0.858	0.868	0.960	0.836	0.848	0.959
	{0.189}	{0.192}	{0.260}	{0.185}	{0.189}	{0.265}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 4. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 2, $x = 0.8$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.893	0.899	0.931	0.884	0.892	0.927
	{0.168}	{0.171}	{0.187}	{0.166}	{0.169}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.842	0.852	0.919	0.827	0.835	0.926
	{0.168}	{0.171}	{0.209}	{0.165}	{0.168}	{0.212}
(ρ_X, ρ_e)=(0.5,0.2)	0.898	0.905	0.934	0.873	0.879	0.925
	{0.167}	{0.171}	{0.192}	{0.165}	{0.168}	{0.193}
(ρ_X, ρ_e)=(0.5,0.5)	0.826	0.835	0.930	0.802	0.809	0.925
	{0.167}	{0.171}	{0.219}	{0.164}	{0.168}	{0.223}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 5. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 2, $x = 0.4$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.991	0.992	0.997	0.988	0.989	0.998
	{0.137}	{0.138}	{0.157}	{0.134}	{0.136}	{0.158}
(ρ_X, ρ_e)=(0.2,0.5)	0.975	0.978	0.999	0.969	0.972	1.000
	{0.136}	{0.139}	{0.182}	{0.134}	{0.136}	{0.184}
(ρ_X, ρ_e)=(0.5,0.2)	0.990	0.991	0.998	0.991	0.992	0.998
	{0.136}	{0.138}	{0.160}	{0.133}	{0.135}	{0.161}
(ρ_X, ρ_e)=(0.5,0.5)	0.973	0.976	0.998	0.962	0.967	0.999
	{0.136}	{0.138}	{0.188}	{0.133}	{0.135}	{0.192}

Note: Lengths of confidence intervals are shown in curly brackets.

confidence intervals in several scenarios. As in Section 9.1, we consider 8 different scenarios with all possible combinations of $n_G \in \{20, 100\}$, $\rho_X \in \{0.2, 0.5\}$ and $\rho_e \in \{0.2, 0.5\}$.

In Setup 1 (Table 3, homoskedastic errors), CI_{CR} has slightly better coverages than CI does although both confidence intervals have severe under-coverage values when $\rho_e = 0.5$. These confidence intervals work more poorly for the case $\max_{g \leq G} n_g = 100$. On the other hand, CI_λ

performs the best among the three methods. It has accurate coverage (95-96%) for every data-generating process.

For Setup 2 (heteroskedastic errors), we consider two different points (Table 4 for $x = 0.8$ and Table 5 for $x = 0.4$). Table 4 shows that CI_λ improves the accuracy greatly, and it attains coverage ratios close to 95%. CI_{CR} and CI fail to reach even 90% coverage ratios for almost all cases. However, Table 5 shows that all three methods have 95% coverage ratios, and CI_λ has over-coverage values at $x = 0.4$. Differences between Table 4 and Table 5 come from the functional form of the error term $\sigma\left(X_{gj}\right) e_{gj}$. Since $\sigma\left(x\right) = \left(2 + \cos\left(2\pi x\right)\right)/5$ takes a large value at $x = 0.8$ and a small value at $x = 0.4$, the conditional variance and covariance of error terms also do so. Overall, CI_λ is the most conservative choice among the three methods. Our proposed confidence interval CI_λ performs well even if we ignore the estimation bias of nonparametric estimators (see Appendix D).

We recommend CI_λ because it works the best for homoskedastic errors, and it provides a conservative interval for heteroskedastic errors in our simulation.

10. Empirical Illustration

In this section, we will apply our methods to a dataset from Alatas *et al.* (2012),[3] which ran an experiment in 640 Indonesian villages with heterogeneous cluster sizes from 17 to 72. The purpose is to investigate a good way to target people with low incomes. In their village-level randomized assignments, they compare three different ways of targeting: using demographic characteristics as proxies of income, using the community knowledge on the ranking of wealth (community targeting), and using a hybrid of them. The wealth ranking for community targeting was measured as follows. In each village, villagers were asked to rank everyone in the community from the richest to the poorest. A facilitator used randomly ordered index cards, each representing a household. Starting the first two cards, the facilitator asked the community which household was better off in terms of wealth. Based on the community's response, the cards were placed with the wealth order. By sequentially adding one more index card to the comparison, the facilitator continued the process until all the households had been ranked.

One concern for this ranking process is that human errors could happen since it took 1.68 hours on average. Alatas *et al.* (2012) investigated this concern by running a nonparametric regression of the mistarget rate (Y_{gj}) on the card order in the ranking process (X_{gj}). The mistarget rate is calculated based on the household's per capita consumption. The card orders in the ranking process are scaled from 0 to 1. Error terms of nonparametric regression can be dependent on the same cluster. For example, some villages can be more patient than others, and their mistarget rate can be less variant across the order in the ranking process. Thus, we revisit Alatas *et al.* (2012) with theoretically justified methods for cluster sampling. We will use the local linear regression with the Epachenikov kernel while Alatas *et al.* (2012) used the local linear regression with the quartic kernel (what they call nonparametric Fan regression).

By the random card order, it is reasonable to assume that the regressor X_{gj} is independent within the cluster (village). Since the distribution of X_{gj} does not follow from $U[0, 1]$ due to the

[3] Their replication package, including datasets, is available on the AEA website.

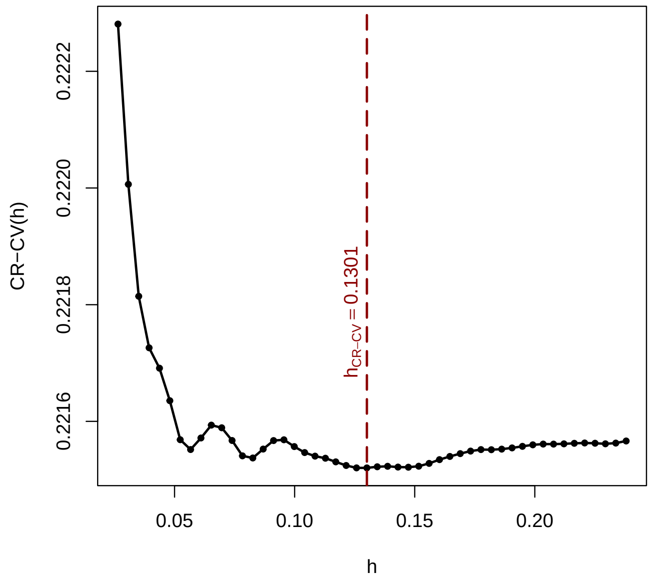

FIGURE 3. Cluster-robust cross-validation function CV(h)

lack of observations on the mistarget rate Y_{gj}, we also estimate it nonparametrically. Thanks to the independence of the regressor, we can estimate the joint density by the product of marginal densities. Other detailed calculations for the bandwidth selection and standard errors are done in the same way as in Section 9.

The sub-dataset for the above regression contains $n = 3784$ observations, $G = 431$ villages, and each village has from 4 to 9 observations. Thus, $\max_{g \le G} n_g$ is 9. The selected bandwidth by CR-CV was 0.1301 while Alatas *et al.* (2012) choose it to be $(\max(X_{gj}) - \min(X_{gj}))/5 = 0.1979$. We plot the cluster-robust cross-validation function in Figure 3.

We calculated three 95% confidence intervals: CI, CI_{CR}, and CI_λ. We calculate $\widehat{\lambda} \approx 1.148$. Since CI and CI_{CR} are almost identical, we only draw CI_{CR} on the plot. Figure 4 shows the estimated nonparametric regression values and estimated pointwise confidence intervals. We found that CI_λ is slightly wider than CI_{CR}. We still have significant pointwise differences between the first few households and the household in the middle of the ranking process (mistargeting rate rises 5-10%) even under wider confidence intervals CI_λ. The conclusions are similar to Alatas *et al.* (2012).

11. CONCLUSION

This article has developed a comprehensive theoretical framework for nonparametric regression analysis under cluster sampling. Our contributions are threefold, addressing critical aspects of cluster-dependent data analysis that have significant implications for econometric methodologies and applied research. First, we allow both growing and bounded size clusters. This extension is crucial, as growing cluster sizes introduce a non-negligible within-cluster dependence, necessitating the inclusion of an additional term in the asymptotic variance to capture this phenomenon

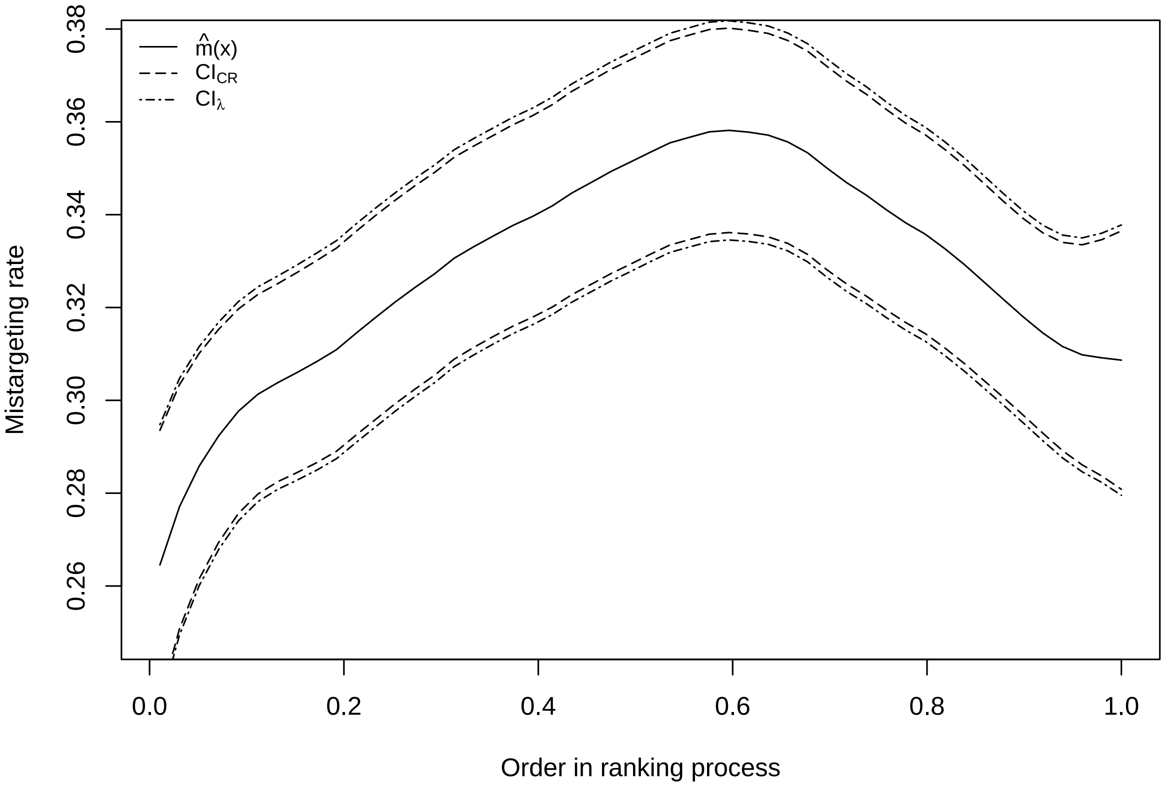

FIGURE 4. Local linear estimation and 95% CIs on Alatas *et al.* (2012)'s dataset

accurately. Second, we cover the case where regressors contain common variables within the same clusters. These cluster-level regressors are the extreme case of cluster-dependent regressors, and they require the careful estimation of the joint density function. Third, our proposed inference is valid with heterogeneous and growing cluster sizes. The simulation studies illustrate the critical role of accounting for within-cluster dependence, affirming the practical relevance of our theoretical insights.

While this article establishes a foundation for nonparametric regression analysis under cluster sampling, several avenues for future research emerge. Theoretical work on other nonparametric estimators, such as local polynomial regressions and series regressions, would be an interesting extension. Investigating boundary analysis is crucial due to its impact on estimator bias. Additionally, developing cluster bootstrap inference methods for nonparametric regressions is important since it would provide more practical statistical inference for clustered data.

Appendix A. Proofs for main results

In this section, we will provide technical lemmas and proofs for the main results. The proofs for technical lemmas are in Appendix B.

Let $K_h(\cdot) = \frac{1}{h^d} K\left(\frac{\cdot}{h}\right)$.

Lemma 2. *Under Assumptions 1 and 2,*

$$F_0(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right) = f\left(x\right) + o_p(1)$$

$$F_1(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right) = o_p\left(h\right)\mathbf{1}_d,$$

$$F_2(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right)\left(X_{gj}-x\right)^\top = h^2 f\left(x\right)\kappa_2\mathbf{I}_{d\times d} + o_p\left(h^2\right)\mathbf{1}_d\mathbf{1}_d^\top.$$

Lemma 3. *Under Assumptions 1-3,*

$$\begin{aligned} J_0(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left\{m\left(X_{gj}\right)-m(x)\right\} \\ &= h^2\kappa_2\sum_{q=1}^{d}\left(\frac{1}{2}\partial_{qq}m(x) + f(x)^{-1}\partial_q f(x)\partial_q m(x)\right) + o_p\left(h^2\right) + O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right), \\ J_1(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left\{m\left(X_{gj}\right)-m(x)\right\}\left(X_{gj}-x\right) \\ &= h^2 f(0)\kappa_2\nabla m\left(0\right) + o_p\left(h^3\right)\mathbf{1}_d + O_p\left(\sqrt{\frac{1}{nh^{d-4}}}\right)\mathbf{1}_d. \end{aligned}$$

Lemma 4. *Under Assumptions 1-3,*

$$H_0(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right)\sigma^2\left(X_{gj}\right) = \frac{1}{h^d} f\left(x\right)\sigma^2\left(x\right)R_k^d + o_p\left(h^{-d}\right),$$

$$H_1(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}-x\right) = o_p\left(h^{-d+1}\right)\mathbf{1}_d,$$

$$\begin{aligned} H_2(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}-x\right)\left(X_{gj}-x\right)^\top \\ &= \frac{1}{h^{d-2}} f\left(x\right)\sigma^2\left(x\right)\left\{\int_{\mathbb{R}^d} K^2\left(T\right)TT^\top \mathrm{d}T\right\} + o_p\left(h^{-d+2}\right)\mathbf{1}_d\mathbf{1}_d^\top. \end{aligned}$$

Lemma 5. *Under Assumptions 1-4,*

$$\begin{aligned}
I_0(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} K_h\left(X_{gj}-x\right)K_h\left(X_{g\ell}-x\right)\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\\
&=\frac{1}{2h^d}\lambda R_k^{d_{\mathrm{cls}}}f_2\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)\sigma\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)+o_p\left(h^{-d}\right),\\
I_1(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} K_h\left(X_{gj}-x\right)K_h\left(X_{g\ell}-x\right)\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{gj}-x\right)\\
&=o_p\left(h^{-d+1}\right)\mathbf{1}_d,\\
I_2(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} K_h\left(X_{gj}-x\right)K_h\left(X_{g\ell}-x\right)\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{g\ell}-x\right)\left(X_{gj}-x\right)^\top\\
&=O_p\left(h^{-d+2}\right)\mathbf{1}_d\mathbf{1}_d^\top.
\end{aligned}$$

Lemma 6. *Under Assumptions 1-4,*

$$\begin{aligned}
\mathcal{E}_0(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)e_{gj}=O_p\left(\sqrt{\frac{1}{nh^d}}\right),\\
\mathcal{E}_1(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)e_{gj}\left(X_{gj}-x\right)=O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)\mathbf{1}_d.
\end{aligned}$$

Lemma 7. *Under Assumptions 1, 2 and 6,*

$$\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)\left\{\left(X_{gj}-x\right)^\top\nabla^2 m(x)\left(X_{gj}-x\right)\right\}=h^2\kappa_2 f(x)\sum_{q=1}^{d}\partial_{qq}m(x)+o_p\left(h^2\right),$$

$$\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right)\left\{\left(X_{gj}-x\right)^\top\nabla^2 m(x)\left(X_{gj}-x\right)\right\}=O_p\left(h^3\right)\mathbf{1}_d.$$

A.1. **Proof for Theorem 1.**

Proof. Lemma 2 for $F_0(x)$ implies the result. □

A.2. **Proof for Theorem 2.**

Proof. Since observations belonging to different clusters are mutually independent and $\mathbb{E}\left[Y_{gj}\mid \mathbf{X}_g\right]=m\left(X_{gj}\right)$,

$$\begin{aligned}
\mathbb{E}\left[\hat{m}_{\mathrm{nw}}(x)\mid \mathbf{X}_1,\cdots,\mathbf{X}_G\right]&=\frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g}K\left(\frac{X_{gj}-x}{h}\right)m\left(X_{gj}\right)}{\sum_{g=1}^{G}\sum_{j=1}^{n_g}K\left(\frac{X_{gj}-x}{h}\right)}\\
&=m(x)+\frac{J_0(x)}{\hat{f}(x)}.
\end{aligned}$$

Theorem 1 implies $\widehat{f}(x)\xrightarrow{p}f(x)>0$. Thus, the continuous mapping theorem and Lemma 3 imply the result. □

A.3. Proof for Theorem 3.

Proof. Since $e_{gj} = Y_{gj} - m\left(X_{gj}\right)$,

$$
\begin{aligned}
& \operatorname{Var}\left[\hat{m}_{\mathrm{nw}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \; & \mathbb{E}\left[\left(\hat{m}_{\mathrm{nw}}(x) - \mathbb{E}\left[\left(\hat{m}_{\mathrm{nw}}(x)\right) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]\right)^2 \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \; & \mathbb{E}\left[\left(\frac{\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)\left(Y_{gj} - m\left(X_{gj}\right)\right)}{\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)}\right)^2 \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \; & \frac{\mathbb{E}\left[\left(\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) e_{gj}\right)^2 \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]}{\left(\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)\right)^2} \\
= \; & \frac{\sum_{g=1}^G \mathbb{E}\left[\left(\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) e_{gj}\right)^2 \mid \mathbf{X}_g\right]}{\left(\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)\right)^2} \\
= \; & \frac{\sum_{g=1}^G \left\{\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)^2 \mathbb{E}\left[e_{gj}^2 \mid \mathbf{X}_g\right] + 2\sum_{1 \leq j < \ell \leq n_g} K\left(\frac{X_{gj}-x}{h}\right) K\left(\frac{X_{g\ell}-x}{h}\right) \mathbb{E}\left[e_{gj} e_{g\ell} \mid \mathbf{X}_g\right]\right\}}{\left(\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)\right)^2} \\
= \; & \frac{\sum_{g=1}^G \left\{\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)^2 \sigma^2\left(X_{gj}\right) + 2\sum_{1 \leq j < \ell \leq n_g} K\left(\frac{X_{gj}-x}{h}\right) K\left(\frac{X_{g\ell}-x}{h}\right) \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)\right\}}{\left(\sum_{g=1}^G \sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)\right)^2} \\
= \; & \frac{h^d \left\{H_0(x) + 2I_0(x)\right\}}{nh^d \left(\hat{f}(x)\right)^2},
\end{aligned}
$$

where the fourth equality follows from the mutual independence between clusters.

Theorem 1 implies $\widehat{f}(x) \xrightarrow{p} f(x) > 0$. Lemmas 4 and 5 for $H_0(x)$ and $I_0(x)$ and the continuous mapping theorem together imply that

$$
\begin{aligned}
& \operatorname{Var}\left[\hat{m}_{\mathrm{nw}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \; & \frac{1}{nh^d} \frac{f(x)\sigma^2(x) R_k^d + \lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) + o_p(1)}{f(x)^2 + o_p(1)} \\
= \; & \frac{R_k^d \sigma^2(x)}{f(x) nh^d} + \frac{\lambda R_k^{d_{\mathrm{cls}}} f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right) \sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)}{f(x)^2 nh^d} + o_p\left(\frac{1}{nh^d}\right).
\end{aligned}
$$

□

A.4. Proof for Theorem 4.

Proof.

$$\begin{aligned}\hat{m}_{\text{nw}}(x) &= \frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) Y_{gj}}{\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right)} \\ &= \frac{\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left\{m(x)+m\left(X_{gj}\right)-m(x)+e_{gj}\right\}}{\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)} \\ &= m(x)+\frac{J_0(x)}{\widehat{f}(x)}+\frac{\mathcal{E}_0(x)}{\widehat{f}(x)} \qquad (47) \\ &\xrightarrow{p} 0\end{aligned}$$

by Theorem 1 and Lemmas 3 and 6. □

A.5. **Proof for Theorem 5.**

Proof. Since we have (47), Theorem 1, Lemma 3, and (12),

$$\begin{aligned}\sqrt{nh^d}\left(\widehat{m}_{\text{nw}}(x)-m(x)-h^2B_{\text{nw}}(x)\right) &= \sqrt{nh^d}\left(\frac{\mathcal{E}_0(x)}{\widehat{f}(x)}\right)+\sqrt{nh^d}\left(\frac{J_0(x)}{\widehat{f}(x)}-h^2B_{\text{nw}}(x)\right) \\ &= \frac{\sqrt{nh^d}\mathcal{E}_0(x)}{\widehat{f}(x)}+\sqrt{nh^d}\left(o_p\left(h^2\right)+O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)\right) \\ &= \frac{\sqrt{nh^d}\mathcal{E}_0(x)}{\widehat{f}(x)}+\left(o_p\left(\sqrt{nh^{d+4}}\right)+O_p(h)\right) \\ &= \frac{\sqrt{nh^d}\mathcal{E}_0(x)}{f(x)+o_p(1)}+o_p(1).\end{aligned}$$

Define $\widetilde{\mathbf{Z}}_{ng}=\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) e_{gj}$. Note that $\left\{\widetilde{\mathbf{Z}}_{ng}\right\}_{g=1}^{G}$ are independent and $\mathbb{E}\left[\widetilde{\mathbf{Z}}_{ng}\right]=0$. We can express $\sqrt{nh^d}\mathcal{E}_0(x)=\frac{1}{\sqrt{nh^d}}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(\frac{X_{gj}-x}{h}\right) e_{gj}=\frac{1}{\sqrt{nh^d}}\sum_{g=1}^{G}\widetilde{\mathbf{Z}}_{ng}$. Denote $s_n^2=\operatorname{Var}\left[\frac{1}{\sqrt{nh^d}}\sum_{g=1}^{G}\widetilde{\mathbf{Z}}_{ng}\right]$. By the proof for Lemma 6,

$$\begin{aligned}s_n^2 &= nh^d\operatorname{Var}\left[\mathcal{E}_0(x)\right]=h^d\mathbb{E}\left[H_0(x)+2I_0(x)\right] \\ &= f(x)\sigma^2(x)R_k^d+\lambda R_k^{d_{\text{cls}}} f_2\left(x^{(\text{ind})},x^{(\text{ind})};x^{(\text{cls})}\right)\sigma\left(x^{(\text{ind})},x^{(\text{ind})};x^{(\text{cls})}\right)+o(1).\end{aligned}$$

By assumption, this implies that s_n^2 deterministically converges to some positive constant. The conclusion follows by applying the Lindeberg Central Limit Theorem and the Slutsky's Lemma. Thus, it is sufficient to verify the Lindeberg condition:

$$\frac{1}{nh^d s_n^2}\sum_{g=1}^{G}\mathbb{E}\left[\widetilde{\mathbf{Z}}_{ng}^2\mathbf{1}\left\{\widetilde{\mathbf{Z}}_{ng}^2\geq nh^d s_n^2\varepsilon\right\}\right]=o(1) \qquad (48)$$

for all $\varepsilon>0$.

Pick any $\varepsilon > 0$ and any $\delta > 0$.

$$\begin{aligned}
& \sup_i \mathbb{E}\left[\left|\frac{1}{n^{1/4}h^{d/2}} K\left(\frac{X_i - x}{h}\right) e_i\right|^r \mathbf{1}\left\{\frac{1}{n^{1/4}h^{d/2}}\left|K\left(\frac{X_i - x}{h}\right) e_i\right| \geq B\right\}\right] \\
\leq \; & \sup_i \frac{1}{n^{r/4}h^{dr/2}} \mathbb{E}\left[K^r\left(\frac{X_i - x}{h}\right)|e_i|^r \mathbf{1}\left\{|e_i| \geq \frac{n^{1/4}h^{d/2}B}{\bar{K}}\right\}\right] \\
= \; & \sup_i \frac{1}{n^{r/4}h^{dr/2}} \mathbb{E}\left[K^r\left(\frac{X_i - x}{h}\right) \mathbb{E}\left[|e_i|^r \mathbf{1}\left\{|e_i| \geq \frac{n^{1/4}h^{d/2}B}{\bar{K}}\right\} \mid X_i\right]\right] \\
= \; & \sup_i \frac{1}{n^{r/4}h^{dr/2-d}} \int K^r\left(T_i\right) \mathbb{E}\left[|e_i|^r \mathbf{1}\left\{|e_i| \geq \frac{n^{1/4}h^{d/2}B}{\bar{K}}\right\} \mid X_i = x + hT_i\right] f\left(x + hT_i\right) \mathrm{d}T_i \\
\leq \; & \int K^r\left(T\right) f\left(x + hT\right) \mathrm{d}T \frac{1}{n^{r/4}h^{dr/2-d}} \overline{v}^2 \left|\frac{\bar{K}}{n^{1/4}h^{d/2}B}\right|^r \\
\leq \; & \bar{K}^{2r-2} R_K^d \left(f(x) + o(1)\right) \frac{1}{n^{r/2}h^{dr-d}} \overline{v}^2 \frac{1}{|B|^r} \\
\leq \; & O(1) \cdot \frac{1}{|B|^r},
\end{aligned}$$

where the first and third inequality follow from the definition of the kernel function (Definition 1) $K(u) \leq \bar{K} < \infty$, the first equality follows from the law of iterated expectations, the second equality follows from the change of variables $\left(X_i - x\right)/h = T_i$, the second inequality follows from

$$\begin{aligned}
\mathbb{E}\left[|e_i|^r \mathbf{1}\left\{|e_i| \geq \frac{n^{1/4}h^{d/2}B}{\bar{K}}\right\} \mid X_i = x + hT_i\right] &= \mathbb{E}\left[\frac{|e_i|^{2r}}{|e_i|^r} \mathbf{1}\left\{|e_i| \geq \frac{n^{1/4}h^{d/2}B}{\bar{K}}\right\} \mid X_i = x + hT_i\right] \\
&\leq \mathbb{E}\left[|e_i|^{2r} \mid X_i = x + hT_i\right] \left|\frac{\bar{K}}{n^{1/4}h^{d/2}B}\right|^r \\
&\leq \overline{v}^2 \left|\frac{\bar{K}}{n^{1/4}h^{d/2}B}\right|^r, \qquad \because (9)
\end{aligned}$$

and the fourth inequality follows by (11). Thus,

$$\lim_{B \to \infty} \sup_i \mathbb{E}\left[\left|\frac{1}{n^{1/4}h^{d/2}} K\left(\frac{X_i - x}{h}\right) e_i\right|^r \mathbf{1}\left\{\frac{1}{n^{1/4}h^{d/2}}\left|K\left(\frac{X_i - x}{h}\right) e_i\right| \geq B\right\}\right] = 0$$

holds. By Lemma 1 of Hansen and Lee (2019), this equation implies

$$\lim_{B \to \infty} \sup_g \mathbb{E}\left[\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|^r \mathbf{1}\left\{\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right| \geq B\right\}\right] = 0.$$

Hence, we can pick B large enough so that

$$\mathbb{E}\left[\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|^r \mathbf{1}\left\{\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right| \geq B\right\}\right] \leq \frac{s_n^r \varepsilon^{r/2-1}}{C^r} \delta \tag{49}$$

for large enough n. Now, let's verify the Lindeberg condition:

$$\begin{aligned}
& \frac{1}{nh^d s_n^2}\sum_{g=1}^{G}\mathbb{E}\left[\widetilde{\mathbf{Z}}_{ng}^2\mathbf{1}\left\{\widetilde{\mathbf{Z}}_{ng}^2\geq nh^d s_n^2\varepsilon\right\}\right] \\
= & \frac{1}{nh^d s_n^2}\sum_{g=1}^{G}\mathbb{E}\left[\widetilde{\mathbf{Z}}_{ng}^2\mathbf{1}\left\{\left|\widetilde{\mathbf{Z}}_{ng}\right|\geq \left(nh^d s_n^2\varepsilon\right)^{1/2}\right\}\right] \\
= & \frac{1}{nh^d s_n^2}\sum_{g=1}^{G}\mathbb{E}\left[\frac{\left|\widetilde{\mathbf{Z}}_{ng}\right|^r}{\left|\widetilde{\mathbf{Z}}_{ng}\right|^{r-2}}\mathbf{1}\left\{\left|\widetilde{\mathbf{Z}}_{ng}\right|\geq \left(nh^d s_n^2\varepsilon\right)^{1/2}\right\}\right] \\
\leq & \frac{1}{nh^d s_n^2\left(\left(nh^d s_n^2\varepsilon\right)^{1/2}\right)^{r-2}}\sum_{g=1}^{G}\mathbb{E}\left[\left|\widetilde{\mathbf{Z}}_{ng}\right|^r\mathbf{1}\left\{\left|\widetilde{\mathbf{Z}}_{ng}\right|\geq \left(nh^d s_n^2\varepsilon\right)^{1/2}\right\}\right] \\
= & \frac{1}{n^{r/4}s_n^r\varepsilon^{r/2-1}}\sum_{g=1}^{G}n_g^r\mathbb{E}\left[\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|^r\mathbf{1}\left\{\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|\geq\frac{s_n n^{1/4}\varepsilon^{1/2}}{n_g}\right\}\right] \\
\leq & \frac{1}{n^{r/4}s_n^r\varepsilon^{r/2-1}}\sum_{g=1}^{G}n_g^r\mathbb{E}\left[\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|^r\mathbf{1}\left\{\left|\frac{\widetilde{\mathbf{Z}}_{ng}}{n^{1/4}h^{d/2}n_g}\right|\geq B\right\}\right] \\
\leq & \frac{\sum_{g=1}^{G}n_g^r}{n^{r/4}C^r}\delta \\
\leq & \delta,
\end{aligned}$$

where the second inequality holds for sufficiently large n since (13) enables us to pick large enough n^* to satisfy

$$\frac{1}{B}\geq\max_{g\leq G}\frac{n_g}{s_n n^{1/4}\varepsilon^{1/2}}\text{ for any }n\geq n^*,$$

the third inequality follows by (49), and the fourth inequality follows by (10). □

A.6. Proof for Theorem 6.

Proof. Define $\mathbf{M}=\left[m\left(X_1\right),\ldots,m\left(X_n\right)\right]^\top$ and

$$\mathbf{D}_h=\begin{pmatrix}1 & 0\\ 0 & h^{-2}\mathbf{I}_{d\times d}\end{pmatrix}.$$

Then, we can rewrite

$$\begin{aligned}
\mathbb{E}\left[\hat{m}_{\mathrm{LL}}(x)\mid\mathbf{X}_1,\cdots,\mathbf{X}_G\right] &= \mathbf{e}_1^\top\left(\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^\top\mathbf{W}_x\mathbf{M} \\
&= \mathbf{e}_1^\top\left(\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{M}
\end{aligned}$$

by $\mathbb{E}\left[e_{gj}\mid\mathbf{X}_g\right]=0$. Let $\mathbf{Q}_m(x)$ be a $n\times 1$ vector

$$\mathbf{Q}_m(x)=\left[\left(X_1-x\right)^\top\nabla^2 m(x)\left(X_1-x\right),\ldots,\left(X_n-x\right)^\top\nabla^2 m(x)\left(X_n-x\right)\right]^\top.$$

By Taylor expansion of $\mathbf{M}$ around x,

$$\mathbf{M}=\mathbf{X}_x\left[m(x),\nabla m(x)^\top\right]^\top+\frac{1}{2}\mathbf{Q}_m(x)+\mathbf{R}_m(x),$$

where $\mathbf{R}_m(x)$ is a $n \times 1$ vector of remainder terms. Compact support of K implies that there exists some constant $C > 0$ such that we essentially use observations with $\left|X_i^{(q)} - x^{(q)}\right| \leq C \cdot h$ for any $i = 1, \ldots, n$ and any $q = 1, \ldots, d$. Thus, by the multivariate Taylor expansion, we can evaluate a scalar random variable as

$$\mathbf{e}_1^\top \left(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^\top \mathbf{W}_x \mathbf{R}_m(x) = o_p\left(h^2\right). \tag{50}$$

By Lemma 2, we can calculate

$$\begin{aligned}
&\left(\frac{1}{n}\mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \\
= &\left\{\mathbf{D}_h \begin{bmatrix} \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right) & \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right)^\top \\ \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right) & \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right)\left(X_{gj}-x\right)^\top \end{bmatrix}\right\}^{-1} \\
= &\begin{bmatrix} f(x) + o_p(1) & o_p\left(h\right)\mathbf{1}_d^\top \\ o_p\left(h^{-1}\right)\mathbf{1}_d & f(x)\kappa_2 I_{d\times d} + o_p\left(1\right)\mathbf{1}_d\mathbf{1}_d^\top \end{bmatrix}^{-1} \\
= &\begin{bmatrix} f(x)^{-1} + o_p(1) & o_p\left(h\right)\mathbf{1}_d^\top \\ o_p\left(h^{-1}\right)\mathbf{1}_d & \left(f(x)\kappa_2 I_{d\times d}\right)^{-1} + o_p\left(1\right)\mathbf{1}_d\mathbf{1}_d^\top \end{bmatrix}
\end{aligned} \tag{51}$$

Also, by Lemma 7,

$$\begin{aligned}
&\frac{1}{n}\mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \mathbf{Q}_m(x) \\
= \ &\mathbf{D}_h \begin{bmatrix} \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left\{\left(X_{gj}-x\right)^\top \nabla^2 m(x)\left(X_{gj}-x\right)\right\} \\ \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(X_{gj}-x\right)\left\{\left(X_{gj}-x\right)^\top \nabla^2 m(x)\left(X_{gj}-x\right)\right\} \end{bmatrix} \\
= \ &\mathbf{D}_h \begin{bmatrix} h^2\kappa_2 f(x)\sum_{q=1}^{d}\partial_{qq}m(x) + o_p\left(h^2\right) \\ O_p\left(h^3\right)\mathbf{1}_d \end{bmatrix} \\
= \ &\begin{bmatrix} h^2\kappa_2 f(x)\sum_{q=1}^{d}\partial_{qq}m(x) + o_p\left(h^2\right) \\ O_p\left(h\right)\mathbf{1}_d \end{bmatrix}.
\end{aligned} \tag{52}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\hat{m}_{\mathrm{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] - m(x) \\
= \ &\mathbf{e}_1^\top\left(\mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \left(\frac{1}{2}\mathbf{Q}_m(x) + \mathbf{R}_m(x)\right) \\
= \ &\frac{1}{2}\mathbf{e}_1^\top\left(\frac{1}{n}\mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1}\left(\frac{1}{n}\mathbf{D}_h \mathbf{X}_x^\top \mathbf{W}_x \mathbf{Q}_m(x)\right) + o_p\left(h^2\right). \\
= \ &h^2\frac{\kappa_2}{2}\sum_{q=1}^{d}\partial_{qq}m(x) + o_p\left(h^2\right),
\end{aligned}$$

where the first equality holds since $\mathbf{e}_1^\top\left(\mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^\top \mathbf{W}_x \mathbf{X}_x\left[m(x), \nabla m(x)^\top\right]^\top = m(x)$, the second equality follows from (50), and the third equality follows from (51) and (52). □

A.7. **Proof for Theorem 7.**

Proof. Let $\mathbf{Y} = \left[Y_1, \ldots, Y_n\right]^\top$. Then,

$$\begin{aligned}
&\operatorname{Var}\left[\hat{m}_{\mathrm{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \ &\mathbf{e}_1^{\top}\left(\mathbf{X}_x^{\top} \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{X}_x^{\top} \mathbf{W}_x \operatorname{Var}\left[\mathbf{Y} \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \mathbf{W}_x \mathbf{X}_x\left(\mathbf{X}_x^{\top} \mathbf{W}_x \mathbf{X}_x\right)^{-1} \mathbf{e}_1.
\end{aligned}$$

Here, $\operatorname{Var}\left[\mathbf{Y} \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]$ is a $n \times n$ matrix having the following structure.

$$\operatorname{Var}\left[\mathbf{Y} \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right]=\left[\begin{array}{ccc}
\mathbf{V}_1 & & O \\
& \ddots & \\
O & & \mathbf{V}_G
\end{array}\right],$$

where $\mathbf{V}_g$ (for $g=1, \ldots, G$) is a matrix with

$$\mathbf{V}_g=\left[\mathbb{E}\left[e_{gj} e_{g\ell} \mid \mathbf{X}_g\right]\right]_{n_g \times n_g} .$$

The upper-left scalar element of

$$n^{-1} \mathbf{X}_x^{\top} \mathbf{W}_x \operatorname{Var}\left[\mathbf{Y} \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \mathbf{W}_x \mathbf{X}_x \equiv\left[\begin{array}{cc}
\Omega_{11} & \boldsymbol{\Omega}_{12} \\
\boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}
\end{array}\right]$$

is

$$\begin{aligned}
\Omega_{11} = \ & \frac{1}{n} \sum_{g=1}^G\left\{\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right) \sigma^2\left(X_{gj}\right)\right. \\
& \left.+2 \sum_{1 \leq j<\ell \leq n_g} K_h\left(X_{gj}-x\right) K_h\left(X_{g\ell}-x\right) \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})} ; X_g^{(\mathrm{cls})}\right)\right\},
\end{aligned}$$

the lower-left $d \times 1$ block is

$$\begin{aligned}
\boldsymbol{\Omega}_{21} = \ & \frac{1}{n} \sum_{g=1}^G\left\{\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right) \sigma^2\left(X_{gj}\right)\left(X_{gj}-x\right)\right. \\
& \left.+2 \sum_{1 \leq j<\ell \leq n_g} K_h\left(X_{gj}-x\right) K_h\left(X_{g\ell}-x\right) \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})} ; X_g^{(\mathrm{cls})}\right)\left(X_{gj}-x\right)\right\},
\end{aligned}$$

and the lower-right $d \times d$ block is

$$\begin{aligned}
\boldsymbol{\Omega}_{22} = \ & \frac{1}{n} \sum_{g=1}^G\left\{\sum_{j=1}^{n_g} K_h^2\left(X_{gj}-x\right) \sigma^2\left(X_{gj}\right)\left(X_{gj}-x\right)\left(X_{gj}-x\right)^{\top}\right. \\
& \left.+2 \sum_{1 \leq j<\ell \leq n_g} K_h\left(X_{gj}-x\right) K\left(X_{g\ell}-x\right) \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})} ; X_g^{(\mathrm{cls})}\right)\left(X_{g\ell}-x\right)\left(X_{gj}-x\right)^{\top}\right\}.
\end{aligned}$$

Here, we defined $\boldsymbol{\Omega}_{12} = \boldsymbol{\Omega}_{21}^{\top}$. By Lemma 4 and 5,

$$\begin{aligned}
\Omega_{11} &= H_0(x) + 2I_0(x) \\
&= \frac{1}{h^d}\left\{f(x)\sigma^2(x)R_k^d + \lambda R_k^{d_{\text{cls}}} f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right) + o_p(1)\right\}, \quad (53)\\
\boldsymbol{\Omega}_{21} &= H_1(x) + 2I_1(x) = o_p\left(h^{-d+1}\right)\mathbf{1}_d, \quad (54)\\
\boldsymbol{\Omega}_{22} &= H_2(x) + 2I_2(x) = O_p\left(h^{-d+2}\right)\mathbf{1}_d\mathbf{1}_d^{\top}. \quad (55)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\operatorname{Var}\left[\hat{m}_{\text{LL}}(x) \mid \mathbf{X}_1, \cdots, \mathbf{X}_G\right] \\
= \; & \frac{1}{n}\mathbf{e}_1^{\top}\left(\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{D}_h\begin{bmatrix}\Omega_{11} & \boldsymbol{\Omega}_{12}\\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22}\end{bmatrix}\mathbf{D}_h\left(\frac{1}{n}\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\mathbf{D}_h\right)^{-1}\mathbf{e}_1 \\
= \; & \frac{1}{n}\begin{bmatrix} f(x)^{-1} + o_p(1) \\ o_p\left(h^{-1}\right)\mathbf{1}_d\end{bmatrix}^{\top}\begin{bmatrix}\Omega_{11} & o_p\left(h^{-d+1}\right)\mathbf{1}_d^{\top}\\ o_p\left(h^{-d+1}\right)\mathbf{1}_d & O_p\left(h^{-d+2}\right)\mathbf{1}_d\mathbf{1}_d^{\top}\end{bmatrix}\begin{bmatrix} f(x)^{-1} + o_p(1) \\ o_p\left(h^{-1}\right)\mathbf{1}_d\end{bmatrix} \\
= \; & \frac{1}{n}\left\{f(x)^{-1} + o_p(1)\right\}^2\Omega_{11} + o_p\left(n^{-1}h^{-d+1}\right) \\
= \; & \frac{R_k^d\sigma^2(x)}{f(x)nh^d} + \frac{\lambda R_k^{d_{\text{cls}}} f_2\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)\sigma\left(x^{(\text{ind})}, x^{(\text{ind})}; x^{(\text{cls})}\right)}{f(x)^2nh^d} + o_p\left(\frac{1}{nh^d}\right).
\end{aligned}$$

where the second equality follows from (51), (54), and (55) and the last equality follows from (53). □

A.8. **Proof for Theorem 8.**

Proof. Let $\boldsymbol{\mathcal{E}} = [e_1, \ldots, e_n]^{\top}$. Then, by Theorem 6,

$$\begin{aligned}
\hat{m}_{\text{LL}}(x) &= \mathbf{e}_1^{\top}\left(\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{Y} \\
&= \mathbf{e}_1^{\top}\left(\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{M} + \mathbf{e}_1^{\top}\left(\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^{\top}\mathbf{W}_x\boldsymbol{\mathcal{E}} \\
&= m(x) + o_p(1) + \mathbf{e}_1^{\top}\left(\mathbf{D}_h\mathbf{X}_x^{\top}\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{D}_h\mathbf{X}_x^{\top}\mathbf{W}_x\boldsymbol{\mathcal{E}}.
\end{aligned}$$

Here,

$$\begin{aligned}
\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^{\top}\mathbf{W}_x\boldsymbol{\mathcal{E}} &= \mathbf{D}_h\begin{bmatrix}\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj} - x\right)e_{gj} \\ \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj} - x\right)\left(X_{gj} - x\right)e_{gj}\end{bmatrix} = \mathbf{D}_h\begin{bmatrix}\mathcal{E}_0(x)\\ \mathcal{E}_1(x)\end{bmatrix} \\
&= \begin{bmatrix} O_p\left(\sqrt{\frac{1}{nh^d}}\right) \\ h^{-2}O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)\mathbf{1}_d\end{bmatrix}. \quad (56)
\end{aligned}$$

Thus, (51) and (56) together imply that

$$
\begin{aligned}
&\mathbf{e}_1^\top \left(\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\boldsymbol{\mathcal{E}}\\
=\;& \mathbf{e}_1^\top \begin{bmatrix} f(x)^{-1}+o_p(1) & o_p\left(h\right)\mathbf{1}_d^\top \\ o_p\left(h^{-1}\right)\mathbf{1}_d & \left(f(x)\kappa_2 I_{d\times d}\right)^{-1}+o_p\left(1\right)\mathbf{1}_d\mathbf{1}_d^\top \end{bmatrix}\begin{bmatrix} O_p\left(\sqrt{\frac{1}{nh^d}}\right) \\ h^{-1}O_p\left(\sqrt{\frac{1}{nh^d}}\right)\mathbf{1}_d \end{bmatrix}\\
=\;& O_p\left(\sqrt{\frac{1}{nh^d}}\right)+o_p\left(\sqrt{\frac{1}{nh^d}}\right)=o_p(1).
\end{aligned}
$$

Hence, $\hat{m}_{\mathrm{LL}}\left(x\right)\xrightarrow{p} m\left(x\right)$. □

A.9. **Proof for Theorem 9.**

Proof. Theorem 6 and $\hat{m}_{\mathrm{LL}}(x)=\mathbf{e}_1^\top\left(\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^\top\mathbf{W}_x\left\{\mathbf{M}+\boldsymbol{\mathcal{E}}\right\}$ together imply that

$$
\begin{aligned}
\sqrt{nh^d}\left(\hat{m}_{\mathrm{LL}}(x)-m(x)-h^2B_{\mathrm{LL}}(x)\right) &= \mathbf{e}_1^\top\left(\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\sqrt{nh^d}\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\boldsymbol{\mathcal{E}}+\sqrt{nh^d}o_p\left(h^2\right)\\
&= \mathbf{e}_1^\top\left(\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\sqrt{nh^d}\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\boldsymbol{\mathcal{E}}+o_p\left(1\right),
\end{aligned}
$$

where the second equality follows from $nh^{d+4}=O(1)$.

Equations (51) and (56) together imply that the first term on the displayed equation will be

$$
\begin{aligned}
&\mathbf{e}_1^\top\left(\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\sqrt{nh^d}\frac{1}{n}\mathbf{D}_h\mathbf{X}_x^\top\mathbf{W}_x\boldsymbol{\mathcal{E}}\\
=\;& \mathbf{e}_1^\top \begin{bmatrix} f(x)^{-1}+o_p(1) & o_p\left(h\right)\mathbf{1}_d^\top \\ o_p\left(h^{-1}\right)\mathbf{1}_d & \left(f(x)\kappa_2 I_{d\times d}\right)^{-1}+o_p\left(1\right)\mathbf{1}_d\mathbf{1}_d^\top \end{bmatrix}\begin{bmatrix} \sqrt{nh^d}\mathcal{E}_0(x) \\ \sqrt{nh^d}h^{-2}O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)\mathbf{1}_d \end{bmatrix}\\
=\;& \left\{f(x)^{-1}+o_p(1)\right\}\sqrt{nh^d}\mathcal{E}_0(x)+o_p\left(1\right)\\
=\;& \frac{\sqrt{nh^d}\mathcal{E}_0(x)}{f(x)}+o_p\left(1\right).
\end{aligned}
$$

We conclude with a similar argument to the proof of Theorem 5. □

A.10. **Proof for Theorem 10.**

Proof. We will show the theorem by the following three steps. The proof modifies time series results (Theorem 2 of Hansen (2008); Theorem 4.1 of Vogt (2012)) to the cluster sampling case.

Let $\tau_n=C_\tau n^{1/s}$, where $C_\tau>0$ will be chosen in Step 1 below.[4] Decompose $\hat{\psi}\left(x\right)$ into the tail $\hat{\psi}_2\left(x\right)$ and the truncated part $\hat{\psi}_1\left(x\right)$.

[4] The choice of τ_n is different from Hansen (2008). For discussions on it, the reader can refer to the proof of Lemma B-1 in Cattaneo, Crump and Jansson (2013) and the proof of Theorem 4.1 in Vogt (2012).

$$\begin{aligned}\hat{\psi}(x) &= \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h(X_{gj}-x) W_{gj}\mathbf{1}\{|W_{gj}|\leq\tau_n\} \\ &\quad+\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h(X_{gj}-x) W_{gj}\mathbf{1}\{|W_{gj}|>\tau_n\} \\ &\equiv \hat{\psi}_1(x)+\hat{\psi}_2(x).\end{aligned}$$

Then,

$$\hat{\psi}(x)-\mathbb{E}\left[\hat{\psi}(x)\right]=\hat{\psi}_1(x)-\mathbb{E}\left[\hat{\psi}_1(x)\right]+\hat{\psi}_2(x)-\mathbb{E}\left[\hat{\psi}_2(x)\right].$$

Step 1: Evaluate the tail part $\hat{\psi}_2(x)-\mathbb{E}\left[\hat{\psi}_2(x)\right]$

The tail part has the following bounds. Pick any $\varepsilon>0$.

$$\begin{aligned}\mathbb{P}\left(\sup_x\left|\hat{\psi}_2(x)\right|>a_n\right) &\leq \mathbb{P}(|W_i|>\tau_n \text{ for some } i) \\ &\leq n\mathbb{P}(|W|>\tau_n) \\ &\leq n\mathbb{E}[|W|^s]\tau_n^{-s} \\ &\leq nB_1\tau_n^{-s}\leq B_1/C_\tau^s,\end{aligned}$$

where the first inequality follows from the construction of $\hat{\psi}_2(x)$, the second inequality follows from the union bound, the third inequality follows from Markov's inequality, the fourth inequality follows from (19), and the last equality follows from the definition of τ_n. Then, we can choose a large enough number C_τ such that $\mathbb{P}\left(\sup_x\left|\hat{\psi}_2(x)\right|>a_n\right)\leq\varepsilon$. Hence, $\left|\hat{\psi}_2(x)\right|=O_p(a_n)$ uniformly. Note that C_τ depends on ε, but does not on n.

Also,

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{\psi}_2(x)\right|\right] \\ \leq\;& \frac{1}{h^d}\int_{\mathbb{R}^d}K\left(\frac{X-x}{h}\right)\mathbb{E}[|W|\mathbf{1}\{|W|>\tau_n\}\mid X]f(X)\,\mathrm{d}X \\ \leq\;& \int_{\mathbb{R}^d}K(T)\mathbb{E}[|W|\mathbf{1}\{|W|>\tau_n\}\mid X=x+hT]f(x+hT)\,\mathrm{d}T \\ \leq\;& \frac{1}{\tau_n^{s-1}}\int_{\mathbb{R}^d}K(T)\mathbb{E}[|W|^s\mathbf{1}\{|W|>\tau_n\}\mid X=x+hT]f(x+hT)\,\mathrm{d}T \\ \leq\;& \frac{1}{\tau_n^{s-1}}B_2=O(a_n),\end{aligned}$$

where the fourth inequality follows uniformly from (20), and the last equality follows from

$$\begin{aligned}\frac{1}{\tau_n^{s-1}} &= O\left(n^{1/s-1}\right) \\ &\leq O(a_n). \qquad \because (21)\end{aligned}$$

In the next two steps, we evaluate the truncated part $\hat{\psi}_1(x)-\mathbb{E}\left[\hat{\psi}_1(x)\right]$.

Step 2: Bound the supremum over $\|x\|\leq c_n$ with the maximum over a finite grid

We can cover the region $\left\{x \in \mathbb{R}^d : \|x\| \leq c_n\right\}$ with $N_{\text{ball}} \leq c_n^d h^{-d} a_n^{-d}$ balls

$$B_{a_n h}\left(x_k\right) = \left\{x \in \mathbb{R}^d : \|x - x_k\| \leq a_n h\right\},$$

where x_k is the midpoint of $B_{a_n h}\left(x_k\right)$. Assumption 9 implies that for all $\|x - x'\| \leq a \leq L$, there exists some function $K^*(\cdot)$ and some constant $A > 0$ such that

$$\left|K(x) - K(x')\right| \leq aAK^*(x') \tag{57}$$

where $K^*(u) = \prod_{q=1}^d k^*\left(u^{(q)}\right)$ and $k^*(\cdot)$ satisfies the definition of the kernel function (Definition 1). To construct such functions, we can define

$$\begin{aligned} K^*(u) &\equiv \prod_{q=1}^{d} 1/(4L)\mathbf{1}\left\{\left|u^{(q)}\right| \leq 2L\right\} \\ &\equiv \prod_{q=1}^{d} k^*\left(u^{(q)}\right) \end{aligned}$$

and set $A = 4^d L^d \Lambda$. Also, let $K_h^*(\cdot) = \frac{1}{h^d} K^*\left(\frac{\cdot}{h}\right)$.[5]

Then, for any $x \in B_{a_n h}\left(x_k\right)$ equation (57) implies

$$\left|K\left(\frac{X_{gj} - x}{h}\right) - K\left(\frac{X_{gj} - x_k}{h}\right)\right| \leq a_n A K^*\left(\frac{X_{gj} - x_k}{h}\right) \tag{58}$$

since

$$\left\|\frac{X_{gj} - x}{h} - \frac{X_{gj} - x_k}{h}\right\| = \frac{\|x - x_k\|}{h} \leq a_n,$$

and $a_n \leq L$ for large enough n.

Define $\widetilde{\psi}_1(x)$ by replacing $K_h(\cdot)$ on $\hat{\psi}_1(x)$ with $K_h^*(\cdot)$,

$$\widetilde{\psi}_1(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^*\left(X_{gj} - x\right) W_{gj}\mathbf{1}\left\{\left|W_{gj}\right| \leq \tau_n\right\}.$$

Then, $\mathbb{E}\left[\widetilde{\psi}_1(x)\right]$ is bounded since

$$\begin{aligned} \mathbb{E}\left[\widetilde{\psi}_1(x)\right] &= \frac{1}{h^d}\mathbb{E}\left[K^*\left(\frac{X_{gj} - x}{h}\right)\mathbb{E}\left[\left|W_{gj}\right|\mathbf{1}\left\{\left|W_{gj}\right| \leq \tau_n\right\} \mid X_{gj}\right]\right] \\ &\leq \int_{\mathbb{R}^d} K^*\left(u_{gj}\right)\mathbb{E}\left[\left|W_{gj}\right| \mid X_{gj} = x + hu_{gj}\right] f\left(x + hu_{gj}\right) \mathrm{d}u_{gj} \\ &\leq B_3 \int_{\mathbb{R}^d} K^*\left(u_{gj}\right) \mathrm{d}u_{gj} = B_3 < \infty, \end{aligned}$$

[5] Under Assumption 9,

$$\begin{aligned} \left|K(x) - K(x')\right| &\leq \Lambda\left\|x - x'\right\|\mathbf{1}\left\{\left\|x'\right\| \leq 2L\right\} \leq a\Lambda\mathbf{1}\left\{\left\|x'\right\| \leq 2L\right\} \\ &\leq a\Lambda\prod_{q=1}^{d}\mathbf{1}\left\{\left|x'^{(q)}\right| \leq 2L\right\} = a\Lambda 4^d L^d K^*\left(x'\right), \end{aligned}$$

where the first inequality follows from the support of K, the second inequality follows from $\|x - x'\| \leq a$, and the third inequality follows from $\left|x'^{(q)}\right| \leq \sqrt{\sum_{p=1}^d \left(x'^{(p)}\right)^2} = \|x'\|$.
Since k^* is bounded, symmetric, and has finite moments, it satisfies the definition of the kernel function.

where B_3 exists by (20). Thus, $A\mathbb{E}\left[\widetilde{\psi}_1(x)\right] < M$ for large enough M, and within each ball $B_{a_n h}(x_k)$,

$$\begin{aligned}
& \sup_{x\in B_{a_n h}(x_k)} \left|\hat{\psi}_1(x) - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| \\
= & \sup_{x\in B_{a_n h}(x_k)} \left|\hat{\psi}_1(x) - \hat{\psi}_1(x_k) + \hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right] + \mathbb{E}\left[\hat{\psi}_1(x_k)\right] - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| \\
\leq & \left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| + \sup_{x\in B_{a_n h}(x_k)} \left|\hat{\psi}_1(x) - \hat{\psi}_1(x_k)\right| + \sup_{x\in B_{a_n h}(x_k)} \left|\mathbb{E}\left[\hat{\psi}_1(x_k)\right] - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| \\
\leq & \left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| + a_n A\left\{\left|\widetilde{\psi}_1(x_k)\right| + \mathbb{E}\left[\left|\widetilde{\psi}_1(x_k)\right|\right]\right\} \\
\leq & \left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| + a_n A\left|\widetilde{\psi}_1(x_k) - \mathbb{E}\left[\widetilde{\psi}_1(x_k)\right]\right| + 2a_n A\mathbb{E}\left[\left|\widetilde{\psi}_1(x_k)\right|\right] \\
\leq & \left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| + \left|\widetilde{\psi}_1(x_k) - \mathbb{E}\left[\widetilde{\psi}_1(x_k)\right]\right| + 2a_n M,
\end{aligned}$$

where the first and third inequalities follow from the triangle inequality, the second inequality follows from (58), and the last inequality comes from $a_n \leq A^{-1}$ for large enough n and $A\mathbb{E}\left[\widetilde{\psi}_1(x)\right] < M$.

As a consequence,

$$\begin{aligned}
& \mathbb{P}\left[\sup_{\|x\|\leq c_n} \left|\hat{\psi}_1(x) - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| > 4Ma_n\right] \\
\leq & N_{\text{ball}} \max_{1\leq k\leq N_{\text{ball}}} \mathbb{P}\left[\sup_{x\in B_{a_n h}(x_k)} \left|\hat{\psi}_1(x) - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| > 4Ma_n\right] \\
\leq & N_{\text{ball}} \max_{1\leq k\leq N_{\text{ball}}} \left\{\mathbb{P}\left[\left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| > Ma_n\right] + \mathbb{P}\left[\left|\widetilde{\psi}_1(x_k) - \mathbb{E}\left[\widetilde{\psi}_1(x_k)\right]\right| > Ma_n\right]\right\}.
\end{aligned}$$

Since we can evaluate both of $\mathbb{P}\left[\left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| > Ma_n\right]$ and $\mathbb{P}\left[\left|\widetilde{\psi}_1(x_k) - \mathbb{E}\left[\widetilde{\psi}_1(x_k)\right]\right| > Ma_n\right]$ in the same way, we will focus on $\mathbb{P}\left[\left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| > Ma_n\right]$ in the next step.

Step 3: Apply the Bernstein's inequality.

Define

$$\widetilde{\mathbf{U}}_g = \sum_{j=1}^{n_g}\left\{K\left(\frac{X_{gj}-x}{h}\right)W_{gj}\mathbf{1}\left\{|W_{gj}|\leq\tau_n\right\} - \mathbb{E}\left[K\left(\frac{X_{gj}-x}{h}\right)W_{gj}\mathbf{1}\left\{|W_{gj}|\leq\tau_n\right\}\right]\right\}.$$

Then,

$$\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right] = \frac{1}{nh^d}\sum_{g=1}^{G}\widetilde{\mathbf{U}}_g.$$

Since

$$\left|K\left(\frac{X_{gj}-x}{h}\right)W_{gj}\mathbf{1}\left\{|W_{gj}|\leq\tau_n\right\} - \mathbb{E}\left[K\left(\frac{X_{gj}-x}{h}\right)W_{gj}\mathbf{1}\left\{|W_{gj}|\leq\tau_n\right\}\right]\right| \leq 2\bar{K}\tau_n$$

and

$$\operatorname{Var}\left(\sum_{g=1}^{G}\widetilde{\mathbf{U}}_g\right) = n^2h^{2d}\operatorname{Var}\left(\hat{\psi}(x)\right) \leq nh^d\overline{V}, \qquad \because \text{Assumption 7}$$

the Bernstein's inequality for cluster sampling (Lemma 1) implies

$$\begin{aligned}
\mathbb{P}\left[\left|\hat{\psi}_1(x_k) - \mathbb{E}\left[\hat{\psi}_1(x_k)\right]\right| > Ma_n\right] &= \mathbb{P}\left[\left|\sum_{g=1}^{G} \widetilde{\mathbf{U}}_g\right| > Ma_n nh^d\right] \\
&\leq 2\exp\left\{-\frac{1}{2}\frac{M^2 a_n^2 n^2 h^{2d}}{nh^d\overline{V} + 2\left(\max_{g\leq G} n_g\right)\bar{K}\tau_n M a_n nh^d/3}\right\} \\
&\leq 2\exp\left\{-\frac{1}{2}\frac{M^2 a_n^2 nh^d}{\overline{V} + 2C_a\bar{K}M/3}\right\} \\
&= 2\exp\left\{-\frac{1}{2}\frac{M^2 \log n}{\overline{V} + 2C_a\bar{K}M/3}\right\} \\
&\leq 2\exp\left\{-\frac{6M\log n}{3 + 2C_a\bar{K}}\right\} \\
&= 2n^{-6M/\left(3+2C_a\bar{K}\right)},
\end{aligned}$$

where the second inequality with some $C_a > 0$ follows from $\left(\max_{g\leq G} n_g\right)\tau_n a_n = O(1)$ by (21), the second equality follows from $a_n^2 = \log n/(nh^d)$, the third inequality follows by choosing $M > \overline{V}$. Thus,

$$\begin{aligned}
\mathbb{P}\left[\sup_{\|x\|\leq c_n}\left|\hat{\psi}_1(x) - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| > 4Ma_n\right] &\leq 4N_{\text{ball}}n^{-6M/\left(3+2C_a\bar{K}\right)} \\
&\leq O(T_n), \qquad (59)
\end{aligned}$$

where $T_n = c_n^d h^{-d} a_n^{-d} n^{-6M/\left(3+2C_a\bar{K}\right)}$. We can evaluate

$$\begin{aligned}
c_n^d h^{-d} &= O\left(\frac{\left(\max_{g\leq G} n_g\right)^2 \log n}{h^d}\right) \qquad \because (22) \\
&= O\left(n^{1-(2/s)}\right) \qquad \because (21)
\end{aligned}$$

and

$$a_n^{-d} = \left(\frac{nh^d}{\log n}\right)^{d/2} = o\left(n^{d/2}\right).$$

Thus,

$$\begin{aligned}
T_n &= o\left(n^{1-(2/s)+(d/2)-(6M/(3+2C_a\bar{K}))}\right) \\
&\leq o(1),
\end{aligned}$$

where the inequality holds for large enough M. Therefore, (59) implies $\sup_{\|x\|\leq c_n}\left|\hat{\psi}_1(x) - \mathbb{E}\left[\hat{\psi}_1(x)\right]\right| = O_p(a_n)$. □

A.11. **Proof for Lemma 1.**

Proof. By the triangle inequality, $\left|\widetilde{\mathbf{Y}}_g\right| = \left|\sum_{j=1}^{n_g} Y_{gj}\right| \leq n_g B$. Thus,

$$\max_{g\leq G}\left|\widetilde{\mathbf{Y}}_g\right| \leq \left(\max_{g\leq G} n_g\right) B.$$

The result follows from the standard Bernstein's inequality for the independent and zero mean random variables $\widetilde{\mathbf{Y}}_1, \ldots, \widetilde{\mathbf{Y}}_G$. □

A.12. Proof for Theorem 11.

Proof. As the proof for Lemma 2, we can prove that

$$\begin{aligned} \operatorname{Var}\left[\hat{f}(x)\right] &= \operatorname{Var}\left[F_0(x)\right] \\ &\leq O\left(n^{-1}h^{-d}\right) + O\left(\frac{1}{n}\left(\max_g n_g\right)\right) = O\left(\frac{1}{nh^d}\right). \end{aligned}$$

Under Assumption 10, this bound holds uniformly for any $x \in \mathbb{R}^d$. Thus, Assumption 7 for $\hat{\psi}(x) = \hat{f}(x)$ with $W_{gj} = 1$ is satisfied. Since we also have Assumption 8 with $s = \infty$, all assumptions for Theorem 10 are satisfied. Hence,

$$\sup_{\|x\| \leq c_n} \left|\hat{f}(x) - \mathbb{E}\left[\hat{f}(x)\right]\right| = O_p(a_n). \tag{60}$$

As the proof for Lemma 2, we can also show

$$\sup_{x \in \mathbb{R}^d} \left|\mathbb{E}\left[\hat{f}(x)\right] - f(x)\right| = O\left(h^2\right), \tag{61}$$

where we have the sup bound under Assumption 10. The triangle inequality, (60), and (61) together imply the result. □

A.13. Proof for Theorem 12.

Proof. **For the case $\widehat{m}_*(x) = \widehat{m}_{\mathrm{nw}}(x)$.**

First, Theorem 11 implies

$$\sup_{\|x\| \leq c_n} \left|\frac{\hat{f}(x)}{f(x)} - 1\right| \leq \frac{\sup_{\|x\| \leq c_n}\left|\hat{f}(x) - f(x)\right|}{\inf_{\|x\| \leq c_n} f(x)} = O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right). \tag{62}$$

Next, define

$$\widehat{\phi}(x) = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj} - x\right) Y_{gj}.$$

Then,

$$\begin{aligned} & \operatorname{Var}\left[\widehat{\phi}(x)\right] \\ = \ & \operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K\left(X_{gj} - x\right)\left\{m\left(X_{gj}\right) - m(x) + e_{gj} + m(x)\right\}\right] \\ = \ & \operatorname{Var}\left[J_0(x) + m(x)F_0(x) + \mathcal{E}_0(x)\right] \\ \leq \ & \left(\sqrt{\operatorname{Var}\left[J_0(x)\right]} + m(x)\sqrt{\operatorname{Var}\left[F_0(x)\right]} + \sqrt{\operatorname{Var}\left[\mathcal{E}_0(x)\right]}\right)^2, \end{aligned}$$

where the inequality follows since the absolute value of covariance is bonded by the product of the square root of variances. By the similar way as in the proof of Lemmas 2, 3, and 6, we can evaluate

$$\begin{aligned}
\mathrm{Var}\left[F_0(x)\right] &\leq O\left(n^{-1}h^{-d}\right) + O\left(n^{-1}\left(\max_g n_g\right)\right) \leq O\left(\frac{1}{nh^d}\right), \\
\mathrm{Var}\left[J_0(x)\right] &\leq O\left(\frac{h^2}{nh^d}\right), \\
\mathrm{Var}\left[\mathcal{E}_0(x)\right] &\leq O\left(\frac{1}{nh^d}\right).
\end{aligned}$$

Under Assumption 10, these bounds hold uniformly for any $x \in \mathbb{R}$. Combining these equations and the uniform boundedness of $m(x)$, we have

$$\mathrm{Var}\left[\widehat{\phi}\left(x\right)\right] \leq O\left(\frac{1}{nh^d}\right)$$

uniformly for any $x \in \mathbb{R}^d$. Then, all assumptions for Theorem 10 are satisfied. Hence,

$$\sup_{\|x\| \leq c_n} \left|\widehat{\phi}\left(x\right) - \mathbb{E}\left[\widehat{\phi}\left(x\right)\right]\right| = O_p\left(a_n\right). \tag{63}$$

Also,

$$\begin{aligned}
& \sup_{\|x\| \leq c_n} \left|\mathbb{E}\left[\widehat{\phi}\left(x\right)\right] - m(x)f(x)\right| \\
= \;& \sup_{\|x\| \leq c_n} \left|\mathbb{E}\left[J_0(x)\right] + \mathbb{E}\left[\mathcal{E}_0(x)\right] + m(x)\mathbb{E}\left[F_0(x)\right] - m(x)f(x)\right| \\
\leq \;& \sup_{\|x\| \leq c_n} \left|\mathbb{E}\left[J_0(x)\right]\right| + \sup_{\|x\| \leq c_n} \left|\mathbb{E}\left[\mathcal{E}_0(x)\right]\right| + \sup_{\|x\| \leq c_n} \left|m(x)\right| \sup_{\|x\| \leq c_n} \left|\mathbb{E}\left[F_0(x)\right] - f(x)\right| \\
\leq \;& O\left(h^2\right) + 0 + O\left(1\right)O\left(h^2\right) \\
= \;& O\left(h^2\right),
\end{aligned} \tag{64}$$

where the first inequality follows from the triangle inequality, the second inequality can be shown as in the proof of Lemmas 2, 3, and 6, and Assumption 10 implies these bounds hold uniformly for any $x \in \mathbb{R}^d$. Hence,

$$\begin{aligned}
& \sup_{\|x\|\le c_n} \left|\widehat{m}_{\mathrm{nw}}(x) - m(x)\right| \\
= & \sup_{\|x\|\le c_n} \left|\frac{\widehat{\phi}(x)}{f(x)} \cdot \frac{f(x)}{\widehat{f}(x)} - m(x)\right| \\
= & \sup_{\|x\|\le c_n} \left|\left(\frac{\widehat{\phi}(x)}{f(x)} - \frac{m(x)\widehat{f}(x)}{f(x)}\right)\frac{f(x)}{\widehat{f}(x)}\right| \\
\le & \sup_{\|x\|\le c_n} \left|\frac{\widehat{\phi}(x)}{f(x)} - \frac{m(x)\widehat{f}(x)}{f(x)}\right| \sup_{\|x\|\le c_n}\left|\frac{f(x)}{\widehat{f}(x)}\right| \\
\le & \sup_{\|x\|\le c_n} \left|\widehat{\phi}(x) - m(x)\widehat{f}(x)\right| \delta_n^{-1}\left\{1 + O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)\right\} \\
\le & \left\{\sup_{\|x\|\le c_n}\left|\widehat{\phi}(x) - \mathbb{E}\left[\widehat{\phi}(x)\right]\right| + \sup_{\|x\|\le c_n}\left|\mathbb{E}\left[\widehat{\phi}(x)\right] - m(x)f(x)\right| + \sup_{\|x\|\le c_n}\left|m(x)f(x) - m(x)\widehat{f}(x)\right|\right\} \\
& \times\delta_n^{-1}\left\{1 + O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)\right\} \\
\le & \left\{O_p(a_n) + O\left(h^2\right) + O_p(a_n)\right\}\delta_n^{-1}\left\{1 + O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)\right\} \\
\le & O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)\left\{1 + O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)\right\} \\
= & O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right),
\end{aligned}$$

where the second inequality follows from (26) and (62), the third inequality follows from the triangle inequality, and the fourth inequality follows from (63), the uniform boundedness of $m(x)$, the result of Theorem 11, and (64).

For the case $\widehat{m}_*(x) = \widehat{m}_{\mathrm{LL}}(x)$.

Using the partition matrix inversion, we can rewrite

$$\begin{aligned}
\widehat{m}_{\mathrm{LL}}(x) &= \mathbf{e}_1^\top\left(\mathbf{X}_x^\top\mathbf{W}_x\mathbf{X}_x\right)^{-1}\mathbf{X}_x^\top\mathbf{W}_x\mathbf{Y} \\
&= \frac{\widehat{f}(x)\widehat{m}_{\mathrm{nw}}(x) - S(x)^\top M(x)^{-1}N(x)}{\widehat{f}(x) - S(x)^\top M(x)^{-1}S(x)},
\end{aligned} \tag{65}$$

where

$$\begin{aligned}
S(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj} - x\right)\left(\frac{X_{gj} - x}{h}\right), \\
M(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj} - x\right)\left(\frac{X_{gj} - x}{h}\right)\left(\frac{X_{gj} - x}{h}\right)^\top, \\
N(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj} - x\right)\left(\frac{X_{gj} - x}{h}\right)Y_{gj}.
\end{aligned}$$

Define

$$S^{(q)}(x) \equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(\frac{X_{gj}^{(q)}-x^{(q)}}{h}\right) = h^{-1}F_1^{(q)}(x),$$

$$\begin{aligned} M^{(p,q)}(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(\frac{X_{gj}^{(p)}-x^{(p)}}{h}\right)\left(\frac{X_{gj}^{(q)}-x^{(q)}}{h}\right)^{\top} \\ &= \begin{cases} h^{-2}F_2^{(q)}(x) & \text{if } p=q \\ h^{-2}F^{(p,q)}(x) & \text{if } p\neq q \end{cases}, \end{aligned}$$

$$\begin{aligned} N^{(q)}(x) &\equiv \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}-x\right)\left(\frac{X_{gj}^{(q)}-x^{(q)}}{h}\right)\left\{m\left(X_{gj}\right)-m(x)+m(x)+e_{gj}\right\} \\ &= h^{-1}J_1^{(q)}(x)+h^{-1}m(x)F_1^{(q)}(x)+h^{-1}\mathcal{E}_1^{(q)}(x). \end{aligned}$$

Similar way as in the proof of Lemmas 2, 3, and 6, we can evaluate

$$\operatorname{Var}\left[h^{-1}F_1^{(q)}\right] \leq O\left(n^{-1}h^{-d}\right)+O\left(n^{-1}\left(\max_g n_g\right)\right)=O\left(n^{-1}h^{-d}\right),$$

$$\operatorname{Var}\left[h^{-2}F_2^{(q)}\right] \leq O\left(n^{-1}h^{-d}\right)+O\left(n^{-1}\left(\max_g n_g\right)\right)=O\left(n^{-1}h^{-d}\right),$$

$$\operatorname{Var}\left[h^{-2}F^{(p,q)}(x)\right] \leq O\left(n^{-1}h^{-d}\right)+O\left(n^{-1}\left(\max_g n_g\right)\right)=O\left(n^{-1}h^{-d}\right),$$

$$\operatorname{Var}\left[h^{-1}J_1^{(q)}(x)\right] \leq O\left(n^{-1}h^{2-d}\right)+O\left(n^{-1}\left(\max_g n_g\right)h^2\right)=O\left(n^{-1}h^{-d}\right),$$

and

$$\operatorname{Var}\left[h^{-1}\mathcal{E}_1^{(q)}(x)\right]=O\left(n^{-1}h^{-d}\right).$$

Since these bounds are uniform for any $x\in\mathbb{R}$ under Assumption 10 and the compact kernel function enables us to treat $(X_{gj}^{(q)}-x^{(q)})/h$ as bounded in $S^{(q)}(x)$, $M^{(p,q)}(x)$, and $N^{(q)}(x)$, we can apply Theorem 10:

$$\sup_{\|x\|\leq c_n}\left|S^{(q)}(x)-\mathbb{E}\left[h^{-1}F_1^{(q)}(x)\right]\right| = \sup_{\|x\|\leq c_n}\left|S^{(q)}(x)-h\partial_q f(x)\kappa_2+O\left(h^2\right)\right|=O_p\left(a_n\right),$$

$$\sup_{\|x\|\leq c_n}\left|M^{(q,q)}(x)-\mathbb{E}\left[h^{-2}F_2^{(q)}(x)\right]\right| = \sup_{\|x\|\leq c_n}\left|M^{(q,q)}(x)-f(x)\kappa_2+O\left(h^2\right)\right|=O_p\left(a_n\right),$$

$$\sup_{\|x\|\leq c_n}\left|M^{(p,q)}(x)-\mathbb{E}\left[h^{-2}F^{(p,q)}(x)\right]\right| = \sup_{\|x\|\leq c_n}\left|M^{(p,q)}(x)+O\left(h^2\right)\right|=O_p\left(a_n\right),$$

and

$$\begin{aligned} &\sup_{\|x\|\leq c_n}\left|N^{(q)}(x)-\mathbb{E}\left[N^{(q)}(x)\right]\right| \\ =\ &\sup_{\|x\|\leq c_n}\left|N^{(q)}(x)-hf(x)\partial_q m(x)\kappa_2-hm(x)\partial_q f(x)\kappa_2-0+O\left(h^2\right)\right|=O_p\left(a_n\right). \end{aligned}$$

By element-wise comparisons, we obtain

$$\begin{aligned}
S(x) &= h\kappa_2 \nabla f\left(x\right) + O_p\left(a_n + h^2\right)\mathbf{1}_d,\\
M(x) &= f(x)\kappa_2 \mathbf{I}_{d\times d} + O_p\left(a_n + h^2\right)\mathbf{1}_d\mathbf{1}_d^\top,\\
N(x) &= h\kappa_2 \nabla\left\{f\left(x\right)m\left(x\right)\right\} + O_p\left(a_n + h^2\right)\mathbf{1}_d,
\end{aligned}$$

where asymptotic orders are uniform over $\|x\| \le c_n$.

Therefore, by the same matrix calculations as Hansen (2008), we obtain

$$\widehat{m}_{\mathrm{LL}}(x) = m(x) + O_p\left(\delta_n^{-1}\left(a_n + h^2\right)\right)$$

uniform over $\|x\| \le c_n$. □

A.14. **Proof for Theorem 13.**

Proof. Necessary condition is

$$\frac{\partial}{\partial h}\mathrm{AIMSE} = 4h^3\bar{B} - \frac{dR_k^d\bar{\sigma}^2}{nh^{d+1}} = 0.$$

We obtain h_0 by solving this equation since by

$$\frac{\partial}{\partial h^2}\mathrm{AIMSE} = 12h^2\bar{B} + \frac{d(d+1)R_k^d\bar{\sigma}^2}{nh^{d+2}} > 0,$$

the first-order condition is sufficient. □

A.15. **Proof for Theorem 14.**

Proof. For any g and j,

$$\begin{aligned}
\mathbb{E}\left[\tilde{e}_{gj}\left(h\right)^2 w\left(X_{gj}\right)\right] &= \mathbb{E}\left[e_{gj}^2 w\left(X_{gj}\right)\right] + \mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\}^2 w\left(X_{gj}\right)\right]\\
&\quad + 2\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\} e_{gj} w\left(X_{gj}\right)\right]\\
&\overset{\text{(i)}}{=} \overline{\sigma}_w^2 + \mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\}^2 w\left(X_{gj}\right)\right]\\
&\overset{\text{(ii)}}{=} \overline{\sigma}_w^2 + \mathbb{E}_{-g}\left[\int_{\mathbb{R}^d}\left\{m\left(x\right) - \widetilde{m}_{-g}\left(x, h\right)\right\}^2 f\left(x\right) w\left(x\right)\mathrm{d}x\right]
\end{aligned}$$

where (i) follows from the definition of $\overline{\sigma}_w^2$ and

$$\begin{aligned}
\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\} e_{gj} w\left(X_{gj}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\} e_{gj} w\left(X_{gj}\right) \mid \mathbf{X}_g\right]\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\} w\left(X_{gj}\right) \mid \mathbf{X}_g\right]\mathbb{E}\left[e_{gj} \mid \mathbf{X}_g\right]\right]\\
&= 0
\end{aligned}$$

since $\widetilde{m}_{-g}\left(X_{gj}, h\right)$ is independent of e_{gj} after conditioning X_g, and (ii) follows from

$$\begin{aligned}
\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\}^2 w\left(X_{gj}\right)\right] &= \mathbb{E}_{-g}\left[\mathbb{E}\left[\left\{m\left(X_{gj}\right) - \widetilde{m}_{-g}\left(X_{gj}, h\right)\right\}^2 w\left(X_{gj}\right) \mid \mathbf{Y}_{-g}, \mathbf{X}_{-g}\right]\right]\\
&= \mathbb{E}_{-g}\left[\int_{\mathbb{R}^d}\left\{m\left(x\right) - \widetilde{m}_{-g}\left(x, h\right)\right\}^2 f\left(x\right) w\left(x\right)\mathrm{d}x\right].
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{CV}(h)\right] &= \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[\tilde{e}_{gj}\left(h\right)^2 w\left(X_{gj}\right)\right] \\
&= \overline{\sigma}_w^2 + \sum_{g=1}^{G}\frac{n_g}{n}\mathbb{E}_{-g}\left[\int_{\mathbb{R}^d}\left\{m\left(x\right)-\widetilde{m}_{-g}\left(x,h\right)\right\}^2 f\left(x\right)w\left(x\right)\mathrm{d}x\right] \\
&= \overline{\sigma}_w^2 + \mathrm{IMSE}_{G-1}(h).
\end{aligned}$$

□

A.16. **Proof for Theorem 15.**

Proof. We can interpret (39) as a standard nonparametric density estimator. Under Assumption 11, Theorem 1 is applicable for $(2d_{\text{ind}}+d_{\text{cls}})$-dimensional regressors and $n_g(n_g-1)/2$ size clusters. □

A.17. **Proof for Theorem 16.**

Proof. First, we show that the feasible estimator of conditional variance can be asymptotically replaced with the infeasible estimator, i.e.,

$$\left|\widehat{\sigma}_{\text{nw}}^2\left(x\right)-\widehat{\sigma}_{\text{nw}}^{2*}\left(x\right)\right| = o_p(1). \tag{66}$$

Here,

$$\begin{aligned}
\left|\widehat{\sigma}_{\text{nw}}^2\left(x\right)-\widehat{\sigma}_{\text{nw}}^{2*}\left(x\right)\right| &\le \frac{\left|\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)\left(\widehat{e}_{gj}^2-e_{gj}^2\right)\right|}{\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}-x\right)} \\
&\le \max_g\max_j\left|\widehat{e}_{gj}^2-e_{gj}^2\right|.
\end{aligned}$$

and

$$\begin{aligned}
&\max_g\max_j\left|\widehat{e}_{gj}^2-e_{gj}^2\right| \\
\le\;& \max_g\max_j\left|\left\{e_{gj}+m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}^2-e_{gj}^2\right| \\
=\;& \max_g\max_j\left|2e_{gj}\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}+\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}^2\right| \\
\le\;& 2\max_g\max_j\left|e_{gj}\right|\cdot\max_g\max_j\left|\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}\right| \\
&+\left\{\max_g\max_j\left|m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right|\right\}^2.
\end{aligned}$$

Pick any $\varepsilon > 0$. By Theorem 12 and Assumption 11 (v),

$$\begin{aligned}
& \Pr\left(\max_g \max_j \left|\{m(X_{gj}) - \widehat{m}_*(X_{gj})\}\right| > \varepsilon\right) \\
\leq\; & \Pr\left(\max_g \max_j \left|\{m(X_{gj}) - \widehat{m}_*(X_{gj})\}\right| > \varepsilon \mid \|X_{gj}\| \leq c_n\right) \Pr\left(\|X_{gj}\| \leq c_n\right) \\
& + o(1) \\
\leq\; & \Pr\left(\sup_{\|x\|\leq c_n} \left|m(x) - \widehat{m}_*(x)\right| > \varepsilon\right) + o(1) \\
\leq\; & o(1).
\end{aligned}$$

We also know that assumptions for Theorem 12 imply

$$\max_g \max_j |Y_{gj}| = o_p\left(n^{-1/s}\right),$$

and

$$\max_g \max_j |m(X_{gj})| = O(1),$$

thus

$$\begin{aligned}
\max_g \max_j |e_{gj}| &= \max_g \max_j |Y_{gj}| + \max_g \max_j |m(X_{gj})| \\
&\leq O_p(1).
\end{aligned}$$

Hence,

$$\max_g \max_j \left|\widehat{e}_{gj}^2 - e_{gj}^2\right| = o_p(1).$$

Thus, it is sufficient to show that

$$\widehat{\sigma}^{2*}_{\mathrm{nw}}(x) \xrightarrow{p} \sigma^2(x). \tag{67}$$

Let $v_{gj} = e_{gj}^2 - \sigma^2(X_{gj})$. Since $\sigma^2(x) = \mathbb{E}\left[e^2 \mid X = x\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[v_{gj} \mid \mathbf{X}_g\right] &= 0, \\
\mathbb{E}\left[v_{gj}^2 \mid \mathbf{X}_g\right] &= \mathbb{E}\left[v_{gj}^2 \mid X_{gj}\right] = \mathbb{E}\left[\left\{e_{gj}^2 - \sigma^2(X_{gj})\right\}^2 \mid X_{gj}\right] = \mathbb{E}\left[e_{gj}^4 \mid X_{gj}\right] - \left\{\sigma^2(X_{gj})\right\}^2 \\
&= \varsigma^2(X_{gj}) - \left\{\sigma^2(X_{gj})\right\}^2, \\
\mathbb{E}\left[v_{gj} v_{g\ell} \mid \mathbf{X}_g\right] &= \mathbb{E}\left[v_{gj} v_{g\ell} \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right] \\
&= \mathbb{E}\left[\left\{e_{gj}^2 - \sigma^2(X_{gj})\right\}\left\{e_{g\ell}^2 - \sigma^2(X_{g\ell})\right\} \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right] \\
&= \mathbb{E}\left[e_{gj}^2 e_{g\ell}^2 \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right] - \sigma^2(X_{g\ell})\, \sigma^2(X_{g\ell}) \\
&= \varsigma\left(X_{gj}^{(\mathrm{ind})}, X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right) - \sigma^2(X_{g\ell})\, \sigma^2(X_{g\ell}).
\end{aligned}$$

Under Assumption 11, we can apply Theorem 4 after replacing $m(x)$ with $\sigma^2(x)$ and obtain (67). □

A.18. **Proof for Theorem 17.**

Proof. We will first show

$$\left|\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)-\widehat{\sigma}_{\mathrm{nw}}^{2}\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)\right|=o_p(1), \tag{68}$$

and then show that

$$\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)\xrightarrow{p}\sigma\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right). \tag{69}$$

For the first step,

$$\begin{aligned}
&\left|\widehat{\sigma}_{\mathrm{nw}}\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)-\widehat{\sigma}_{\mathrm{nw}}^{2}\left(x^{(\mathrm{ind})},x^{(\mathrm{ind})};x^{(\mathrm{cls})}\right)\right|\\
\leq\;&\frac{\left|\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g}K\left(\frac{\left(X_{gj}^{(\mathrm{ind})\top},X_{g\ell}^{(\mathrm{ind})\top},X_{g}^{(\mathrm{cls})\top}\right)^{\top}-\left(x^{(\mathrm{ind})\top},x^{(\mathrm{ind})\top},x^{(\mathrm{cls})\top}\right)^{\top}}{b}\right)\left(\widehat{e}_{gj}\widehat{e}_{g\ell}-e_{gj}e_{g\ell}\right)\right|}{\sum_{g:n_g\geq 2}\sum_{1\leq j<\ell\leq n_g}K\left(\frac{\left(X_{gj}^{(\mathrm{ind})\top},X_{g\ell}^{(\mathrm{ind})\top},X_{g}^{(\mathrm{cls})\top}\right)^{\top}-\left(x^{(\mathrm{ind})\top},x^{(\mathrm{ind})\top},x^{(\mathrm{cls})\top}\right)^{\top}}{b}\right)}\\
\leq\;&\max_g\max_{j,\ell}\left|\widehat{e}_{gj}\widehat{e}_{g\ell}-e_{gj}e_{g\ell}\right|,
\end{aligned}$$

and

$$\begin{aligned}
&\max_g\max_{j,\ell}\left|\widehat{e}_{gj}\widehat{e}_{g\ell}-e_{gj}e_{g\ell}\right|\\
\leq\;&\max_g\max_{j,\ell}\left|\left\{e_{gj}+m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}\left\{e_{g\ell}+m\left(X_{g\ell}\right)-\widehat{m}_*\left(X_{g\ell}\right)\right\}-e_{gj}e_{g\ell}\right|\\
\leq\;&\max_g\max_{j,\ell}\left|e_{gj}\left\{m\left(X_{g\ell}\right)-\widehat{m}_*\left(X_{g\ell}\right)\right\}\right|\\
&+\max_g\max_{j,\ell}\left|e_{g\ell}\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}\right|\\
&+\max_g\max_{j,\ell}\left|\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}\left\{m\left(X_{g\ell}\right)-\widehat{m}_*\left(X_{g\ell}\right)\right\}\right|\\
\leq\;&2\max_g\max_{j}\left|e_{gj}\right|\cdot\max_g\max_{j}\left|\left\{m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right\}\right|\\
&+\max_g\max_{j}\left|m\left(X_{gj}\right)-\widehat{m}_*\left(X_{gj}\right)\right|^2
\end{aligned}$$

Thus, similarly to the proof of Theorem 16, we can show that

$$\max_g\max_{j,\ell}\left|\widehat{e}_{gj}\widehat{e}_{g\ell}-e_{gj}e_{g\ell}\right|=o_p(1),$$

and (68) is shown.

Next, let's prove (69). Let $u_{gj\ell}=e_{gj}e_{g\ell}-\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right)$. Since $\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right)=\mathbb{E}\left[e_{gj}e_{g\ell}\mid X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[u_{gj\ell}\mid\mathbf{X}_g\right]&=0,\\
\mathbb{E}\left[u_{gj\ell}^2\mid\mathbf{X}_g\right]&=\mathbb{E}\left[u_{gj\ell}^2\mid X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right]\\
&=\mathbb{E}\left[\left\{e_{gj}e_{g\ell}-\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right)\right\}^2\mid X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right]\\
&=\mathbb{E}\left[e_{gj}^2e_{g\ell}^2\mid X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right]-\sigma^2\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right)\\
&=\varsigma\left(X_{gj}^{(\mathrm{ind})},X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right)-\sigma^2\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_{g}^{(\mathrm{cls})}\right),
\end{aligned}$$

and for $(j,\ell) \neq (t,s)$,

$$
\begin{aligned}
& \mathbb{E}\left[u_{gj\ell}u_{gts} \mid \mathbf{X}_g\right] \\
= \; & \mathbb{E}\left[u_{gj\ell}u_{gts} \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}, X_{gt}^{(\mathrm{ind})}, X_{gs}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right] \\
= \; & \mathbb{E}\left[\left\{e_{gj}e_{g\ell} - \sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right)\right\}\right. \\
& \quad \left. \times \left\{e_{gt}e_{gs} - \sigma\left(X_{gt}^{(\mathrm{ind})}, X_{gs}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right)\right\} \mid X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}, X_{gt}^{(\mathrm{ind})}, X_{gs}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right] \\
= \; & \varsigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}, X_{gt}^{(\mathrm{ind})}, X_{gs}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right) \\
& -\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right)\sigma\left(X_{gt}^{(\mathrm{ind})}, X_{gs}^{(\mathrm{ind})}; X_{g}^{(\mathrm{cls})}\right).
\end{aligned}
$$

Under Assumption 11, we can apply Theorem 4 for $(2d_{\mathrm{ind}} + d_{\mathrm{cls}})$-dimensional regressors and $n_g(n_g-1)/2$ size clusters. □

A.19. **Proof for Corollary 1.**

Proof. Apply the Slutsky's Lemma. □

Appendix B. Proofs for technical lemmas

For the following proofs, we focus on the case $x = \left(x^{(\mathrm{ind})\top}, x^{(\mathrm{cls})\top}\right)^\top = 0$ to make notation lighter. We also suppress subscripts such as g and j if the meaning is implied from the context.

B.1. Proof for Lemma 2.

Proof. Define $F_r^{(q)} = \frac{1}{n}\sum_{g=1}^G \sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r$, $F^{(p,q)} = \frac{1}{n}\sum_{g=1}^G \sum_{j=1}^{n_g} K_h\left(X_{gj}\right) X_{gj}^{(p)} X_{gj}^{(q)}$ for $p \neq q$, and $\nu_r^{(q)} = \int_{\mathbb{R}^d} K\left(T\right)\left(T^{(q)}\right)^r \mathrm{d}T = \int_{-\infty}^{\infty} k\left(T^{(q)}\right)\left(T^{(q)}\right)^r \mathrm{d}T^{(q)}$ for $r = 0, 1, 2$. [6] Note that

$$\nu_r^{(q)} = \begin{cases} 1 & \text{if } r = 0 \\ 0 & \text{if } r = 1 \\ \kappa_2 & \text{if } r = 2 \end{cases}.$$

We will evaluate expectations and variances of $F^{(p,q)}$ and $F_r^{(q)}$ and obtain a conclusion by Markov's inequality. For expectations, we have

$$\begin{aligned}
\mathbb{E}\left[F_r^{(q)}\right] &= h^r \int_{\mathbb{R}^d} K\left(T\right)\left(T^{(q)}\right)^r f\left(hT\right)\mathrm{d}T \\
&= h^r \int_{\mathbb{R}^d} K\left(T\right)\left(T^{(q)}\right)^r \left\{f\left(0\right) + hT^\top \nabla f\left(0\right) + \frac{h^2}{2} T^\top \nabla^2 f\left(h\tilde{T}\right) T\right\}\mathrm{d}T \\
&= h^r \int_{\mathbb{R}^d} K\left(T\right)\left(T^{(q)}\right)^r \left\{f\left(0\right) + hT^\top \nabla f\left(0\right)\right\}\mathrm{d}T + O\left(h^{r+2}\right) \\
&= \begin{cases} h^r f\left(0\right)\nu_r^{(q)} + O\left(h^{r+2}\right) & \text{if } r \text{ is even} \\ h^{r+1}\partial_q f\left(0\right)\int_{-\infty}^{\infty} k\left(T^{(q)}\right)\left(T^{(q)}\right)^{r+1}\mathrm{d}T^{(q)} + O\left(h^{r+1}\right) & \text{if } r \text{ is odd} \end{cases},
\end{aligned}$$

by the identical marginal distribution, the change of variables $T = X/h$, the Taylor expansion ($\tilde{T}$ is between 0 and T), the dominated convergence theorem, and the symmetry of the kernel function.[7] Thus,

$$\mathbb{E}\left[F_r^{(q)}\right] = \begin{cases} f\left(0\right) + o\left(1\right) & \text{if } r = 0 \\ o\left(h\right) & \text{if } r = 1 \\ h^2 f\left(0\right)\kappa_2 + o\left(h^2\right) & \text{if } r = 2 \end{cases}.$$

Similarly, for $p \neq q$,

$$\begin{aligned}
\mathbb{E}\left[F^{(p,q)}\right] &= h^2 \int_{\mathbb{R}^d} K\left(T\right) T^{(p)} T^{(q)} f\left(hT\right)\mathrm{d}T \\
&= h^2 \int_{\mathbb{R}^d} K\left(T\right) T^{(p)} T^{(q)} \left\{f\left(0\right) + hT^\top \nabla f\left(0\right) + \frac{h^2}{2} T^\top \nabla^2 f\left(0\right) T\right\}\mathrm{d}T + o\left(h^4\right) \\
&= h^2 f\left(0\right)\int_{\mathbb{R}^d} K\left(T\right) T^{(p)} T^{(q)}\mathrm{d}T + h^3 \int_{\mathbb{R}^d} K\left(T\right) T^{(p)} T^{(q)} T^\top \nabla f\left(0\right)\mathrm{d}T + O\left(h^4\right) \\
&= O\left(h^4\right).
\end{aligned}$$

[6] When $r = 0$, $F_r^{(q)}$ does not depend on q because $\left(X_{gj}^{(q)}\right)^r = 1$.

[7] We use the continuity of $\nabla^2 f\left(x\right)$ in some neighborhood $\mathcal{N}$ of $x = 0$. The continuity implies $\nabla^2 f\left(h\tilde{T}\right) \to \nabla^2 f\left(0\right)$ as $h \to 0$. Since $\nabla^2 f\left(0\right)$ exists, it is bounded. Thus, we can apply the dominated convergence theorem.

For variances,

$$
\begin{aligned}
& \operatorname{Var}\left[F_r^{(q)}\right] \\
= & \operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right] = \frac{1}{n^2}\sum_{g=1}^{G}\operatorname{Var}\left[\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right] \\
= & \frac{1}{n^2}\sum_{g=1}^{G}\left\{\sum_{j=1}^{n_g}\operatorname{Var}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right] + 2\sum_{1\leq j<\ell\leq n_g}\operatorname{Cov}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r, K_h\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]\right\} \\
\leq & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[K_h^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^{2r}\right] \\
& + \frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}\left(\mathbb{E}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r K_h\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]\right. \\
& \left. - \underbrace{\mathbb{E}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right]\mathbb{E}\left[K_h\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]}_{=\mathbb{E}\left[F_r^{(q)}\right]^2}\right),
\end{aligned}
$$

where the second equality follows from the independence between clusters and the inequality follows from $\operatorname{Var}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right] \leq \mathbb{E}\left[K_h^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^{2r}\right]$. We will bound the following two expectations

$$
\mathbb{E}\left[K_h^2(X)\left(X^{(q)}\right)^{2r}\right], \tag{70}
$$

$$
\mathbb{E}\left[K_h\left(X_j\right)\left(X_j^{(q)}\right)^r K_h\left(X_\ell\right)\left(X_\ell^{(q)}\right)^r\right]. \tag{71}
$$

$$
\begin{aligned}
(70): \quad \mathbb{E}\left[K_h^2(X)\left(X^{(q)}\right)^{2r}\right] &= \frac{1}{h^{2d}}\int_{\mathbb{R}^d} K\left(\frac{X}{h}\right)^2\left(X^{(q)}\right)^{2r} f(X)\,\mathrm{d}X \\
= \quad & \frac{1}{h^{d-2r}}\int_{\mathbb{R}^d} K(T)^2\left(T^{(q)}\right)^{2r} f(Th)\,\mathrm{d}T \\
= \quad & \frac{1}{h^{d-2r}}\int_{\mathbb{R}^d} K(T)^2\left(T^{(q)}\right)^{2r} f(0)\mathrm{d}T + o\left(h^{2r-d}\right) \\
= \quad & O\left(h^{2r-d}\right),
\end{aligned}
$$

where the second equality follows from the change of variables $T = X/h$ and the third equality follows from the continuity.

Also,

$$\begin{aligned}
(71): \quad & \mathbb{E}\left[K_h\left(X_j\right)\left(X_j^{(q)}\right)^r K_h\left(X_\ell\right)\left(X_\ell^{(q)}\right)^r\right] \\
= \quad & \frac{1}{h^{2d}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(\frac{X_j}{h}\right) K\left(\frac{X_\ell}{h}\right)\left(X_j^{(q)}\right)^r\left(X_\ell^{(q)}\right)^r \\
& \qquad \times f_2\left(X_j^{(\text{ind})}, X_\ell^{(\text{ind})} ; X^{(\text{cls})}\right) \mathrm{d} X_j^{(\text{ind})} \mathrm{d} X_\ell^{(\text{ind})} \mathrm{d} X^{(\text{cls})} \\
= \quad & h^{2r-d_{\text{cls}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(T_j\right) K\left(T_\ell\right)\left(T_j^{(q)}\right)^r\left(T_\ell^{(q)}\right)^r \\
& \qquad \times f_2\left(h T_j^{(\text{ind})}, h T_\ell^{(\text{ind})} ; h T^{(\text{cls})}\right) \mathrm{d} T_j^{(\text{ind})} \mathrm{d} T_\ell^{(\text{ind})} \mathrm{d} T^{(\text{cls})} \\
= \quad & h^{2r-d_{\text{cls}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(T_j\right) K\left(T_\ell\right)\left(T_j^{(q)}\right)^r\left(T_\ell^{(q)}\right)^r \\
& \qquad \times f_2(0,0 ; 0) \mathrm{d} T_j^{(\text{ind})} \mathrm{d} T_\ell^{(\text{ind})} \mathrm{d} T^{(\text{cls})}+o\left(h^{2r-d_{\text{cls}}}\right) \\
= \quad & O\left(h^{2r-d_{\text{cls}}}\right),
\end{aligned}$$

where the second equality follows from the change of variables $T_j^{(\text{ind})} = X_j^{(\text{ind})}/h$, $T_\ell^{(\text{ind})} = X_\ell^{(\text{ind})}/h$, and $T^{(\text{cls})} = X^{(\text{cls})}/h$ (we define $T_j = \left(T_j^{(\text{ind})\top}, T^{(\text{cls})\top}\right)^\top$, $T_\ell = \left(T_\ell^{(\text{ind})\top}, T^{(\text{cls})\top}\right)^\top$), and the third equality follows from the continuity.

Thus, since $\sum_{g=1}^G n_g = n$ and $\left(\max_g n_g\right) / \left(n h^{d_{\text{cls}}}\right) = \left(\max_g n_g h^{d_{\text{ind}}}\right) / \left(n h^d\right) = o(1)$,

$$\begin{aligned}
\operatorname{Var}\left[F_r^{(q)}\right] & \leq \frac{1}{n^2} \sum_{g=1}^G \sum_{j=1}^{n_g} O\left(h^{2r-d}\right)+\frac{1}{n^2} \sum_{g=1}^G \sum_{1 \leq j<\ell \leq n_g} O\left(h^{2r-d_{\text{cls}}}\right) \\
& \leq O\left(n^{-1} h^{2r-d}\right)+\frac{1}{n}\left(\max_g n_g\right) O\left(h^{2r-d_{\text{cls}}}\right)=o\left(h^{2r}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \operatorname{Var}\left[F^{(p,q)}\right] \\
\leq \; & \frac{1}{n^2} \sum_{g=1}^G \sum_{j=1}^{n_g} \mathbb{E}\left[K_h^2\left(X_{gj}\right)\left(X_{gj}^{(p)}\right)^2\left(X_{gj}^{(q)}\right)^2\right] \\
& +\frac{2}{n^2} \sum_{g=1}^G \sum_{1 \leq j<\ell \leq n_g}\left(\mathbb{E}\left[K_h\left(X_{gj}\right) X_{gj}^{(p)} X_{gj}^{(q)} K_h\left(X_{g\ell}\right) X_{g\ell}^{(p)} X_{g\ell}^{(q)}\right]\right. \\
& \left.\qquad\qquad -\underbrace{\mathbb{E}\left[K_h\left(X_{gj}\right) X_{gj}^{(p)} X_{gj}^{(q)}\right] \mathbb{E}\left[K_h\left(X_{g\ell}\right) X_{g\ell}^{(p)} X_{g\ell}^{(q)}\right]}_{=\mathbb{E}\left[F^{(p,q)}\right]^2}\right) \\
\leq \; & \frac{1}{n^2} \sum_{g=1}^G \sum_{j=1}^{n_g} O\left(h^{4-d}\right)+\frac{1}{n^2} \sum_{g=1}^G \sum_{1 \leq j<\ell \leq n_g} O\left(h^{4-d_{\text{cls}}}\right) \\
\leq \; & O\left(n^{-1} h^{4-d}\right)+O\left(n^{-1}\left(\max_g n_g\right) h^{4-d_{\text{cls}}}\right)=o\left(h^4\right).
\end{aligned}$$

Therefore, by Markov's inequality and Jensen's inequality,

$$\begin{aligned}
\mathbb{P}\left[\left|h^2\left(F_2^{(q)}-h^2 f(0)\kappa_2\right)\right|>\delta\right] &\leq \frac{\mathbb{E}\left[\left|F_2^{(q)}-h^2 f(0)\kappa_2\right|\right]}{h^2\delta} \leq \frac{\mathbb{E}\left[\left(F_2^{(q)}-h^2 f(0)\kappa_2\right)^2\right]^{1/2}}{h^2\delta} \\
&= \frac{\left|\mathbb{E}\left[F_2^{(q)}\right]-h^2 f(0)\kappa_2\right|+\sqrt{\operatorname{Var}\left[F_2^{(q)}\right]}}{h^2\delta} \\
&\leq o(1) \qquad \text{for any } \delta,
\end{aligned}$$

which implies that $F_2^{(q)}=h^2 f(0)\kappa_2+o_p\left(h^2\right)$. Similarly, we have $F_0^{(q)}=f(0)+o_p(1)$, $F_1^{(q)}=o_p(h)$, and $F^{(p,q)}=o_p\left(h^2\right)$. We conclude by element-wise comparisons. □

B.2. **Proof for Lemma 3.**

Proof. Define $J_r^{(q)}=\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r$ for $r=0,1$. For expectations,

$$\begin{aligned}
&\mathbb{E}\left[J_r^{(q)}\right] \\
=\ &\mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r\right] \\
=\ &\frac{1}{h^d}\int_{\mathbb{R}^d}K\left(\frac{X}{h}\right)\left\{m(X)-m(0)\right\}\left(X^{(q)}\right)^r f(X)\,\mathrm{d}X \\
=\ &h^r\int_{\mathbb{R}^d}K(T)\left\{m(hT)-m(0)\right\}\left(T^{(q)}\right)^r f(hT)\,\mathrm{d}T \\
=\ &h^r\int_{\mathbb{R}^d}K(T)\left(T^{(q)}\right)^r\left\{hT^\top\nabla m(0)+\frac{h^2}{2}T^\top\nabla^2 m\left(h\tilde{T}\right)T\right\}\left\{f(0)+hT^\top\nabla f\left(h\grave{T}\right)\right\}\mathrm{d}T \\
=\ &h^{r+1}f(0)\int_{\mathbb{R}^d}\left(T^{(q)}\right)^r T^\top\nabla m(0)K(T)\,\mathrm{d}T \\
&+\frac{h^{r+2}}{2}f(0)\int_{\mathbb{R}^d}\left(T^{(q)}\right)^r T^\top\nabla^2 m(0)TK(T)\,\mathrm{d}T \\
&+h^{r+2}\int_{\mathbb{R}^d}\left(T^{(q)}\right)^r T^\top\nabla m(0)T^\top\nabla f(0)K(T)\,\mathrm{d}T+O\left(h^{r+3}\right)+o\left(h^{r+2}\right) \\
=\ &\begin{cases} h^2\sum_{q=1}^{d}\left\{\frac{1}{2}f(0)\partial_{qq}m(0)+\partial_q m(0)\partial_q f(0)\right\}\kappa_2+o\left(h^2\right) & \text{if } r=0 \\ h^2 f(0)\partial_q m(0)\kappa_2+o\left(h^3\right) & \text{if } r=1 \end{cases},
\end{aligned}$$

where the second equality follows from the linearity of the expectation and the identical marginal distribution, the third equality follows from the change of variables $T=X/h$, the fourth equality follows from the Taylor expansion ($\tilde{T}$ and $\grave{T}$ are between 0 and T), the fifth equality follows from the dominated convergence theorem, and the sixth equality follows from the symmetry of the kernel function.

For variances,

$$
\begin{aligned}
& \operatorname{Var}\left[J_r^{(q)}\right] \\
= \; & \operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r\right] \\
= \; & \frac{1}{n^2}\sum_{g=1}^{G}\operatorname{Var}\left[\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r\right] \\
= \; & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\operatorname{Var}\left[K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r\right] \\
& +2\frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\operatorname{Cov}\left[K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r, K_h\left(X_{g\ell}\right)\left\{m\left(X_{g\ell}\right)-m(0)\right\}\left(X_{g\ell}^{(q)}\right)^r\right] \\
\le \; & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[K_h^2\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}^2\left(X_{gj}^{(q)}\right)^{2r}\right] \\
& +\frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\mathbb{E}\left[K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r K_h\left(X_{g\ell}\right)\left\{m\left(X_{g\ell}\right)-m(0)\right\}\left(X_{g\ell}^{(q)}\right)^r\right] \\
& -\frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\underbrace{\mathbb{E}\left[K_h\left(X_{gj}\right)\left\{m\left(X_{gj}\right)-m(0)\right\}\left(X_{gj}^{(q)}\right)^r\right]\mathbb{E}\left[K_h\left(X_{g\ell}\right)\left\{m\left(X_{g\ell}\right)-m(0)\right\}\left(X_{g\ell}^{(q)}\right)^r\right]}_{=\mathbb{E}\left[J_r^{(q)}\right]^2},
\end{aligned}
$$

where the second equality follows from the independence between clusters. We will bound the following two expectations

$$
\mathbb{E}\left[K_h^2(X)\left\{m(X)-m(0)\right\}^2\left(X^{(q)}\right)^{2r}\right], \tag{72}
$$

$$
\mathbb{E}\left[K_h\left(X_j\right)\left\{m\left(X_j\right)-m(0)\right\}\left(X_j^{(q)}\right)^r K_h\left(X_\ell\right)\left\{m\left(X_\ell\right)-m(0)\right\}\left(X_\ell^{(q)}\right)^r\right]. \tag{73}
$$

$$
\begin{aligned}
(72): \; & \mathbb{E}\left[K_h^2(X)\left\{m(X)-m(0)\right\}^2\left(X^{(q)}\right)^{2r}\right] \\
= \; & \frac{1}{h^{2d}}\int_{\mathbb{R}^d} K\left(\frac{X}{h}\right)^2\left\{m(X)-m(0)\right\}^2\left(X^{(q)}\right)^{2r} f(X)\,\mathrm{d}X \\
= \; & \frac{1}{h^{d-2r}}\int_{\mathbb{R}^d} K(T)^2\left(T^{(q)}\right)^{2r}\left\{m(hT)-m(0)\right\}^2 f(hT)\,\mathrm{d}T \\
= \; & \frac{1}{h^{d-2(r+1)}}\int_{\mathbb{R}^d} K(T)^2\left(T^{(q)}\right)^{2r}\left\{T^\top\nabla m(0)\right\}^2 f(0)\,\mathrm{d}T+o\left(h^{2(r+1)-d}\right) \\
= \; & O\left(h^{2(r+1)-d}\right).
\end{aligned}
$$

where the second equality follows from the change of variables $T = X/h$, and the third equality follows from the Taylor expansion and the dominated convergence theorem.

Also,

$$
\begin{aligned}
(73): \quad & \mathbb{E}\left[K_h\left(X_j\right)\left\{m\left(X_j\right)-m(0)\right\}\left(X_j^{(q)}\right)^r K_h\left(X_\ell\right)\left\{m\left(X_\ell\right)-m(0)\right\}\left(X_\ell^{(q)}\right)^r\right] \\
= \quad & \frac{1}{h^{2d}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(\frac{X_j}{h}\right) K\left(\frac{X_\ell}{h}\right)\left(X_j^{(q)}\right)^r\left(X_\ell^{(q)}\right)^r \\
& \quad \times\left\{m\left(X_j\right)-m(0)\right\}\left\{m\left(X_{g\ell}\right)-m(0)\right\} \\
& \quad \times f_2\left(X_j^{(\text{ind})}, X_\ell^{(\text{ind})} ; X^{(\text{cls})}\right) \mathrm{d}X_j^{(\text{ind})} \mathrm{d}X_\ell^{(\text{ind})} \mathrm{d}X^{(\text{cls})} \\
= \quad & h^{2r-d_{\text{cls}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(T_j\right) K\left(T_\ell\right)\left(T_j^{(q)}\right)^r\left(T_\ell^{(q)}\right)^r \\
& \quad \times\left\{m\left(hT_j\right)-m(0)\right\}\left\{m\left(hT_\ell\right)-m(0)\right\} \\
& \quad \times f_2\left(hT_j^{(\text{ind})}, hT_\ell^{(\text{ind})} ; hT^{(\text{cls})}\right) \mathrm{d}T_j^{(\text{ind})} \mathrm{d}T_\ell^{(\text{ind})} \mathrm{d}T^{(\text{cls})} \\
= \quad & h^{2r+2-d_{\text{cls}}} f_2(0,0;0) \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{ind}}}} \int_{\mathbb{R}^{d_{\text{cls}}}} K\left(T_j\right) K\left(T_\ell\right)\left(T_j^{(q)}\right)^r\left(T_\ell^{(q)}\right)^r \\
& \quad \times\left\{T_j^{\top} \nabla m(0)\right\}\left\{T_\ell^{\top} \nabla m(0)\right\} \mathrm{d}T_j^{(\text{ind})} \mathrm{d}T_\ell^{(\text{ind})} \mathrm{d}T^{(\text{cls})} \\
& +o\left(h^{2r+2-d_{\text{cls}}}\right) \\
= \quad & \begin{cases} O\left(h^{2-d_{\text{cls}}}\right) & \text{if } r=0 \\ O\left(h^{4-d_{\text{cls}}}\right) & \text{if } r=1 \end{cases},
\end{aligned}
$$

where the second equality follows from the change of variables $T_j^{(\text{ind})} = X_j^{(\text{ind})}/h$, $T_\ell^{(\text{ind})} = X_\ell^{(\text{ind})}/h$, and $T^{(\text{cls})} = X^{(\text{cls})}/h$ (we define $T_j = \left(T_j^{(\text{ind})\top}, T^{(\text{cls})\top}\right)^\top$, $T_\ell = \left(T_\ell^{(\text{ind})\top}, T^{(\text{cls})\top}\right)^\top$), and the third equality follows from the Taylor expansion and the dominated convergence theorem. Thus, for $r=0$,

$$
\begin{aligned}
\operatorname{Var}\left[J_0^{(q)}\right] & = \frac{1}{n^2} \sum_{g=1}^G\left[\sum_{j=1}^{n_g} O\left(h^{2-d}\right)+2 \sum_{1 \leq j<\ell \leq n_g} O\left(h^{2-d_{\text{cls}}}\right)\right] \\
& \leq O\left(n^{-1} h^{2-d}\right)+O\left(n^{-1}\left(\max_g n_g\right) h^{2-d_{\text{cls}}}\right) \\
& = O\left(\frac{h^2}{nh^d}\right)+\left\{\left(\max_{g \leq G} n_g\right) h^{d_{\text{ind}}}\right\} O\left(\frac{h^2}{nh^d}\right) \\
& = O\left(\frac{h^2}{nh^d}\right),
\end{aligned}
$$

and for $r=1$,

$$
\begin{aligned}
\operatorname{Var}\left[J_1^{(q)}\right] \quad & \leq \quad \frac{1}{n^2} \sum_{g=1}^G\left[\sum_{j=1}^{n_g} O\left(h^{4-d}\right)+2 \sum_{1 \leq j<\ell \leq n_g} O\left(h^{4-d_{\text{cls}}}\right)\right] \\
& \leq \quad O\left(\frac{h^4}{nh^d}\right).
\end{aligned}
$$

Therefore, by Markov's inequality, $J_0^{(q)} = h^2\kappa_2 \sum_{q=1}^{d} \left\{\frac{1}{2}f(0)\partial_{qq}m\left(0\right) + \partial_q m\left(0\right)\partial_q f\left(0\right)\right\} + o_p\left(h^2\right) + O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)$, $J_1^{(q)} = h^2\kappa_2 f(0)\partial_q m\left(0\right) + o_p\left(h^3\right) + O_p\left(\sqrt{\frac{1}{nh^{d-4}}}\right)$. We conclude by element-wise comparisons. □

B.3. **Proof for Lemma 4.**

Proof. Define $H_r^{(q)} = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r$, $H^{(p,q)} = \frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}\right)$ $\sigma^2\left(X_{gj}\right)X_{gj}^{(p)}X_{gj}^{(q)}$ for $p \neq q$, and $\pi_r^{(q)} = \int_{\mathbb{R}^d} K^2\left(T\right)\left(T^{(q)}\right)^r \mathrm{d}T$ for $r = 0, 1, 2$. For expectations,

$$\begin{aligned}
\mathbb{E}\left[H_r^{(q)}\right] &= \frac{1}{h^{2d}}\int_{\mathbb{R}^d} K^2\left(\frac{X}{h}\right)\sigma^2\left(X\right)\left(X^{(q)}\right)^r f\left(X\right)\mathrm{d}X \\
&= \frac{1}{h^{d-r}}\int_{\mathbb{R}^d} K^2\left(T\right)\sigma^2\left(hT\right)\left(T^{(q)}\right)^r f\left(hT\right)\mathrm{d}T \\
&= \frac{1}{h^{d-r}}\int_{\mathbb{R}^d} K^2\left(T\right)\left(T^{(q)}\right)^r \sigma^2\left(0\right)\left\{f\left(0\right) + hT^\top\nabla f\left(0\right)\right\}\mathrm{d}T \\
&\quad +o\left(h^{r+1-d}\right) \\
&= \begin{cases} \frac{1}{h^{d-r}} f\left(0\right)\sigma^2\left(0\right)\pi_r^{(q)} + o\left(h^{r+1-d}\right) & \text{if } r \text{ is even} \\ O\left(h^{r+1-d}\right) & \text{if } r \text{ is odd} \end{cases},
\end{aligned}$$

where the second equality follows from the change of variables $T = X/h$, and the third equality follows from the Taylor expansion. Since $\pi_0^{(q)} = R_k^d$,

$$\mathbb{E}\left[H_r^{(q)}\right] = \begin{cases} \frac{1}{h^d}\left\{f\left(0\right)\sigma^2\left(0\right)R_k^d + o\left(1\right)\right\} & \text{if } r = 0 \\ O\left(h^{-d+2}\right) & \text{if } r = 1 \\ \frac{1}{h^{d-2}} f\left(0\right)\sigma^2\left(0\right)\left\{\int_{\mathbb{R}^d} K^2\left(T\right)\left(T^{(q)}\right)^2 \mathrm{d}T\right\} + o\left(h^{-d+2}\right) & \text{if } r = 2 \end{cases}.$$

Similarly, for $p \neq q$,

$$\begin{aligned}
\mathbb{E}\left[H^{(p,q)}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)X_{gj}^{(p)}X_{gj}^{(q)}\right] = \frac{1}{h^{2d}}\int_{\mathbb{R}^d} K^2\left(\frac{X}{h}\right)\sigma^2\left(X\right)X^{(p)}X^{(q)}f\left(X\right)\mathrm{d}X \\
&= \frac{1}{h^{d-2}}\int_{\mathbb{R}^d} K^2\left(T\right)\sigma^2\left(hT\right)T^{(p)}T^{(q)}f\left(hT\right)\mathrm{d}T \\
&= \frac{1}{h^{d-2}} f\left(0\right)\sigma^2\left(0\right)\int_{\mathbb{R}^d} K^2\left(T\right)T^{(p)}T^{(q)}\mathrm{d}T + o\left(\frac{1}{h^{d-2}}\right) \\
&= o\left(h^{-d+2}\right).
\end{aligned}$$

For variances,

$$\begin{aligned}
&\operatorname{Var}\left[H_r^{(q)}\right]\\
=\ &\operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right]=\frac{1}{n^2}\sum_{g=1}^{G}\operatorname{Var}\left[\sum_{j=1}^{n_g}K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right]\\
=\ &\frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\operatorname{Var}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right]\\
&+2\frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\operatorname{Cov}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r,K_h^2\left(X_{g\ell}\right)\sigma^2\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]\\
\le\ &\frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[K_h^4\left(X_{gj}\right)\left(\sigma^2\left(X_{gj}\right)\right)^2\left(X_{gj}^{(q)}\right)^{2r}\right]\\
&+\frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^rK_h^2\left(X_{g\ell}\right)\sigma^2\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]\\
&-\frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\underbrace{\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\right]\mathbb{E}\left[K_h^2\left(X_{g\ell}\right)\sigma^2\left(X_{g\ell}\right)\left(X_{g\ell}^{(q)}\right)^r\right]}_{=\mathbb{E}\left[H_r^{(q)}\right]^2},
\end{aligned}$$

where the second equality follows from the independence between clusters. We will bound the following two expectations

$$\mathbb{E}\left[K_h^4\left(X\right)\left(\sigma^2\left(X\right)\right)^2\left(X^{(q)}\right)^{2r}\right],\tag{74}$$

$$\mathbb{E}\left[K_h^2\left(X_j\right)\sigma^2\left(X_j\right)\left(X_j^{(q)}\right)^rK_h^2\left(X_\ell\right)\sigma^2\left(X_\ell\right)\left(X_\ell^{(q)}\right)^r\right].\tag{75}$$

$$\begin{aligned}
(74):\ \ &\mathbb{E}\left[K_h^4\left(X\right)\left(\sigma^2\left(X\right)\right)^2\left(X^{(q)}\right)^{2r}\right]=\frac{1}{h^{4d}}\int_{\mathbb{R}^d}K\left(\frac{X}{h}\right)^4\left(\sigma^2\left(X\right)\right)^2\left(X^{(q)}\right)^{2r}f\left(X\right)\mathrm{d}X\\
=\ &\frac{1}{h^{3d-2r}}\int_{\mathbb{R}^d}K\left(T\right)^4\left(T^{(q)}\right)^{2r}\left(\sigma^2\left(hT\right)\right)^2f\left(hT\right)\mathrm{d}T\\
=\ &\frac{1}{h^{3d-2r}}\int_{\mathbb{R}^d}K\left(T\right)^4\left(T^{(q)}\right)^{2r}\left(\sigma^2\left(0\right)\right)^2\left\{f(0)+hT^\top\nabla f(0)\right\}\mathrm{d}T+o\left(h^{2r-3d+1}\right)\\
=\ &O\left(h^{2r-3d}\right),
\end{aligned}$$

where the second equality follows from the change of variables $T=X/h$, and the third equality follows from the Taylor expansion.

Also,

$$
\begin{aligned}
&(75): \quad \mathbb{E}\left[K_h^2\left(X_j\right)\sigma^2\left(X_j\right)\left(X_j^{(q)}\right)^r K_h^2\left(X_\ell\right)\sigma^2\left(X_\ell\right)\left(X_\ell^{(q)}\right)^r\right] \\
= \quad & \frac{1}{h^{4d}}\int_{\mathbb{R}^{d_{\text{ind}}}}\int_{\mathbb{R}^{d_{\text{ind}}}}\int_{\mathbb{R}^{d_{\text{cls}}}} K^2\left(\frac{X_j}{h}\right)\sigma^2\left(X_j\right)\left(X_j^{(q)}\right)^r K^2\left(\frac{X_\ell}{h}\right)\sigma^2\left(X_\ell\right)\left(X_\ell^{(q)}\right)^r \\
& \qquad \times f_2\left(X_j^{(\text{ind})}, X_\ell^{(\text{ind})}; X^{(\text{cls})}\right)\mathrm{d}X_j^{(\text{ind})}\mathrm{d}X_\ell^{(\text{ind})}\mathrm{d}X^{(\text{cls})} \\
= \quad & \frac{1}{h^{2d-2r+d_{\text{cls}}}}\int_{\mathbb{R}^{d_{\text{ind}}}}\int_{\mathbb{R}^{d_{\text{ind}}}}\int_{\mathbb{R}^{d_{\text{cls}}}} K^2\left(T_j\right)\sigma^2\left(hT_j\right)\left(T_j^{(q)}\right)^r K^2\left(T_\ell\right)\left(T_\ell^{(q)}\right)^r\sigma^2\left(hT_\ell\right) \\
& \qquad \times f_2\left(hT_j^{(\text{ind})}, hT_\ell^{(\text{ind})}; hT^{(\text{cls})}\right)\mathrm{d}T_j^{(\text{ind})}\mathrm{d}T_\ell^{(\text{ind})}\mathrm{d}T^{(\text{cls})} \\
= \quad & O\left(h^{2r-2d-d_{\text{cls}}}\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\operatorname{Var}\left[F_r^{(q)}\right] & \leq \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g} O\left(h^{2r-3d}\right) + \frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} O\left(h^{2r-2d-d_{\text{cls}}}\right) \\
& \leq O\left(n^{-1}h^{2r-3d}\right) + O\left(n^{-1}\left(\max_g n_g\right)h^{2r-2d-d_{\text{cls}}}\right) = o\left(h^{2r-2d}\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
& \operatorname{Var}\left[F^{(p,q)}\right] \\
\leq \quad & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[K_h^4\left(X_{gj}\right)\left(\sigma^2\left(X_{gj}\right)\right)^2\left(X_{gj}^{(q)}\right)^2\right] \\
& + \frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)X_{gj}^{(p)}X_{gj}^{(q)}K_h^2\left(X_{g\ell}\right)\sigma^2\left(X_{g\ell}\right)X_{g\ell}^{(p)}X_{g\ell}^{(q)}\right] \\
& - \frac{2}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}\underbrace{\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)X_{gj}^{(p)}X_{gj}^{(q)}\right]\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{g\ell}\right)X_{g\ell}^{(p)}X_{g\ell}^{(q)}\right]}_{=\mathbb{E}\left[H^{(p,q)}\right]^2} \\
\leq \quad & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g} O\left(h^{4-3d}\right) + \frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} O\left(h^{4-2d-d_{\text{cls}}}\right) \\
\leq \quad & O\left(n^{-1}h^{4-3d}\right) + O\left(n^{-1}\left(\max_g n_g\right)h^{4-2d-d_{\text{cls}}}\right) = o\left(h^{4-2d}\right).
\end{aligned}
$$

Therefore, by Markov's inequality, $H_0^{(q)} = \frac{1}{h^d}f(0)\sigma^2(0)R_k^d + o_p\left(h^{-d}\right)$, $H_1^{(q)} = o_p\left(h^{-d+1}\right)$, $H_2^{(q)} = \frac{1}{h^{d-2}}f(0)\sigma^2(0)\left\{\int_{\mathbb{R}^d}K^2(T)\left(T^{(q)}\right)^2\mathrm{d}T\right\} + o_p\left(h^{-d+2}\right)$, and $H^{(p,q)} = o_p\left(h^{-d+2}\right)$. We conclude by element-wise comparisons. □

B.4. **Proof for Lemma 5.**

Proof. Define $I_r^{(q)} = \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}; X_g^{(\text{cls})}\right)\left(X_{gj}^{(q)}\right)^r$ for $r = 0, 1$, and $I^{(p,q)} = \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})}, X_{g\ell}^{(\text{ind})}; X_g^{(\text{cls})}\right)X_{gj}^{(p)}X_{g\ell}^{(q)}$

for any p and q (allow $p = q$ here). For expectations,

$$\begin{aligned}\mathbb{E}\left[I_r^{(q)}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g} K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)\left(X_{gj}^{(q)}\right)^r\right]\\ &= \frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g} \mathbb{E}\left[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)\left(X_{gj}^{(q)}\right)^r\right],\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\left[I^{(p,q)}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g} K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)X_{gj}^{(p)}X_{g\ell}^{(q)}\right]\\ &= \frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g} \mathbb{E}\left[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)X_{gj}^{(p)}X_{g\ell}^{(q)}\right].\end{aligned}$$

We will evaluate

$$\mathbb{E}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right], \tag{76}$$

$$\mathbb{E}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)X_j^{(p)}X_\ell^{(q)}\right]. \tag{77}$$

Denote

$$\begin{aligned}\nabla_1 f_2\left(0,0;0\right) &= \left.\frac{\partial f_2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)}{\partial X_j^{(\mathrm{ind})}}\right|_{\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)=(0,0;0)},\\ \nabla_2 f_2\left(0,0;0\right) &= \left.\frac{\partial f_2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)}{\partial X_\ell^{(\mathrm{ind})}}\right|_{\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)=(0,0;0)},\\ \nabla_c f_2\left(0,0;0\right) &= \left.\frac{\partial f_2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)}{\partial X^{(\mathrm{cls})}}\right|_{\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)=(0,0;0)}.\end{aligned}$$

$$
\begin{aligned}
(76):\quad & \mathbb{E}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right] \\
= \quad & \frac{1}{h^{2d}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(\frac{X_j}{h}\right)\left(X_j^{(q)}\right)^r K\left(\frac{X_\ell}{h}\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right) \\
& \qquad \times f_2\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\mathrm{d}X_j^{(\mathrm{ind})}\mathrm{d}X_\ell^{(\mathrm{ind})}\mathrm{d}X^{(\mathrm{cls})} \\
= \quad & h^{r-d_{\mathrm{cls}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(T_j\right)\left(T_j^{(q)}\right)^r K\left(T_\ell\right)\sigma\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right) \\
& \qquad \times f_2\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right)\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})} \\
= \quad & h^{r-d_{\mathrm{cls}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(T_j\right)\left(T_j^{(q)}\right)^r K\left(T_\ell\right)\sigma\left(0,0;0\right) \\
& \qquad \times\left\{f_2\left(0,0;0\right)+hT_j^{(\mathrm{ind})\top}\nabla_1 f_2\left(0,0;0\right)+hT_\ell^{(\mathrm{ind})\top}\nabla_2 f_2\left(0,0;0\right)+hT^{(\mathrm{cls})\top}\nabla_c f_2\left(0,0;0\right)\right\} \\
& \qquad \times\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})}+o\left(h^{r+1-d_{\mathrm{cls}}}\right) \\
= \quad & \begin{cases} h^{-d_{\mathrm{cls}}}R_k^{d_{\mathrm{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)+O\left(h\right) & \text{if } r=0 \\ O\left(h^{2-d_{\mathrm{cls}}}\right) & \text{if } r=1 \end{cases},
\end{aligned}
$$

where the second equality follows from the change of variables $T_j^{(\mathrm{ind})}=X_j^{(\mathrm{ind})}/h$, $T_\ell^{(\mathrm{ind})}=X_\ell^{(\mathrm{ind})}/h$, and $T^{(\mathrm{cls})}=X^{(\mathrm{cls})}/h$ (we define $T_j=\left(T_j^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$, $T_\ell=\left(T_\ell^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$), and the third equality follows from the Taylor expansion.

Similarly,

$$
\begin{aligned}
(77):\quad & \mathbb{E}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)X_j^{(p)}X_\ell^{(q)}\right] \\
= \quad & \frac{1}{h^{2d}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(\frac{X_j}{h}\right)X_j^{(p)}K\left(\frac{X_\ell}{h}\right)X_\ell^{(q)}\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right) \\
& \qquad\qquad \times f_2\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\mathrm{d}X_j^{(\mathrm{ind})}\mathrm{d}X_\ell^{(\mathrm{ind})}\mathrm{d}X^{(\mathrm{cls})} \\
= \quad & h^{2-d_{\mathrm{cls}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(T_j\right)T_j^{(p)}K\left(T_\ell\right)T_\ell^{(q)}\sigma\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right) \\
& \qquad\qquad \times f_2\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right)\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})} \\
= \quad & h^{2-d_{\mathrm{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K\left(T_j\right)T_j^{(p)}K\left(T_\ell\right)T_\ell^{(q)}\mathrm{d}T_j\mathrm{d}T_\ell+o\left(h^{2-d_{\mathrm{cls}}}\right) \\
= \quad & O\left(h^{2-d_{\mathrm{cls}}}\right).
\end{aligned}
$$

Thus, since

$$
\frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}1=\frac{1}{n}\sum_{g=1}^{G}\left(1+2+\cdots+\left(n_g-1\right)\right)=\frac{1}{n}\sum_{g=1}^{G}\frac{n_g(n_g-1)}{2}=\left(\frac{1}{n}\sum_{g=1}^{G}\frac{n_g^2}{2}\right)-\frac{1}{2}
$$

and

$$
\left(\frac{1}{n}\sum_{g=1}^{G}n_g^2-1\right)h^{d_{\mathrm{ind}}}=\lambda+o(1),
$$

we have for $r = 0$,

$$\begin{aligned}\mathbb{E}\left[I_0^{(q)}\right] &= \frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\left\{h^{-d_{\text{cls}}}R_k^{d_{\text{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)+O\left(h^{1-d_{\text{cls}}}\right)\right\}\\ &= \frac{1}{2}\left(\frac{1}{n}\sum_{g=1}^{G}n_g^2-1\right)h^{-d_{\text{cls}}}R_k^{d_{\text{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)+o\left(\left(\max_g n_g\right)h^{-d_{\text{cls}}}\right)\\ &= h^{-d}\left\{\frac{\lambda}{2}R_k^{d_{\text{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)+o\left(1\right)\right\},\end{aligned}$$

and for $r = 1$,

$$\mathbb{E}\left[I_1^{(q)}\right] = \frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}O\left(h^{2-d_{\text{cls}}}\right) = O\left(\left(\max_g n_g\right)h^{2-d_{\text{cls}}}\right) = O\left(h^{-d+2}\right).$$

Also,

$$\mathbb{E}\left[I^{(p,q)}\right] = \frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}O\left(h^{2-d_{\text{cls}}}\right) = O\left(h^{-d+2}\right).$$

For variances, by the mutual independence between clusters,

$$\begin{aligned}&\operatorname{Var}\left[I_r^{(q)}\right]\\ =\;& \operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})},X_{g\ell}^{(\text{ind})};X_g^{(\text{cls})}\right)\left(X_{gj}^{(q)}\right)^r\right]\\ =\;& \frac{1}{n^2}\sum_{g=1}^{G}\operatorname{Var}\left[\sum_{1\le j<\ell\le n_g}K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})},X_{g\ell}^{(\text{ind})};X_g^{(\text{cls})}\right)\left(X_{gj}^{(q)}\right)^r\right].\end{aligned}$$

Here,

$$\begin{aligned}&\operatorname{Var}\left[\sum_{1\le j<\ell\le n_g}K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})},X_{g\ell}^{(\text{ind})};X_g^{(\text{cls})}\right)\left(X_{gj}^{(q)}\right)^r\right]\\ =\;& \sum_{1\le j<\ell\le n_g}\sum_{1\le t<s\le n_g}\operatorname{Cov}\left[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\text{ind})},X_{g\ell}^{(\text{ind})};X_g^{(\text{cls})}\right)\left(X_{gj}^{(q)}\right)^r,\right.\\ &\qquad\left.K_h\left(X_{gt}\right)K_h\left(X_{gs}\right)\sigma\left(X_{gj}^{(\text{ind})},X_{g\ell}^{(\text{ind})};X_g^{(\text{cls})}\right)\left(X_{gt}^{(q)}\right)^r\right],\end{aligned}$$

and there are the following three cases (i) $j = t$ and $\ell = s$, (ii) $j = t, \ell \neq s$, (iii) $j \neq t, \ell = s$, (iv) $j \neq t$ and $\ell \neq s$.

(i) When $j = t$ and $\ell = s$,

$$
\begin{aligned}
& \mathrm{Cov}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r,\right.\\
&\qquad\qquad\left. K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]\\
= & \mathrm{Var}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]\\
\leq & \mathbb{E}\left[K_h^2\left(X_j\right)K_h^2\left(X_\ell\right)\sigma^2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^{2r}\right]\\
= & \frac{1}{h^{4d}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K^2\left(\frac{X_j}{h}\right)K^2\left(\frac{X_\ell}{h}\right)\sigma^2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^{2r}\\
&\qquad \times f_2\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\mathrm{d}X_j^{(\mathrm{ind})}\mathrm{d}X_\ell^{(\mathrm{ind})}\mathrm{d}X^{(\mathrm{cls})}\\
= & \frac{1}{h^{2d-2r+d_{\mathrm{cls}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K^2\left(T_j\right)K^2\left(T_\ell\right)\sigma^2\left(hT_j^{(\mathrm{ind})}, hT_\ell^{(\mathrm{ind})}; hT^{(\mathrm{cls})}\right)\left(T_j^{(q)}\right)^{2r}\\
&\qquad \times f_2\left(hT_j^{(\mathrm{ind})}, hT_\ell^{(\mathrm{ind})}; hT^{(\mathrm{cls})}\right)\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})}\\
= & O\left(h^{2r-2d-d_{\mathrm{cls}}}\right),
\end{aligned}
$$

where the third equality follows from the change of variables $T_j^{(\mathrm{ind})} = X_j^{(\mathrm{ind})}/h$, $T_\ell^{(\mathrm{ind})} = X_\ell^{(\mathrm{ind})}/h$, and $T^{(\mathrm{cls})} = X^{(\mathrm{cls})}/h$ (we define $T_j = \left(T_j^{(\mathrm{ind})\top}, T^{(\mathrm{cls})\top}\right)^\top$, $T_\ell = \left(T_\ell^{(\mathrm{ind})\top}, T^{(\mathrm{cls})\top}\right)^\top$).

(ii) When $j = t, \ell \neq s$,

$$
\begin{aligned}
& \mathrm{Cov}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r,\right.\\
&\qquad\qquad\left. K_h\left(X_j\right)K_h\left(X_s\right)\sigma\left(X_j^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]\\
= & \mathbb{E}\left[K_h^2\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^{2r}K_h\left(X_s\right)\sigma\left(X_j^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\right] \qquad (78)\\
& -\underbrace{\mathbb{E}\left[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]^2}_{=\{(76)\}^2}.
\end{aligned}
$$

We can evaluate an expectation as

$$
\begin{aligned}
(78): & \ \mathbb{E}\left[K_h^2\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^{2r}K_h\left(X_s\right)\sigma\left(X_j^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\right]\\
= & \frac{1}{h^{4d}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K^2\left(\frac{X_j}{h}\right)K\left(\frac{X_\ell}{h}\right)\sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^{2r}\\
&\qquad \times K_h\left(\frac{X_s}{h}\right)\sigma\left(X_j^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\\
&\qquad \times f_3\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\mathrm{d}X_j^{(\mathrm{ind})}\mathrm{d}X_\ell^{(\mathrm{ind})}\mathrm{d}X_s^{(\mathrm{ind})}\mathrm{d}X^{(\mathrm{cls})}\\
= & \frac{1}{h^{d-2r+2d_{\mathrm{cls}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}} K^2\left(T_j\right)K\left(T_\ell\right)\sigma\left(hT_j^{(\mathrm{ind})}, hT_\ell^{(\mathrm{ind})}; hT^{(\mathrm{cls})}\right)\left(T_j^{(q)}\right)^{2r}\\
&\qquad \times K\left(T_s\right)\sigma\left(hT_j^{(\mathrm{ind})}, hT_s^{(\mathrm{ind})}; hT^{(\mathrm{cls})}\right)\\
&\qquad \times f_3\left(hT_j^{(\mathrm{ind})}, hT_\ell^{(\mathrm{ind})}, hT_s^{(\mathrm{ind})}; hT^{(\mathrm{cls})}\right)\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T_s^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})}\\
= & O\left(h^{2r-d-2d_{\mathrm{cls}}}\right),
\end{aligned}
$$

where the second equality follows from the change of variables $T_j^{(\mathrm{ind})} = X_j^{(\mathrm{ind})}/h$, $T_\ell^{(\mathrm{ind})} = X_\ell^{(\mathrm{ind})}/h$, $T_s^{(\mathrm{ind})} = X_s^{(\mathrm{ind})}/h$ and $T^{(\mathrm{cls})} = X^{(\mathrm{cls})}/h$ (we define $T_j = \left(T_j^{(\mathrm{ind})\top}, T^{(\mathrm{cls})\top}\right)^\top$, $T_\ell = \left(T_\ell^{(\mathrm{ind})\top}, T^{(\mathrm{cls})\top}\right)^\top$,

$T_s = \left(T_s^{(\mathrm{ind})\top}, T^{(\mathrm{cls})\top}\right)^{\top}$). Thus,

$$
\begin{aligned}
& \mathrm{Cov}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r,\right. \\
& \qquad\qquad \left. K_h\left(X_j\right) K_h\left(X_s\right) \sigma\left(X_j^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right] \\
= \;& O\left(h^{2r-d-2d_{\mathrm{cls}}}\right).
\end{aligned}
$$

(iii) When $j \neq t, \ell = s$,

$$
\begin{aligned}
& \mathrm{Cov}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r,\right. \\
& \qquad\qquad \left. K_h\left(X_t\right) K_h\left(X_\ell\right) \sigma\left(X_t^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\right] \\
& \mathbb{E}\left[K_h\left(X_j\right) K_h^2\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right. \\
& \qquad \left. \times K_h\left(X_t\right) \sigma\left(X_t^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\right] \\
& - \underbrace{\mathbb{E}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]^2}_{=\{(76)\}^2} \\
= \;& O\left(h^{2r-d-2d_{\mathrm{cls}}}\right)
\end{aligned} \tag{79}
$$

by a similar derivation to the case (ii).

(iv) When $j \neq t, \ell \neq s$,

$$
\begin{aligned}
& \mathrm{Cov}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r,\right. \\
& \qquad\qquad \left. K_h\left(X_t\right) K_h\left(X_s\right) \sigma\left(X_t^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\right] \\
= \;& \mathbb{E}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right. \\
& \qquad \left. \times K_h\left(X_t\right) K_h\left(X_s\right) \sigma\left(X_t^{(\mathrm{ind})}, X_s^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\right] \\
& - \underbrace{\mathbb{E}\left[K_h\left(X_j\right) K_h\left(X_\ell\right) \sigma\left(X_j^{(\mathrm{ind})}, X_\ell^{(\mathrm{ind})}; X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r\right]^2}_{=\{(76)\}^2}.
\end{aligned} \tag{80}
$$

We can evaluate an expectation as

$$
\begin{aligned}
(80):\quad & \mathbb{E}\Big[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r \\
& \qquad \times K_h\left(X_t\right)K_h\left(X_s\right)\sigma\left(X_t^{(\mathrm{ind})},X_s^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\Big] \\
= \quad & \frac{1}{h^{4d}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}} \\
& \qquad \times K\left(\frac{X_j}{h}\right)K\left(\frac{X_\ell}{h}\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r \\
& \qquad \times K\left(\frac{X_t}{h}\right)K\left(\frac{X_s}{h}\right)\sigma\left(X_t^{(\mathrm{ind})},X_s^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r \\
& \qquad \times f_4\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})},X_t^{(\mathrm{ind})},X_s^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\mathrm{d}X_j^{(\mathrm{ind})}\mathrm{d}X_\ell^{(\mathrm{ind})}\mathrm{d}X_t^{(\mathrm{ind})}\mathrm{d}X_s^{(\mathrm{ind})}\mathrm{d}X^{(\mathrm{cls})} \\
= \quad & h^{2r-3d_{\mathrm{cls}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}}\int_{\mathbb{R}^{d_{\mathrm{cls}}}}\int_{\mathbb{R}^{d_{\mathrm{ind}}}} \\
& \qquad \times K\left(T_j\right)K\left(T_\ell\right)\sigma\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right)\left(T_j^{(q)}\right)^r \\
& \qquad \times K\left(T_t\right)K\left(T_s\right)\sigma\left(hT_t^{(\mathrm{ind})},hT_s^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right)\left(T_t^{(q)}\right)^r \\
& \qquad \times f_4\left(hT_j^{(\mathrm{ind})},hT_\ell^{(\mathrm{ind})},hT_t^{(\mathrm{ind})},hT_s^{(\mathrm{ind})};hT^{(\mathrm{cls})}\right)\mathrm{d}T_j^{(\mathrm{ind})}\mathrm{d}T_\ell^{(\mathrm{ind})}\mathrm{d}T_t^{(\mathrm{ind})}\mathrm{d}T_s^{(\mathrm{ind})}\mathrm{d}T^{(\mathrm{cls})} \\
= \quad & O\left(h^{2r-3d_{\mathrm{cls}}}\right),
\end{aligned}
$$

where the second equality follows from the change of variables $T_j^{(\mathrm{ind})}=X_j^{(\mathrm{ind})}/h$, $T_\ell^{(\mathrm{ind})}=X_\ell^{(\mathrm{ind})}/h$, $T_t^{(\mathrm{ind})}=X_t^{(\mathrm{ind})}/h$, $T_s^{(\mathrm{ind})}=X_s^{(\mathrm{ind})}/h$ and $T^{(\mathrm{cls})}=X^{(\mathrm{cls})}/h$ (we define $T_j=\left(T_j^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$, $T_\ell=\left(T_\ell^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$, $T_t=\left(T_t^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$, $T_s=\left(T_s^{(\mathrm{ind})\top},T^{(\mathrm{cls})\top}\right)^\top$). Thus,

$$
\begin{aligned}
& \operatorname{Cov}\Big[K_h\left(X_j\right)K_h\left(X_\ell\right)\sigma\left(X_j^{(\mathrm{ind})},X_\ell^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_j^{(q)}\right)^r, \\
& \qquad\qquad K_h\left(X_t\right)K_h\left(X_s\right)\sigma\left(X_t^{(\mathrm{ind})},X_s^{(\mathrm{ind})};X^{(\mathrm{cls})}\right)\left(X_t^{(q)}\right)^r\Big] \\
= \quad & O\left(h^{2r-3d_{\mathrm{cls}}}\right).
\end{aligned}
$$

Thus, by counting cases (i)-(iv),

$$
\begin{aligned}
& \operatorname{Var}\left[I_r^{(q)}\right] \\
\leq \quad & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}\sum_{1\leq t<s\leq n_g}\operatorname{Cov}\Big[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{gj}^{(q)}\right)^r, \\
& \qquad\qquad K_h\left(X_{gt}\right)K_h\left(X_{gs}\right)\sigma\left(X_{gt}^{(\mathrm{ind})},X_{gs}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{gt}^{(q)}\right)^r\Big] \\
\leq \quad & \frac{1}{n^2}\sum_{g=1}^{G}\left\{n_g^2O\left(h^{2r-2d-d_{\mathrm{cls}}}\right)+n_g^3O\left(h^{2r-d-2d_{\mathrm{cls}}}\right)+n_g^3O\left(h^{2r-d-2d_{\mathrm{cls}}}\right)+n_g^4O\left(h^{2r-3d_{\mathrm{cls}}}\right)\right\} \\
\leq \quad & \frac{\left(\max_g n_g\right)h^{-d_{\mathrm{cls}}}}{n}\left\{O\left(h^{2r-2d}\right)+\left(\max_g n_g\right)h^{-d_{\mathrm{cls}}}O\left(h^{2r-d}\right)+\left(\max_g n_g\right)^2h^{-2d_{\mathrm{cls}}}O\left(h^{2r}\right)\right\} \\
= \quad & o(1)\left\{O\left(h^{2r-2d}\right)+\underbrace{\left(\max_g n_g\right)h^{d_{\mathrm{ind}}}}_{O(1)}O\left(h^{2r-2d}\right)+\underbrace{\left(\max_g n_g\right)^2h^{2d_{\mathrm{ind}}}}_{=O(1)}O\left(h^{2r-2d}\right)\right\}=o\left(h^{2r-2d}\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&\operatorname{Var}\left[I^{(p,q)}\right]\\
\leq\ & \frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g}\sum_{1\leq t<s\leq n_g}\operatorname{Cov}\left[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\sigma\left(X_{gj}^{(\mathrm{ind})},X_{g\ell}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{gj}^{(p)}\right)\left(X_{gj}^{(q)}\right),\right.\\
&\qquad\left.K_h\left(X_{gt}\right)K_h\left(X_{gs}\right)\sigma\left(X_{gt}^{(\mathrm{ind})},X_{gs}^{(\mathrm{ind})};X_g^{(\mathrm{cls})}\right)\left(X_{gt}^{(p)}\right)\left(X_{gs}^{(q)}\right)\right]\\
\leq\ & \frac{1}{n^2}\sum_{g=1}^{G}\left\{n_g^2 O\left(h^{4-2d-d_{\mathrm{cls}}}\right)+n_g^3 O\left(h^{4-d-2d_{\mathrm{cls}}}\right)+n_g^3 O\left(h^{4-d-2d_{\mathrm{cls}}}\right)+n_g^4 O\left(h^{4-3d_{\mathrm{cls}}}\right)\right\}\leq o\left(h^{4-2d}\right).
\end{aligned}
$$

Therefore, by Markov's inequality, $I_0^{(q)} = h^{-d}\left\{\frac{\lambda}{2}R_k^{d_{\mathrm{cls}}}\sigma\left(0,0;0\right)f_2\left(0,0;0\right)+o_p\left(1\right)\right\}$, $I_1^{(q)} = o_p\left(h^{-d+1}\right)$, and $I^{(p,q)} = O_p\left(h^{-d+2}\right)$. We conclude by element-wise comparisons. □

B.5. **Proof for Lemma 6.**

Proof. Define $\mathcal{E}_r^{(q)} = \frac{1}{n}\sum_{g=1}^{G}\sum_{1\leq j<\ell\leq n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}$ for $r=0,1$. We have

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{E}_r^{(q)}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g}K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}\right]\\
&= \mathbb{E}\left[K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r\mathbb{E}\left[e_{gj}\mid X_{gj}\right]\right]\\
&= 0
\end{aligned}
$$

by the law of iterated expectations.

For variances,

$$
\begin{aligned}
&\operatorname{Var}\left[\mathcal{E}_r^{(q)}\right]\\
= &\operatorname{Var}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}\right]\\
= &\mathbb{E}\left[\left\{\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}\right\}^2\right] - \underbrace{\mathbb{E}\left[\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}\right]^2}_{=0}\\
= &\frac{1}{n^2}\sum_{g=1}^{G}\mathbb{E}\left[\mathbb{E}\left[\left\{\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^r e_{gj}\right\}^2 \mid X_g\right]\right]\\
= &\frac{1}{n^2}\sum_{g=1}^{G}\sum_{j=1}^{n_g}\mathbb{E}\left[K_h^2\left(X_{gj}\right)\sigma^2\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^{2r}\right]\\
&+2\frac{1}{n^2}\sum_{g=1}^{G}\sum_{1\le j<\ell\le n_g}\mathbb{E}\left[K_h\left(X_{gj}\right)K_h\left(X_{g\ell}\right)\left(X_{gj}^{(q)}\right)^r\left(X_{g\ell}^{(q)}\right)^r\sigma\left(X_{gj}^{(\mathrm{ind})}, X_{g\ell}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)\right]\\
= &\begin{cases}\frac{1}{n}\mathbb{E}\left[H_0^{(q)} + 2I_0^{(q)}\right] & \text{if } r=0\\ \frac{1}{n}\mathbb{E}\left[H_2^{(q)} + 2I^{(q,q)}\right] & \text{if } r=1\end{cases}\\
= &\begin{cases}O\left(\frac{1}{nh^d}\right) & \text{if } r=0\\ O\left(\frac{1}{nh^{d-2}}\right) & \text{if } r=1\end{cases},
\end{aligned}
$$

where the third equality follows from the mutual independence between clusters.

Therefore, by Markov's inequality, $\mathcal{E}_0^{(q)} = O_p\left(\sqrt{\frac{1}{nh^d}}\right)$ and $\mathcal{E}_1^{(q)} = O_p\left(\sqrt{\frac{1}{nh^{d-2}}}\right)$ We conclude by element-wise comparisons. □

B.6. **Proof for Lemma 7.**

Proof. By the proof of Lemma 2,

$$
\begin{aligned}
&\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left\{X_{gj}^{\top}\nabla^2 m(0) X_{gj}\right\}\\
= &\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\sum_{p=1}^{d}\sum_{q=1}^{d}\partial_{pq}m(0)X_{gj}^{(p)}X_{gj}^{(q)}\\
= &\sum_{q=1}^{d}\partial_{qq}m(0)F_2^{(q)} + 2\sum_{1\le p<q\le d}\partial_{pq}m(0)F^{(p,q)}\\
= &\sum_{q=1}^{d}\partial_{qq}m(0)\left\{h^2 f(0)\kappa_2 + o_p\left(h^2\right)\right\} + 2\sum_{1\le p<q\le d}\partial_{pq}m(0)o_p\left(h^2\right)\\
= &h^2\kappa_2\sum_{q=1}^{d}\partial_{qq}m(0)f(0) + o_p\left(h^2\right).
\end{aligned}
$$

Next, we will evaluate

$$\begin{aligned}
&\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right) X_{gj}\left\{X_{gj}^{\top}\nabla^2 m(0) X_{gj}\right\}\\
= \quad &\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right) X_{gj}\left\{\sum_{q=1}^{d}\partial_{qq}m(0)\left(X_{gj}^{(q)}\right)^2 + 2\sum_{1\le p<q\le d}\partial_{pq}m(0)X_{gj}^{(p)}X_{gj}^{(q)}\right\}
\end{aligned}$$

The compact support of the kernel function implies

$$\begin{aligned}
\partial_{qq}m(0)\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)\left(X_{gj}^{(q)}\right)^3 &= O_p\left(h^3\right),\\
\partial_{qq}m(0)\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)X_{gj}^{(p)}\left(X_{gj}^{(q)}\right)^2 &= O_p\left(h^3\right),\\
\partial_{pq}m(0)\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)X_{gj}^{(p)}\left(X_{gj}^{(q)}\right)^2 &= O_p\left(h^3\right),\\
\partial_{pq}m(0)\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right)X_{gj}^{(p)}X_{gj}^{(q)}X_{gj}^{(q')} &= O_p\left(h^3\right).
\end{aligned}$$

Thus,

$$\frac{1}{n}\sum_{g=1}^{G}\sum_{j=1}^{n_g} K_h\left(X_{gj}\right) X_{gj}\left\{X_{gj}^{\top}\nabla^2 m(0) X_{gj}\right\} = O_p\left(h^3\right)\mathbf{1}_d.$$

□

Appendix C. Technical discussion

As we mentioned in Remark 1, we can relax the identical distribution assumptions for joint densities if we strengthen the continuity assumption for them. For example, we can together replace Assumption 1 (iii) and Assumption 3 (ii) by the following assumptions to show the theorems on Section 4.

- **Assumption** 1 (iii'): *X_{gj} are identically distributed across all g and j with common marginal density $f(x)$. For any cluster g with $n_g \ge 2$, $\left(X_{gj_1}^{(\mathrm{ind})}, X_{gj_2}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)$ are identically distributed across all g, j_1, and j_2 with common joint density*

  $$f_2\left(x_1^{(\mathrm{ind})}, x_2^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right).$$

  For any $\underline{n}_g \in \{3,4\}$ and for any cluster g with $n_g \ge \underline{n}_g$, $\left(X_{gj_1}^{(\mathrm{ind})}, \cdots, X_{gj_{\underline{n}_g}}^{(\mathrm{ind})}; X_g^{(\mathrm{cls})}\right)$ with $j_1 < j_2 < \cdots < j_{\underline{n}_g}$ has the joint density

  $$f_{\left(j_1,j_2,\ldots,j_{\underline{n}_g};g\right)}\left(x_1^{(\mathrm{ind})}, x_2^{(\mathrm{ind})}, \cdots, x_{\underline{n}_g}^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right).$$

- **Assumption** 3 (ii'): *There exists some neighborhood $\mathcal{N}$ of $x = \left(x^{(\mathrm{ind})\top}, x^{(\mathrm{cls})\top}\right)^{\top}$ such that $m(x)$ and $f(x)$ are twice continuously differentiable, $f_2\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ is continuously differentiable, and $\sigma^2(x)$, and $\sigma\left(x^{(\mathrm{ind})}, x^{(\mathrm{ind})}; x^{(\mathrm{cls})}\right)$ are continuous. Moreover,* *for any* $\underline{n}_g \in \{3, 4\}$*,*

$$\left\{ \left\{ f_{\left(j_1, j_2, \ldots, j_{\underline{n}_g}; g\right)} \left(x_1^{(\text{ind})}, x_2^{(\text{ind})}, \cdots, x_{\underline{n}_g}^{(\text{ind})}; x^{(\text{cls})} \right) \right\}_{1 \leq j_1 < j_2 < \cdots < j_{\underline{n}_g} \leq n_g} \right\}_{g : n_g \geq \underline{n}_g}$$

is equicontinuous in the neighborhood $\mathcal{N}$*.*

Appendix D. **Additional simulations**

D.1. **Simulation results for the Nadaraya-Watson estimator.** In this subsection, we will provide simulation results of the Nadaraya-Watson estimator for bandwidth selection and inference.

D.1.1. *Bandwidth selection.* The data-generating processes and the calculations are the same as in Section 9. As we did for local linear estimators in Section 9, we will compare four methods of bandwidth choice. The performance is evaluated by

$$\text{ASE}(h) = \frac{1}{n_{\text{grid}}} \sum_{k=1}^{n_{\text{grid}}} \left\{ \widehat{m}_{\text{nw}} \left(u_k, h \right) - m \left(u_k \right) \right\}^2 ,$$

where $\widehat{m}_{\text{nw}} \left(u_k, h \right)$ is the Nadaraya-Watson estimator with the bandwidth h.

Tables 6 and 7 show means of ASEs for the Nadaraya-Watson estimator and means of selected bandwidths (in curly brackets) across each simulation draw for Setup 1 and 2, respectively. Figures 5 and 6 plot values of the bandwidth h in the x-axis and means of the function ASE(h) in the y-axis, which are calculated from simulation draws for Setups 1 and 2, respectively. We found almost the same implications as in Section 9.1, and the detailed explanations are omitted.

Table 6. Mean of ASE and mean of selected bandwidth (m_{nw}, Setup 1)

	$\max n_g = 20$				$\max n_g = 100$			
	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$
(ρ_X, ρ_e)=(0.2,0.2)	0.0053	0.0053	0.0042	0.0042	0.0053	0.0053	0.0042	0.0042
	{0.0297}	{0.0302}	{0.0471}	{0.0471}	{0.0292}	{0.0297}	{0.0467}	{0.0468}
(ρ_X, ρ_e)=(0.2,0.5)	0.0062	0.0061	0.0050	0.0050	0.0063	0.0062	0.0051	0.0051
	{0.0297}	{0.0302}	{0.0471}	{0.0472}	{0.0292}	{0.0297}	{0.0467}	{0.0468}
(ρ_X, ρ_e)=(0.5,0.2)	0.0055	0.0054	0.0043	0.0043	0.0055	0.0055	0.0043	0.0043
	{0.0292}	{0.0300}	{0.0473}	{0.0476}	{0.0288}	{0.0295}	{0.0472}	{0.0474}
(ρ_X, ρ_e)=(0.5,0.5)	0.0066	0.0065	0.0054	0.0054	0.0068	0.0067	0.0056	0.0056
	{0.0292}	{0.0300}	{0.0475}	{0.0477}	{0.0288}	{0.0295}	{0.0471}	{0.0473}

Note: Means of selected bandwidths are shown in curly brackets.

TABLE 7. Mean of ASE and mean of selected bandwidth (m_{nw}, Setup 2)

	$\max n_g = 20$				$\max n_g = 100$			
	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$	h_{ROT}	$h_{\text{CR-ROT}}$	h_{CV}	$h_{\text{CR-CV}}$
(ρ_X, ρ_e)=(0.2,0.2)	0.0086	0.0072	0.0028	0.0028	0.0081	0.0069	0.0028	0.0028
	{0.0890}	{0.0865}	{0.0467}	{0.0468}	{0.0876}	{0.0853}	{0.0463}	{0.0464}
(ρ_X, ρ_e)=(0.2,0.5)	0.0094	0.0079	0.0033	0.0033	0.0089	0.0076	0.0034	0.0034
	{0.0893}	{0.0868}	{0.0467}	{0.0468}	{0.0878}	{0.0855}	{0.0465}	{0.0467}
(ρ_X, ρ_e)=(0.5,0.2)	0.0088	0.0077	0.0029	0.0029	0.0087	0.0075	0.0029	0.0029
	{0.0896}	{0.0877}	{0.0474}	{0.0475}	{0.0889}	{0.0869}	{0.0469}	{0.0471}
(ρ_X, ρ_e)=(0.5,0.5)	0.0094	0.0083	0.0036	0.0036	0.0094	0.0082	0.0037	0.0037
	{0.0892}	{0.0874}	{0.0471}	{0.0474}	{0.0886}	{0.0866}	{0.0467}	{0.0470}

Note: Means of selected bandwidths are shown in curly brackets.

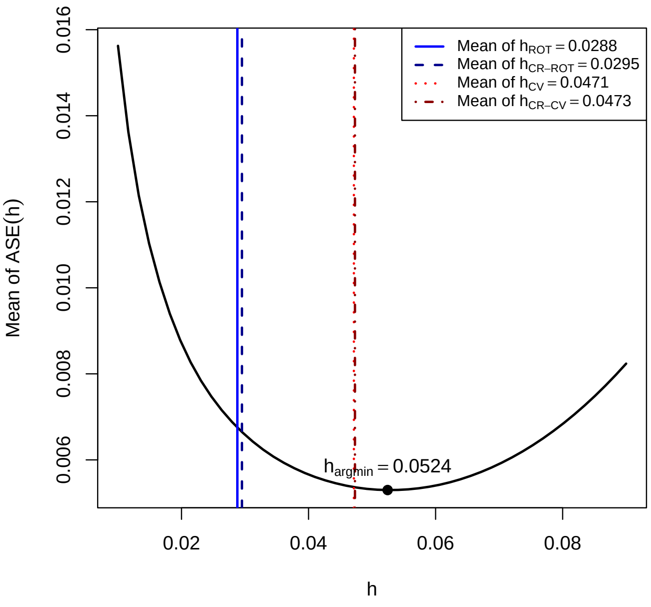

FIGURE 5. Mean of ASE(h) for m_{nw} in Setup 1 with $\max_{g \leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$

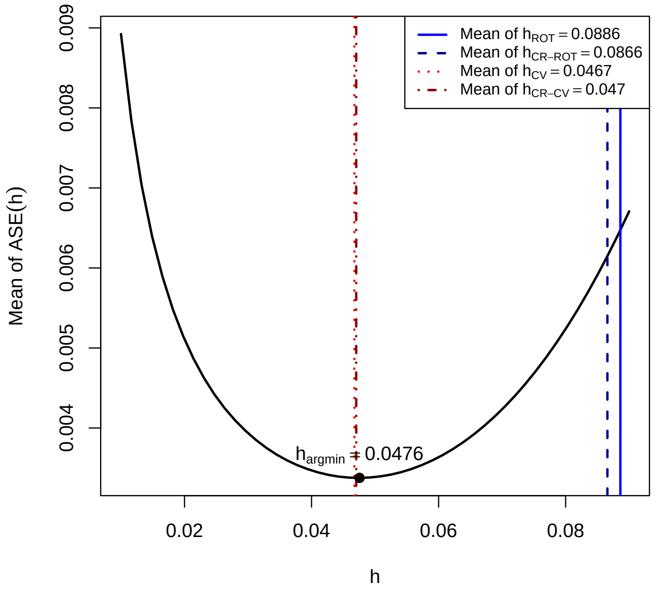

FIGURE 6. Mean of ASE(h) for m_{nw} in Setup 2 with $\max_{g \leq G} n_g = 100$ and $\rho_X = \rho_e = 0.5$

D.1.2. *Inference.* The data-generating processes and the calculations are the same as in Section 9. Tables 8-10 show the coverage ratio for the Nadaraya-Watson estimator and means of the length of confidence intervals (in curly brackets) across each simulation draw for Setup 1, Setup 2 with $x = 0.8$, and Setup 2 with $x = 0.4$, respectively. We found almost the same implications as in Section 9.2, and the detailed explanations are omitted.

TABLE 8. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 1)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_{λ}	CI	CI_{CR}	CI_{λ}
(ρ_X, ρ_e)=(0.2,0.2)	0.925	0.931	0.954	0.915	0.920	0.952
	{0.192}	{0.195}	{0.217}	{0.189}	{0.192}	{0.217}
(ρ_X, ρ_e)=(0.2,0.5)	0.879	0.886	0.960	0.861	0.869	0.951
	{0.192}	{0.195}	{0.246}	{0.188}	{0.192}	{0.250}
(ρ_X, ρ_e)=(0.5,0.2)	0.920	0.925	0.956	0.906	0.908	0.954
	{0.191}	{0.194}	{0.227}	{0.188}	{0.191}	{0.228}
(ρ_X, ρ_e)=(0.5,0.5)	0.857	0.867	0.964	0.833	0.844	0.957
	{0.191}	{0.195}	{0.261}	{0.188}	{0.191}	{0.267}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 9. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.8$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.893	0.897	0.931	0.882	0.886	0.923
	{0.167}	{0.170}	{0.187}	{0.165}	{0.168}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.844	0.850	0.918	0.831	0.836	0.924
	{0.167}	{0.171}	{0.209}	{0.164}	{0.168}	{0.211}
(ρ_X, ρ_e)=(0.5,0.2)	0.903	0.909	0.936	0.878	0.884	0.927
	{0.166}	{0.170}	{0.191}	{0.164}	{0.167}	{0.192}
(ρ_X, ρ_e)=(0.5,0.5)	0.826	0.837	0.932	0.806	0.816	0.924
	{0.166}	{0.170}	{0.218}	{0.164}	{0.167}	{0.223}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 10. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.4$)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.985	0.986	0.995	0.983	0.985	0.995
	{0.136}	{0.138}	{0.157}	{0.134}	{0.135}	{0.157}
(ρ_X, ρ_e)=(0.2,0.5)	0.969	0.971	0.998	0.965	0.968	0.998
	{0.136}	{0.138}	{0.181}	{0.133}	{0.135}	{0.184}
(ρ_X, ρ_e)=(0.5,0.2)	0.982	0.983	0.996	0.980	0.983	0.998
	{0.135}	{0.137}	{0.160}	{0.133}	{0.134}	{0.161}
(ρ_X, ρ_e)=(0.5,0.5)	0.962	0.964	0.996	0.961	0.963	0.997
	{0.135}	{0.137}	{0.188}	{0.132}	{0.134}	{0.192}

Note: Lengths of confidence intervals are shown in curly brackets.

D.2. **Inference without bias corrections.** This subsection presents simulation results for inference methods that do not incorporate bias corrections. We include results for both Nadaraya-Watson and local linear estimators. The data-generating processes for these simulations are the same as in Section 9. The calculations of confidence intervals are basically the same as in Section 9, albeit without correcting the bias. Tables 11-16 show the coverage ratio and means of the length of confidence intervals (in curly brackets) across each simulation draw. These results are the feasible version of our previous inference results, assuming undersmoothing to ignore the bias. Notably, among the evaluated methods, our CI_λ confidence intervals exhibit superior performance.

Table 11. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 1, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.917	0.921	0.952	0.908	0.914	0.951
	{0.190}	{0.193}	{0.215}	{0.187}	{0.189}	{0.215}
(ρ_X, ρ_e)=(0.2,0.5)	0.876	0.886	0.958	0.860	0.869	0.951
	{0.189}	{0.192}	{0.244}	{0.186}	{0.189}	{0.248}
(ρ_X, ρ_e)=(0.5,0.2)	0.914	0.919	0.955	0.905	0.910	0.949
	{0.189}	{0.192}	{0.225}	{0.186}	{0.189}	{0.226}
(ρ_X, ρ_e)=(0.5,0.5)	0.858	0.864	0.961	0.835	0.842	0.957
	{0.189}	{0.192}	{0.260}	{0.185}	{0.189}	{0.265}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 12. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 1, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.918	0.925	0.953	0.907	0.914	0.948
	{0.192}	{0.195}	{0.217}	{0.189}	{0.192}	{0.217}
(ρ_X, ρ_e)=(0.2,0.5)	0.880	0.888	0.956	0.861	0.868	0.949
	{0.192}	{0.195}	{0.246}	{0.188}	{0.192}	{0.250}
(ρ_X, ρ_e)=(0.5,0.2)	0.916	0.921	0.956	0.906	0.910	0.951
	{0.191}	{0.194}	{0.227}	{0.188}	{0.191}	{0.228}
(ρ_X, ρ_e)=(0.5,0.5)	0.859	0.865	0.960	0.837	0.845	0.955
	{0.191}	{0.195}	{0.261}	{0.188}	{0.191}	{0.267}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 13. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 2, $x = 0.8$, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.776	0.788	0.827	0.780	0.787	0.843
	{0.168}	{0.171}	{0.187}	{0.166}	{0.169}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.735	0.745	0.846	0.737	0.746	0.856
	{0.168}	{0.171}	{0.209}	{0.165}	{0.168}	{0.212}
(ρ_X, ρ_e)=(0.5,0.2)	0.753	0.764	0.819	0.755	0.761	0.832
	{0.167}	{0.171}	{0.192}	{0.165}	{0.168}	{0.193}
(ρ_X, ρ_e)=(0.5,0.5)	0.706	0.721	0.851	0.715	0.725	0.855
	{0.167}	{0.171}	{0.219}	{0.164}	{0.168}	{0.223}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 14. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.8$, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.772	0.783	0.821	0.782	0.791	0.840
	{0.167}	{0.170}	{0.187}	{0.165}	{0.168}	{0.187}
(ρ_X, ρ_e)=(0.2,0.5)	0.737	0.745	0.842	0.734	0.743	0.852
	{0.167}	{0.171}	{0.209}	{0.164}	{0.168}	{0.211}
(ρ_X, ρ_e)=(0.5,0.2)	0.748	0.756	0.819	0.752	0.758	0.829
	{0.166}	{0.170}	{0.191}	{0.164}	{0.167}	{0.192}
(ρ_X, ρ_e)=(0.5,0.5)	0.701	0.714	0.846	0.707	0.718	0.853
	{0.166}	{0.170}	{0.218}	{0.164}	{0.167}	{0.223}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 15. Coverage and mean of length of 95% CI for each standard error (m_{LL}, Setup 2, $x = 0.4$, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.954	0.957	0.980	0.945	0.949	0.982
	{0.137}	{0.138}	{0.157}	{0.134}	{0.136}	{0.158}
(ρ_X, ρ_e)=(0.2,0.5)	0.935	0.939	0.988	0.922	0.926	0.985
	{0.136}	{0.139}	{0.182}	{0.134}	{0.136}	{0.184}
(ρ_X, ρ_e)=(0.5,0.2)	0.945	0.948	0.981	0.943	0.945	0.984
	{0.136}	{0.138}	{0.160}	{0.133}	{0.135}	{0.161}
(ρ_X, ρ_e)=(0.5,0.5)	0.924	0.929	0.989	0.909	0.915	0.987
	{0.136}	{0.138}	{0.188}	{0.133}	{0.135}	{0.192}

Note: Lengths of confidence intervals are shown in curly brackets.

Table 16. Coverage and average length of 95% CI for each standard error (m_{nw}, Setup 2, $x = 0.4$, with bias)

	$\max n_g = 20$			$\max n_g = 100$		
	CI	CI_{CR}	CI_λ	CI	CI_{CR}	CI_λ
(ρ_X, ρ_e)=(0.2,0.2)	0.951	0.953	0.978	0.941	0.944	0.978
	{0.136}	{0.138}	{0.157}	{0.134}	{0.135}	{0.157}
(ρ_X, ρ_e)=(0.2,0.5)	0.928	0.934	0.982	0.915	0.918	0.986
	{0.136}	{0.138}	{0.181}	{0.133}	{0.135}	{0.184}
(ρ_X, ρ_e)=(0.5,0.2)	0.937	0.942	0.985	0.935	0.938	0.980
	{0.135}	{0.137}	{0.160}	{0.133}	{0.134}	{0.161}
(ρ_X, ρ_e)=(0.5,0.5)	0.919	0.924	0.988	0.910	0.915	0.987
	{0.135}	{0.137}	{0.188}	{0.132}	{0.134}	{0.192}

Note: Lengths of confidence intervals are shown in curly brackets.

References

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M. (2023) When should you adjust standard errors for clustering?, *The Quarterly Journal of Economics*, **138**, 1–35. [Cited on page 5]

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A. and Tobias, J. (2012) Targeting the poor: evidence from a field experiment in indonesia, *American Economic Review*, **102**, 1206–1240. [Cited on pages 3, 25, 26, and 27]

Armstrong, T. B. and Kolesár, M. (2018) Optimal inference in a class of regression models, *Econometrica*, **86**, 655–683. [Cited on page 19]

Bartalotti, O. and Brummet, Q. (2017) Regression discontinuity designs with clustered data, in *Regression Discontinuity Designs*, Emerald Publishing Limited. [Cited on pages 2 and 20]

Bhattacharya, D. (2005) Asymptotic inference from multi-stage samples, *Journal of Econometrics*, **126**, 145–171. [Cited on pages 2 and 9]

Bugni, F., Canay, I., Shaikh, A. and Tabord-Meehan, M. (2022) Inference for cluster randomized experiments with non-ignorable cluster sizes, *arXiv preprint arXiv:2204.08356*. [Cited on pages 2 and 5]

Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019) nprobust: Nonparametric kernel-based estimation and robust bias-corrected inference, *arXiv preprint arXiv:1906.00198*. [Cited on page 15]

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014) Robust nonparametric confidence intervals for regression-discontinuity designs, *Econometrica*, **82**, 2295–2326. [Cited on page 19]

Cameron, A. C., Gelbach, J. B. and Miller, D. L. (2008) Bootstrap-based improvements for inference with clustered errors, *The Review of Economics and Statistics*, **90**, 414–427. [Cited on page 20]

Cameron, A. C. and Miller, D. L. (2015) A practitioner's guide to cluster-robust inference, *Journal of Human Resources*, **50**, 317–372. [Cited on page 2]

Cattaneo, M. D., Crump, R. K. and Jansson, M. (2013) Generalized jackknife estimators of weighted average derivatives, *Journal of the American Statistical Association*, **108**, 1243–1256. [Cited on page 37]

Djogbenou, A. A., MacKinnon, J. G. and Nielsen, M. Ø. (2019) Asymptotic theory and wild bootstrap inference with clustered errors, *Journal of Econometrics*, **212**, 393–412. [Cited on page 2]

Fan, J. (1992) Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004. [Cited on page 3]

Fan, J. and Gijbels, I. (1992) Variable bandwidth and local linear regression smoothers, *The Annals of Statistics*, pp. 2008–2036. [Cited on pages 10 and 20]

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability*, vol. 66, CRC Press. [Cited on pages 13, 14, and 20]

Hansen, B. E. (2008) Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory*, **24**, 726–748. [Cited on pages 3, 37, and 46]

Hansen, B. E. (2022a) *Econometrics*, Princeton University Press. [Cited on pages 14 and 15]

Hansen, B. E. (2022b) Jackknife standard errors for clustered regression. [Cited on page 19]

Hansen, B. E. and Lee, S. (2019) Asymptotic theory for clustered samples, *Journal of Econometrics*, **210**, 268–290. [Cited on pages 2, 6, 9, and 32]

Hansen, C. B. (2007) Asymptotic properties of a robust variance matrix estimator for panel data when t is large, *Journal of Econometrics*, **141**, 597–620. [Cited on page 2]

Imbens, G. and Kalyanaraman, K. (2012) Optimal bandwidth choice for the regression discontinuity estimator, *The Review of Economic Studies*, **79**, 933–959. [Cited on page 23]

Kai, B., Li, R. and Zou, H. (2010) Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **72**, 49–69. [Cited on page 20]

Kristensen, D. (2009) Uniform convergence rates of kernel estimators with heterogeneous dependent data, *Econometric Theory*, **25**, 1433–1445. [Cited on page 3]

Lee, J. and Robinson, P. M. (2016) Series estimation under cross-sectional dependence, *Journal of Econometrics*, **190**, 1–17. [Cited on page 3]

Li, Q. and Racine, J. S. (2007) *Nonparametric econometrics: theory and practice*, Princeton University Press. [Cited on page 23]

Lin, X. and Carroll, R. J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error, *Journal of the American Statistical Association*, **95**, 520–534. [Cited on page 2]

MacKinnon, J. G., Nielsen, M. Ø. and Webb, M. D. (2022) Cluster-robust inference: A guide to empirical practice, *Journal of Econometrics*. [Cited on page 2]

Menzel, K. (2024) Transfer estimates for causal effects across heterogeneous sites, *arXiv preprint arXiv:2305.01435*. [Cited on page 2]

Robinson, P. M. (1983) Nonparametric estimators for time series, *Journal of Time Series Analysis*, **4**, 185–207. [Cited on page 3]

Robinson, P. M. (2011) Asymptotic theory for nonparametric regression with spatial data, *Journal of Econometrics*, **165**, 5–19. [Cited on page 3]

Ruppert, D. and Wand, M. P. (1994) Multivariate locally weighted least squares regression, *The Annals of Statistics*, pp. 1346–1370. [Cited on pages 3 and 6]

Stone, C. J. (1982) Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics*, pp. 1040–1053. [Cited on pages 3 and 12]

Vogt, M. (2012) Nonparametric regression for locally stationary time series, *The Annals of Statistics*, **40**, 2601–2633. [Cited on pages 3 and 37]

Vogt, M. and Linton, O. (2020) Multiscale clustering of nonparametric regression curves, *Journal of Econometrics*, **216**, 305–325. [Cited on page 3]

Wang, N. (2003) Marginal nonparametric kernel regression accounting for within-subject correlation, *Biometrika*, **90**, 43–52. [Cited on page 2]

Department of Economics, University of Wisconsin, Madison. 1180 Observatory Drive, Madison, WI 53706-1393, USA.

Email address: yuya.shimizu@wisc.edu