Additive partially linear models for massive heterogeneous data

Binhuan Wang

Department of Population Health
New York University School of Medicine
e-mail: binhuan.wang@nyulangone.org

Yixin Fang[*]

Department of Mathematical Sciences
New Jersey Institute of Technology
e-mail: yixin.fang@njit.edu

Heng Lian

Department of Mathematics
City University of Hong Kong
e-mail: henglian@cityu.edu.hk

and

Hua Liang

Department of Statistics
George Washington University
e-mail: hliang@email.gwu.edu

Abstract: We consider an additive partially linear framework for modelling massive heterogeneous data. The major goal is to extract multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. We propose an aggregation type of estimators for the commonality parameters that possess the asymptotic optimal bounds and the asymptotic distributions as if there were no heterogeneity. This oracle result holds when the number of sub-populations does not grow too fast and the tuning parameters are selected carefully. A plug-in estimator for the heterogeneity parameter is further constructed, and shown to possess the asymptotic distribution as if the commonality information were available. Furthermore, we develop a heterogeneity test for the linear components and a homogeneity test for the non-linear components accordingly. The performance of the proposed methods is evaluated via simulation studies and an application to the Medicare Provider Utilization and Payment data.

---

[*]Corresponding author.

Contents

1. Introduction

Recent revolutions in technologies have produced many kinds of massive data, where the number of variables p is fixed but the sample size N is very large. In this paper we consider massive heterogeneous data. The analysis of non-massive heterogeneous data has been well studied in the literature. For example, non-massive heterogeneous data can be handled by fitting mixture models [1] and by modeling variance functions [3]. However, as far as we are aware, [21] is the only work that considers the analysis of massive heterogeneous data. They proposed a partially linear framework for modelling massive heterogeneous data, attempting to extract the common feature across all sub-populations while exploring heterogeneity of each sub-population. But the partially linear framework can only deal with only one common feature. In this paper, we propose an additive partially linear framework for modelling massive heterogeneous data, which can be applied to extract several common features across all sub-populations while exploring heterogeneity of each sub-population.

The additive partially linear models (APLMs) are a generalization of multiple linear regression models, and at the same time they are a special case of generalized additive nonparametric regression models [7]. As discussed in [11], APLMs allow an easier interpretation of the effect of each variable and are preferable to completely nonparametric additive models, since they combine both parametric and nonparametric components when it is believed that the response variable depends on some variables in a linear way but is non-linearly related to the remaining independent variables. Estimation and inference for APLMs have been

well studied in literature [e.g., 2, 13]. Recently, [6] proposed an approach for the analysis of heterogeneous data, fitting both the mean function and variance function using different additive partially linear models.

In this paper, we generalize the partially linear model (PLM) considered in [21] and propose an additive partially linear model (APLM) for modeling massive heterogeneous data. Let $\{(Y_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)\}_{i=1}^N$ be the observations from a sample of N subjects. We assume that there exist s independent sub-populations, and the data from the jth sub-population follow the following additive partially linear model,

$$Y^{(j)} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0^{(j)} + \sum_{k=1}^{K} g_{0k}(Z_k) + \varepsilon, \ \ \text{for} \ \ j = 1, \ldots, s, \tag{1.1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$, $\boldsymbol{\beta}_0^{(j)} = (\beta_{01}^{(j)}, \ldots, \beta_{0d}^{(j)})^{\mathrm{T}}$ is the vector of unknown parameters for jth sub-population, $g_{01}, \ldots, g_{0K}$ are unknown smooth functions, and ε has zero mean and variance σ^2. Under model (1.1), $Y^{(j)}$ depends on $\boldsymbol{X}$ linearly but with coefficients varying across different sub-populations, whereas $Y^{(j)}$ depends on $\boldsymbol{Z}$ through additive non-linear functions that are common to all sub-populations. This model implies that the heterogeneity of the data is coming from the difference among $\boldsymbol{\beta}_0^{(j)}$, $j = 1, \ldots, s$.

Compared with [21], the novelty of this paper is two-fold. First, we consider a more general and practical model. The partially linear model considered in [21] is a special case of (1.1) where $K = 1$. Second, we use a different nonparametric tool (i.e., the regression splines tool) to fit the non-parametric functions than the reproducing kernel Hilbert Space (RKHS) tool that was used in [21], making the theoretical development easier and computational implementation faster. The first fold of novelty is more significant than the second one, because it allows to consider more than one non-parametric components ($K > 1$). However, the theoretical novelty of our approach is limited, because it would be straightforward generate the PLM in [21] to the APLM using the same RKHS tool. Instead, we propose to use the regression splines tool, because the resulting computational implementation is fast, especially for massive data. In order to understand this, let's count the number of knots when the RKHS tool and the regression splines are used, respectively. If the RKHS tool is used, the number of knots is N for each non-parametric component, and therefore the total number of knots is equal to KN. If the regression splines is used, the number of knots is to be denoted as J_N, satisfying that $J_N \ll N$, and therefore the total number of knots is equal to KJ_N. The computational gain is small if the PLM is fitted, but the gain is significant if the APLM is fitted for massive data where N is extremely large.

The rest of the paper is organized as follows. We propose the main methods in Section 2 and some hypothesis testing procedures in Section 3, deriving their asymptotic properties. We evaluate the performance of the proposed methods via simulation studies in Section 4 and a real data application in Section 5. We conclude the paper with a brief summary in Section 6 and relegate all the technical proofs to the Appendix.

2. Methods

2.1. Notation and assumptions

Recall that $\boldsymbol{\beta}_0^{(j)}$ is the true sub-population specific parameter-vector for the jth sub-population, $j = 1, \ldots, s$, and $g_0(\boldsymbol{z}) = g_{01}(z_1) + \cdots + g_{0K}(z_K)$ is the true additive common non-parametric function. Without loss of generality, assume that $g_{0k} = g_{0k}(\cdot)$, $k = 1, \ldots, K$, have a common support $[0, 1]$. We propose to use polynomial splines [2] to approximate smooth function g_{0k}, $k = 1, \ldots, K$. Let $\mathcal{S}_N$ be the space of polynomial splines on $[0, 1]$ of degree $\varrho \geq 1$, with a sequence of J_N interior knots,

$$t_{-\varrho} = \cdots = t_{-1} = t_0 = 0 < t_1 < \cdots < t_{J_N} < 1 = t_{J_N+1} = \cdots = t_{J_N+\varrho+1},$$

where J_N increases with the overall sample size N. Although we can choose different sequences of interior knots for different non-parametric functions in different sub-populations, for simplicity, as in [11], here we consider the same sequence of equally spaced knots and let $h_N = 1/(J_N + 1)$ be the distance between neighboring knots.

Assume that $\boldsymbol{X}_i$ are i.i.d. with $\boldsymbol{X}$ and $\boldsymbol{Z}_i$ are i.i.d. with $\boldsymbol{Z}$. Define $\boldsymbol{T} = (\boldsymbol{X}, \boldsymbol{Z})$. Let $m_0^{(j)}(\boldsymbol{T}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0^{(j)} + g_0(\boldsymbol{Z})$, $\Gamma(\boldsymbol{z}) = E(\boldsymbol{X}|\boldsymbol{Z} = \boldsymbol{z})$, and $\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \Gamma(\boldsymbol{Z})$. And $\boldsymbol{C}^{\otimes 2}$ denotes $\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}}$ for any matrix or vector $\boldsymbol{C}$. Let r be a positive integer and $\nu \in (0, 1]$ such that $p = r + \nu > 2$. Let $\mathcal{H}$ be the collection of functions h on $[0, 1]$ whose rth derivative exists and satisfies the Lipschitz condition of order ν,

$$\left| h^{(r)}(z') - h^{(r)}(z) \right| \leq C|z' - z|^\nu, \forall\, 0 \leq z', z \leq 1,$$

where and hereafter C is a generic positive constant. In order to derive asymptotic results, we make the following mild assumptions.

(A1) Each component function $g_{0k} \in \mathcal{H}, k = 1, \ldots, K$;
(A2) The distribution of $\boldsymbol{Z}$ is absolutely continuous and its density f is bounded away from zero and infinity on $[0, 1]^K$;
(A3) There exists $c > 0$ such that $c\|\boldsymbol{\omega}\|^2 \leq \boldsymbol{\omega}^{\mathrm{T}}E(\boldsymbol{X}^{\otimes 2}|\boldsymbol{Z} = \boldsymbol{z})\boldsymbol{\omega} \leq C\|\boldsymbol{\omega}\|^2$, for any vector $\boldsymbol{\omega} \in \mathcal{R}^d$;
(A4) The number of interior knots J_N satisfies: $N^{1/(4p)} \ll J_N \ll N^{1/4}$;
(A5) The projection function $\Gamma(\boldsymbol{z})$ has the additive form $\Gamma(\boldsymbol{z}) = \Gamma_1(z_1) + \cdots + \Gamma_K(z_K)$, where $\Gamma_k \in \mathcal{H}, E[\Gamma_k(z_k)] = 0$ and $E[\Gamma_k(z_k)]^2 < \infty, k = 1, \ldots, K$.

In addition, to quantify the asymptotic consistencies of the non-parametric estimators, we consider both the empirical norms and the corresponding population norms. Let $\|\boldsymbol{z}\|$ be the Euclidean norm, $\|\boldsymbol{z}\|_\infty$ be the supremum norm, and $\|\boldsymbol{z}\|_1$ be the absolute-value norm of a vector $\boldsymbol{z}$, respectively. For a matrix $\boldsymbol{C}$, its L_2-norm is defined as $\|\boldsymbol{C}\|_2 = \sup_{\|\boldsymbol{u}\| \neq 0} \|\boldsymbol{C}\boldsymbol{u}\|/\|\boldsymbol{u}\|$. Let $\|\varphi\|_\infty = \sup_{x \in [0,1]} |\varphi(x)|$ be the supremum norm of a function φ on $[0, 1]$. Following [9] and [14], for any measurable function ϕ_1 and ϕ_2 on $[0, 1]^K$, the empirical inner

product and norm for the jth sub-sample and the whole sample, respectively, are defined as

$$\langle \phi_1, \phi_2 \rangle_{jn} = \frac{1}{n} \sum_{i \in \mathcal{G}_j} \phi_1(\boldsymbol{Z}_i)\phi_2(\boldsymbol{Z}_i), \quad \|\phi\|_{jn}^2 = \frac{1}{n} \sum_{i \in \mathcal{G}_j} \phi^2(\boldsymbol{Z}_i),$$

$$\langle \phi_1, \phi_2 \rangle_N = \frac{1}{N} \sum_{i=1}^{N} \phi_1(\boldsymbol{Z}_i)\phi_2(\boldsymbol{Z}_i), \quad \|\phi\|_N^2 = \frac{1}{N} \sum_{i=1}^{N} \phi^2(\boldsymbol{Z}_i).$$

If ϕ_1 and ϕ_2 are L^2-integrable, the population inner product and norm are defined as

$$\langle \phi_1, \phi_2 \rangle = \int_{[0,1]^K} \phi_1(\boldsymbol{z})\phi_2(\boldsymbol{z})f(\boldsymbol{z})d\boldsymbol{z}, \quad \|\phi\|_2^2 = \int_{[0,1]^K} \phi^2(\boldsymbol{z})f(\boldsymbol{z})d\boldsymbol{z},$$

where f is the density of $\boldsymbol{Z}$. Similarly, for the kth component of $\boldsymbol{Z}$, Z_k with density f_k, the empirical norm on the jth sub-sample, the empirical norm on the whole sample, and the population norm of any L^2-integrable univariate function φ on $[0,1]$ are defined as

$$\|\varphi\|_{jnk}^2 = \frac{1}{n} \sum_{i \in \mathcal{G}_j} \varphi^2(Z_{ik}), \quad \|\varphi\|_{Nk}^2 = \frac{1}{n} \sum_{i=1}^{N} \varphi^2(Z_{ik}), \quad \|\varphi\|_{2k}^2 = \int_0^1 \varphi^2(z_k)f_k(z_k)dz_k.$$

2.2. Estimations for each sub-population

First we consider the estimations for $\boldsymbol{\beta}_0^{(j)}$ and $g_0 = g_0(\cdot)$ based on the data from the jth sub-population only, $j = 1, \ldots, s$. To this aim, let G_j denotes the index set of all the observations from the sub-population j, and let $\mathcal{G}_n^{(j)} = \{g^{(j)}(\cdot)\}$ be the collection of additive functions with the form that $g^{(j)}(\boldsymbol{z}) = g_1^{(j)}(z_1) + \cdots + g_K^{(j)}(z_K)$, where each component function $g_k^{(j)} \in \mathcal{S}_N$ and $\sum_{i \in G_j} g_k^{(j)}(Z_{ik}) = 0$. Thus $\sum_{i \in G_j} g^{(j)}(\boldsymbol{Z}_i) = 0$ for any $g^{(j)} \in \mathcal{G}_n^{(j)}$. For the jth sub-population, we consider the following estimators,

$$(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{g}^{(j)}) = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathcal{R}^d, \ g \in \mathcal{G}_n^{(j)}} \left\{ L_n^{(j)}(\boldsymbol{\beta}, g) = \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - g(\boldsymbol{Z}_i) \right]^2 \right\}. \quad (2.1)$$

For the kth covariate Z_k, let $b_{m,k}(z_k)$ be the B-spline basis functions of degree ϱ equipped with J_N knots defined above. For any $g \in \mathcal{G}_n^{(j)}$, we can write $g(\boldsymbol{z}) = \boldsymbol{b}(\boldsymbol{z})^{\mathrm{T}}\boldsymbol{\gamma}$, where $\boldsymbol{b}(\boldsymbol{z}) = \{b_{m,k}(z_k), m = -\varrho, \ldots, J_N, k = 1, \ldots, K\}^{\mathrm{T}}$, which is a $K(J_N+\varrho+1)$-dim vector given $\boldsymbol{z}$, along with $K(J_N+\varrho+1)$-dim coefficient-vector $\boldsymbol{\gamma} = \{\gamma_{m,k}, m = -\varrho, \ldots, J_N, k = 1, \ldots, K\}^{\mathrm{T}}$. Therefore, (2.1) is equivalent to

$$\operatorname*{argmin}_{\boldsymbol{\beta} \in \mathcal{R}^d, \ \boldsymbol{\gamma} \in \mathcal{R}^{K(J_N+\varrho+1)}} \left\{ l_n^{(j)}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{b}(\boldsymbol{Z}_i)^{\mathrm{T}}\boldsymbol{\gamma} \right]^2 \right\}, \quad (2.2)$$

if we consider the empirically centered estimator $\widehat{g}^{(j)}(\boldsymbol{z}) = \sum_{k=1}^{K} \widehat{g}_k^{(j)}(\boldsymbol{z})$, where

$$\widehat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho}^{J_N} \widehat{\gamma}_{m,k} b_{m,k}(z_k) - \frac{1}{n} \sum_{i \in G_j} \sum_{m=-\varrho}^{J_N} \widehat{\gamma}_{m,k} b_{m,k}(z_{ik}). \qquad (2.3)$$

We derive some asymptotic results associated with the sub-population specific estimators, summarized in the following theorem.

Theorem 2.1. *Under Assumptions (A1)–(A5), if the number of knots satisfies that $J_N \ll n^{1/2}$, we have, for each sub-population, $j = 1, \ldots, s$,*

$$\|\widehat{g}^{(j)} - g_0\|_2 = O_P\left(J_N^{1/2} n^{-1/2} + h_N^p\right) \quad \text{and}$$

$$\|\widehat{g}^{(j)} - g_0\|_{jn} = O_P\left(J_N^{1/2} n^{-1/2} + h_N^p\right).$$

If the number of knots further satisfies that $J_N \gg n^{1/(2p)}$ we have

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{D}^{-1}\big),$$

where $\boldsymbol{D} = E(\widetilde{\boldsymbol{X}}^{\otimes 2})$.

Remark 1. Assume that we consider $s = O(N^{1-\gamma})$ sub-samples, each sub-sample of $n = O(N^{\gamma})$ observations, where γ is some positive number between 0 and 1. In order to minimize the mean-square error of estimating g_0, $O_P(J_N^{1/2} n^{-1/2} + h_N^p)$, the best selection of J_N is $O(N^{\frac{\gamma}{2p+1}})$, or equivalently, $O(n^{\frac{1}{2p+1}})$. Under this selection, the mean-square error achieves the optimal rate, $O(N^{\frac{p\gamma}{2p+1}})$, or equivalently, $O(n^{\frac{p}{2p+1}})$.

Remark 2. On the other hand, in order to ensure that $\widehat{\boldsymbol{\beta}}^{(j)}$ is $\sqrt{n}$-consistent for estimating $\boldsymbol{\beta}_0^{(j)}$, we should adopt under-smoothing tuning with $J_N \gg n^{1/(2p)}$ and carefully determine a balance between the number of sub-samples and the size of each sub-sample. For example, this can be achieved if we select J_N as $O(N^q)$ with $1/(4p) < q < 1/4$, and consider $s = O(N^{1-\gamma})$ sub-samples, each sub-sample of $n = O(N^{\gamma})$, with $2q < \gamma < 2pq$. The order of J_N is consistent with the existing results in the literature. The recommended balance between s and n provides a guidance for the appropriate application of the divide-and-conquer strategy.

2.3. *Aggregation of commonality*

We consider the aggregated estimator, $\overline{g}(\boldsymbol{z}) = \frac{1}{s} \sum_{j=1}^{s} \widehat{g}^{(j)}(\boldsymbol{z})$, as the final estimator of $g_0(\boldsymbol{z})$ based on the whole sample. Let $\mathcal{G}_N$ be the collection of functions with the additive form $g(\boldsymbol{z}) = g_1(z_1) + \cdots + g_K(z_K)$, where $g_k \in \mathcal{S}_N$ and $\sum_{j=1}^{s} \sum_{i \in G_j} g_k(Z_{ik}) = 0$. Thus, for any $g \in \mathcal{G}_N$, $\sum_{j=1}^{s} \sum_{i \in G_j} g(\boldsymbol{Z}_i) = 0$. In

order to ensure that $\overline{g} \in \mathcal{G}_N$, as in (2.3), we center the individual estimator $\widehat{g}_k^{(j)}(z_k)$ via $\widehat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho}^{J_N} \widehat{\gamma}_{m,k} b_{m,k}(z_k) - \frac{1}{N} \sum_{i=1}^{N} \sum_{m=-\varrho}^{J_N} \widehat{\gamma}_{m,k} b_{m,k}(z_{ik})$. To abuse the notation, we still denote the centered estimator as $\widehat{g}_k^{(j)}(z_k)$ and $\widehat{g}^{(j)}(\boldsymbol{z}) = \sum_{k=1}^{K} \widehat{g}_k^{(j)}(z_k)$. We derive the mean-square error of $\overline{g}$ in the following theorem.

Theorem 2.2. *Under Assumptions (A1)–(A5), if $J_N \ll n^{1/2}$, we have*

$$\|\overline{g} - g_0\|_2 = O_P\left(J_N^{1/2} N^{-1/2} + h_N^p\right), \quad \text{and} \quad \|\overline{g} - g_0\|_N = O_P\left(J_N^{1/2} N^{-1/2} + h_N^p\right).$$

Remark 3. In order to minimize the mean-square error of estimating g_0 using the aggregated estimator, if we select J_N as $O(N^{\frac{1}{2p+1}})$, the mean-square error achieves the optimal rate $O(N^{\frac{p}{2p+1}})$.

Remark 4. We compare the mean-square error of $\overline{g}$ with that of the following "oracle estimator":

$$\widehat{g}_{\text{oracle}} = \underset{g \in \mathcal{G}_N}{\operatorname{argmin}} \frac{1}{2} \sum_{j=1}^{s} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0^{(j)} - g(\boldsymbol{Z}_i)\right]^2.$$

assuming $\boldsymbol{\beta}_0^{(j)}$, $j = 1, \ldots, s$, are known. Following the proof of Theorem 2.1, we can show that $\|\widehat{g}_{\text{oracle}} - g_0\|_2 = O_P\left(J_N^{1/2} N^{-1/2} + h_N^p\right)$. Therefore, as long as $n \gg J_N^2$, the means-square errors of the aggregated estimator $\overline{g}$ and the oracle estimator $\widehat{g}_{\text{oracle}}$ are of the same order.

We conclude this subsection with some results for the massive homogeneous data where $\boldsymbol{\beta}_0^{(j)} \equiv \boldsymbol{\beta}_0, j = 1, \ldots, s$. These results are of their own interest, when the divide-and-conquer strategy is applied to massive homogeneous data, where $\boldsymbol{\beta}_0$ and g_0 are estimated using the aggregated estimators $\overline{\boldsymbol{\beta}} = \frac{1}{s} \sum_{j=1}^{s} \widehat{\boldsymbol{\beta}}^{(j)}$ and $\overline{g}$, respectively. The result for $\overline{g}$ is the same as that in Theorem 2.2 and the result for $\overline{\boldsymbol{\beta}}$ is stated in the following corollary.

Corralary 2.1. *Consider homogeneous massive data where $\boldsymbol{\beta}_0^{(j)} \equiv \boldsymbol{\beta}_0, j = 1, \ldots, s$. Under Assumptions (A1)–(A5), if $J_N \gg N^{1/(2p)}$ and $n \gg N^{1/2}$, we have*

$$\sqrt{N}(\overline{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{D}^{-1}\right).$$

2.4. Efficiency boosting for heterogeneous parameters

The asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}^{(j)}$ derived in Theorem 2.1 shows that there is some room to improve the estimation efficiency, because $\boldsymbol{D}^{-1} = E^{-1}(\widetilde{\boldsymbol{X}}^{\otimes 2})$ is bigger than the Cramer-Rao lower bound, $E^{-1}(\boldsymbol{X}^{\otimes 2})$. Therefore, we re-substitute the aggregated estimator of g, $\overline{g}$, into (2.1) to improve the efficiency

of estimating $\boldsymbol{\beta}_0^{(j)}$. This leads to the following more efficient estimator,

$$\breve{\boldsymbol{\beta}}^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta}^{(j)} \in \mathcal{R}^d} \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}^{(j)} - \overline{g}(\boldsymbol{Z}_i) \right]^2. \tag{2.4}$$

for $j = 1, \cdots, s$. We derive the asymptotic normality of $\breve{\boldsymbol{\beta}}^{(j)}$ in the following theorem.

Theorem 2.3. *Under Assumptions (A1)–(A5), if J_N satisfies the condition that $J_N \ll n^{1/2}$ given in the first part of Theorem 2.1 and the condition that $J_N \gg N^{1/(2p)}$ given in Corollary 2.1, and it further satisfies that $J_N \ll s^{1/2}$, then we have*

$$\sqrt{n}\big(\breve{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\big) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}, \sigma^2 \boldsymbol{A}^{-1}\big),$$

where $\boldsymbol{A} = E(\boldsymbol{X}^{\otimes 2})$.

Remark 5. As in Remarks 1–2, assume that we consider $s = O(N^{1-\gamma})$ sub-samples, each sub-sample of $n = O(N^\gamma)$ observations, where γ is some positive number between 0 and 1. In order to satisfy all the conditions in Theorem 2.3, we can consider $N^{2q} \ll n \ll N^{1-2q}$, with $1/(2p) < q < 1/4$, and select $J_N = O(N^q)$. If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are not independent, then $\boldsymbol{A}^{-1} \prec \boldsymbol{D}^{-1}$, implying that we can achieve such efficiency boosting through balancing between n and s.

2.5. Practical issues

In this subsection, we consider several practical issues, including selection of the number of knots J_N, determination of linear components and non-linear components, and how to conduct statistical inference.

We first consider the selection of J_N. All the theoretical results need Assumption (A4): $N^{\frac{1}{4p}} \ll J_N \ll N^{\frac{1}{4}}$. Besides this, different theorem (or corollary) needs different an extra condition. Here is the list of those conditions:

$$\text{(a)} \quad J_N \ll n^{1/2};$$
$$\text{(b)} \quad J_N \gg n^{1/(2p)};$$
$$\text{(c)} \quad J_N \gg N^{1/(2p)} \text{ and } n \gg N^{1/2};$$
$$\text{(d)} \quad J_N \gg N^{1/(2p)} \text{ and } J_N \ll s^{1/2}.$$

In Theorem 2.1, under Condition (a), we derive the bound for the mean-square error of each sub-population specific estimator $\widehat{g}^{(j)}$, $j = 1, \cdots s$. In Theorem 2.1, under Conditions (a) and (b), we derive the asymptotic normality for each sub-population specific estimator $\widehat{\boldsymbol{\beta}}^{(j)}$, $j = 1, \cdots s$. In Theorem 2.2, under Condition (a), we derive the bound for the mean-square error of the aggregated estimator $\overline{g}$. In Corollary 2.1, under Condition (c) and for the massive homogeneous data, we derive the asymptotic normality for the aggregated estimator $\overline{\boldsymbol{\beta}}$. In Theorem 2.3,

under Condition (d), we derive the asymptotic normality for each sub-population specific efficiency-boosted estimator $\breve{\boldsymbol{\beta}}^{(j)}$. These conditions can be satisfied by carefully selecting the balance between n and s, with some guidance provided in Remarks 1–5.

However, it is hard to use these theoretical requirements on the order of J_N to guide the selection of J_N in practice. If computational power allows, we can utilize cross-validation to select J_N adaptively. If cross-validation is used, as discussed in [20], we should consider distributed cross-validation aiming for global optimality (selecting a single J_N based on the entire dataset) instead of subsample cross-validation aiming for local optimality (selecting an J_N for each subsample separately). As we consider massive data here, we don't recommend any data-driven tuning procedure, which requires heavy computational cost. Instead, as in most studies using regression splines, we recommend the fixed choice of the number of internal knots J_N (e.g., some small number between 2 and 10). Although pre-specifying a fixed J_N which might be sub-optimal, it is much more convenient and computationally efficient than any data-driven procedure. This is another advantage of the regression splines compared with smoothing splines: the number of knots in the regression splines can be determined directly in terms of model construction without looking at the data, while in the smoothing splines the tuning parameter λ, which controls the balance between goodness-of-fit and function smoothness, has to be determined adaptively. In order to determine tuning parameter λ in the smoothing splines, we have to use data-driven procedure as in [20].

Another practical issue is the determination of linear components and non-linear components. In practice, we rely on graphical tools such as box-plot and scatterplot to have a rough idea on which components might be linear or non-linear. For the setting in this paper, we consider parametric function for each heterogeneous component (the sample size is smaller for each subsample, so a simpler model is considered) and non-parametric function for each homogeneous component (the sample size is larger for the entire sample, so a more flexible model is considered). Motivated by [12] and [21], some formal hypothesis testing methods for heterogeneity and homogeneity are proposed in the next section.

The third practical issue is conducting statistical inferences about the regression parameters. In the above theorems, we derive the explicit formulas for their corresponding covariance matrices by examining the asymptotic normality. If the covariance matries were to involve the density of error term, we wouldn't apply the plug-in procedure to estimate them as discussed in Section 4 of [16]. Fortunately, all the asymptotic covariance matrices we developed in the above theorems can be estimated using the plug-in procedure and conducting statistical inferences is straightforward. In addition, as discussed in Remark 2, in order to achieve the estimation efficiency of regression parameters, the non-parametric functions should be under-smoothing. This is another reason that we don't recommend data-driven method for the determination of J_N. If J_N is selected adaptively, the estimation of regression parameters may be sub-efficient and it is also subject to the problem of inference-after-selection.

3. Hypothesis testing

3.1. Testing heterogeneity

As in [21], we develop statistical tests for the heterogeneity of the linear components across sub-populations. First, we consider a general class of pairwise test for heterogeneous parameters. Then, we develop a more general heterogeneity test involving many (up to as many as s) sub-populations.

First, consider the following general class of pairwise test for heterogeneous parameters:

$$H_0: \ \boldsymbol{Q}\big(\boldsymbol{\beta}_0^{(j_1)} - \boldsymbol{\beta}_0^{(j_2)}\big) = \boldsymbol{0} \quad \text{vs.} \quad H_a: \ \boldsymbol{Q}\big(\boldsymbol{\beta}_0^{(j_1)} - \boldsymbol{\beta}_0^{(j_2)}\big) \neq \boldsymbol{0}, \qquad (3.1)$$

where $j_1 \neq j_2 \in \{1, \ldots, s\}$, and $\boldsymbol{Q} = (\boldsymbol{q}_1^{\mathrm{T}}, \ldots, \boldsymbol{q}_{d_1}^{\mathrm{T}})^{\mathrm{T}}$ is a $d_1 \times d$ matrix with $d_1 \leq d$. This class of tests includes testing if either the whole vector or specific entries of $\boldsymbol{\beta}_0^{(j_1)}$ are equal to those of $\boldsymbol{\beta}_0^{(j_2)}$. It is straightforward to consider $\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big)$ or $\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \breve{\boldsymbol{\beta}}^{(j_2)}\big)$ as test statistic, which is based on the estimators from Subsection 2.2 or the estimators from Subsection 2.4, respectively. We summarize the asymptotic properties of these two test statistics in the following theorem, based on which we can conduct the Wald tests.

Theorem 3.1. *If the conditions in Theorem 2.1 hold, under the null hypothesis* (3.1),

$$\sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big) \xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}, 2\sigma^2 \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^{\mathrm{T}}\big).$$

Furthermore, if the conditions in Theorem 2.3 hold, under the null hypothesis (3.1),

$$\sqrt{n}\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \breve{\boldsymbol{\beta}}^{(j_2)}\big) \xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}, 2\sigma^2 \boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{Q}^{\mathrm{T}}\big).$$

Based on Theorem 3.1, we can construct the Wald tests as what follows,

$$\Psi_1 \ = \ I\left\{\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big) \notin \sqrt{\frac{2}{n}}\sigma(\boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^{\mathrm{T}})^{\frac{1}{2}} Z_{1-\alpha/2}\right\},$$

$$\Psi_2 \ = \ I\left\{\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \breve{\boldsymbol{\beta}}^{(j_2)}\big) \notin \sqrt{\frac{2}{n}}\sigma(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{Q}^{\mathrm{T}})^{\frac{1}{2}} Z_{1-\alpha/2}\right\},$$

where α is a given significant level and $Z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of a standard normal distribution. It is clear that Ψ_2 is more powerful than Ψ_1 because of a smaller variance of $\breve{\boldsymbol{\beta}}^{(j)}$ compared with $\widehat{\boldsymbol{\beta}}^{(j)}$. Coverage probabilities, length of confidence intervals, nominal levels and powers are empirically evaluated in simulation studies, which confirm our theoretical results.

Then, we consider a more general heterogeneity test involving many sub-populations indexed by $j, j \in \mathcal{S} \subseteq \{1, \ldots, s\}$. Note that $|\mathcal{S}|$ is allowed to be as large as s, i.e., all sub-populations. The null and alternative hypotheses are formulated as

$$H_0: \ \boldsymbol{Q}\boldsymbol{\beta}_0^{(j)} = \boldsymbol{Q}\boldsymbol{\check{\beta}}_0^{(j)} \ \ \forall j \in \mathcal{S} \quad \text{vs.} \quad H_a: \boldsymbol{Q}\,\boldsymbol{\beta}_0^{(j)} \neq \boldsymbol{Q}\boldsymbol{\check{\beta}}_0^{(j)} \ \ \exists j \in \mathcal{S}, \qquad (3.2)$$

where $\boldsymbol{\check{\beta}}_0^{(j)}$'s are some predetermined values.

To test (3.2), it is natural to define the following test statistic:

$$T_{\mathcal{S}} = \max_{j \in \mathcal{S}} \sqrt{n} \|\boldsymbol{Q}(\boldsymbol{\check{\beta}}^{(j)} - \boldsymbol{\check{\beta}}_0^{(j)})\|_{\infty}.$$

Its distribution can be approximated by bootstrapping the quantity

$$W_{\mathcal{S}} = \max_{j \in \mathcal{S}} \frac{1}{\sqrt{n}} \left\| \sum_{i \in G_j} \boldsymbol{Q}(\boldsymbol{\widehat{A}}_n^{(j)})^{-1} \boldsymbol{X}_i e_i \right\|_{\infty},$$

where $\boldsymbol{\widehat{A}}_n^{(j)} = \frac{1}{n} \sum_{i \in G_j} \boldsymbol{X}_i^{\otimes 2}$ and $e_i \sim \mathcal{N}(0, \sigma^2)$'s are i.i.d. The following theorem summarizes the consistency of the proposed multiplier bootstrap method.

Theorem 3.2. *Suppose the conditions in Theorem 2.3 are satisfied. Under the null hypothesis (3.2), for any $\mathcal{S} \subseteq \{1, \ldots, s\}, |\mathcal{S}| = d$, if $s \gg J_N^2 log(pd)$, $(log(pdn))^7/n \leq c_1 n^{-c_2}$ for some constants $c_1, c_2 > 0$, and $p^2 log(pd)/\sqrt{n} = o(1)$, then we have*

$$\sup_{\alpha \in (0,1)} \big|P\big(T_{\mathcal{S}} > c_{\mathcal{S}}(\alpha)\big) - \alpha\big| = o(1),$$

where $c_{\mathcal{S}} = \inf\{w \in \mathcal{R} : P(W_{\mathcal{S}} \leq w|\boldsymbol{X}) \geq 1 - \alpha\}$.

Remark 6. Actually, the above hypotheses can be extended to test the heterogeneity of all sub-populations without specifying $\boldsymbol{\check{\beta}}_0^{(j)}$'s. With a similar argument as that in [21], we only need to test if the differences of heterogeneity parameters between any two consecutive sub-populations equal to zero. The modified test statistic is

$$T_{\mathcal{S}}' = \max_{1 \leq j \leq s-1} \sqrt{n} \|\boldsymbol{Q}(\boldsymbol{\check{\beta}}^{(j)} - \boldsymbol{\check{\beta}}^{(j+1)})\|_{\infty},$$

and corresponding bootstrap quantity is given by

$$W_{\mathcal{S}}' = \max_{1 \leq j \leq s-1} \frac{1}{\sqrt{n}} \left\| \sum_{i \in G_j} \boldsymbol{Q}(\boldsymbol{\widehat{A}}_n^{(j)})^{-1} \boldsymbol{X}_i e_i - \sum_{i \in G_{j+1}} \boldsymbol{Q}(\boldsymbol{\widehat{A}}_n^{(j+1)})^{-1} \boldsymbol{X}_i e_i \right\|_{\infty}.$$

Similarly, we define $c_{\mathcal{S}}' = \inf\{w \in \mathcal{R} : P(W_{\mathcal{S}}' \leq w|\boldsymbol{X}) \geq 1 - \alpha\}$.

3.2. Testing homogeneity

Now we consider the test of whether the non-linear components, $g_{0k}(Z_k), k = 1, 2, \ldots, K$ are homogeneous across all s sub-populations, which is the necessity of doing aggregation of commonality. With a little abuse of notation, for the jth sub-population, we denote the true unknown smooth functions as $g_{0k}^{(j)}(Z_k), k =$

$1, \ldots, K$. We apply the likelihood ratio principle to the following homogeneity test for the kth smooth function,

$$H_0: \; g_{0k}^{(1)}(Z_k) = \cdots = g_{0k}^{(s)}(Z_k) \quad \text{vs.} \quad H_a: g_{0k}^{(j)}\text{'s are not all the same.} \quad (3.3)$$

Let $g_{-k} = \sum_{k' \neq k} g_{k'}$ and $\boldsymbol{Z}_{-k} = (Z_1, \ldots, Z_{k-1}, Z_{k+1}, \ldots, Z_K)$. The subscript $-k$ means removing the kth component therein. Define

$$L_{nk}^{(j)}(\boldsymbol{\beta}, g_k, g_{-k}) = \frac{1}{2n} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} - g_k(Z_k) - g_{-k}(\boldsymbol{Z}_{i,-k}) \right]^2.$$

We can construct a likelihood ratio test statistic as

$$\text{LRT}_{nk}^s = \sum_{j=1}^{s-1} \left\{ L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{g}_k^{(j)}, \widehat{g}_{-k}^{(j)}) - L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{g}_k^{(j+1)}, \widehat{g}_{-k}^{(j)}) \right\}. \quad (3.4)$$

Before providing the limiting distribution for the above test, extra notations are required. We say that a statistic T_n is nearly $\chi_{b_n}^2$, denoted as $T_n \overset{a}{\sim} \chi_{b_n}^2$, if $(2b_n)^{-1/2}(T_n - b_n)$ weakly converges to $\mathcal{N}(0,1)$ for some sequence $b_n \to \infty$.

Theorem 3.3. *Suppose the conditions in Theorem 2.1 are satisfied. Under the null hypothesis (3.3), if $J_N^{1/2} = o(s)$, we have*

$$-\frac{1}{3\sigma^2} n \cdot \text{LRT}_{nk}^s \overset{a}{\sim} \chi_{u_N}^2,$$

where $u_N = \frac{2}{3}(s-1)J_N$.

Remark 7. $J_N^{1/2} = o(s)$ is a weak requirement, implying the number of sub-populations cannot be too small to borrow sufficient information across all of them. This phenomenon is also observed in [12], in which it is called the "blessing of aggregation". According to Remarks 1 and 2, $J_N = O(N^q)$, where $1/(4p) < q < 1/4$, implying that $J_N^{1/2} = o(s)$ means $\gamma < 1 - q/2$ with the minimum value of $7/8$ of the right hand side.

Remark 8. The above theorem considers testing one smooth function. If we want to test the summation of all smooth functions, the result can be proved similarly, which is summarized in the following theorem by removing the extra condition $J_N^{1/2} = o(s)$ due to no usage of both $\widehat{g}_k^{(j+1)}$ and $\widehat{g}_{-k}^{(j)}$ in the $L_{nk}^{(j)}$.

Theorem 3.4. *Suppose the conditions in Theorem 2.1 are satisfied. Under the null hypothesis $H_0: \; g_0^{(1)}(\boldsymbol{Z}) = \cdots = g_0^{(s)}(\boldsymbol{Z})$, we have*

$$-\frac{1}{3\sigma^2} n \cdot \text{LRT}_{nk}^s \overset{a}{\sim} \chi_{u_N}^2,$$

where $u_N = \frac{2}{3}(s-1)KJ_N$.

4. Simulation Studies

We conduct simulation studies to examine the impact of the balance between sub-population sizes n and the number of sub-population s on the performance of the proposed estimators, $\overline{g}$ and $\breve{\boldsymbol{\beta}}^{(j)}$. We consider the following additive partially linear model with two nonparametric components ($K = 2$) as the data generating model:

$$
\begin{aligned}
Y^{(j)} &= X\beta_0^{(j)} + g_1(Z_1) + g_2(Z_2) + \varepsilon, \\
g_1(Z_1) &= 5\sin\{2\pi(Z_1 + 1)\}, \\
g_2(Z_2) &= 100\left(e^{-1.625(Z_2+1)} - 4e^{-3.25(Z_2+1)} + 3e^{-4.825(Z_2+1)}\right) - C_0,
\end{aligned}
$$

where ε is generated from normal distribution $N(0,1)$, Z_1, Z_2 and W are generated independently from uniform distribution $U(-1,1)$, $X = \frac{1}{2}(W + Z_1)$, and C_0 is taken as $100(1 - e^{-3.25})/3.25 - 400(1 - e^{-6.5})/6.5 + 300(1 - e^{-9.75})/9.75$ to make sure that $E\{g_1(Z_1)\} = E\{g_2(Z_2)\} = 0$. We can show that $\widetilde{X} = W/2$, $\boldsymbol{D} = E(\widetilde{X}^2) = 1/12$, and $\boldsymbol{A} = E(X^2) = 1/6$. In order to generate heterogenous data, we let $\beta_0^{(j)} = j$, for the jth sub-population, $j = 1, \ldots, s$, with $d = 1$.

In order to g_1 and g_2 using polynomial splines, we consider cubic splines ($\varrho = 3$) and equal-spaced knots. We estimate the unknown error variance σ^2 using $\overline{\sigma}^2 = \sum_{j=1}^{s}(\widehat{\sigma}^{(j)})^2/s$, where

$$
(\widehat{\sigma}^{(j)})^2 = \frac{1}{n - d - K(J_N + \varrho)} \sum_{i \in G_j} \left[Y_i - X_i\widehat{\beta}^{(j)} - \widehat{g}^{(j)}(\boldsymbol{Z}_i)\right]^2.
$$

We set the massive sample size N as $2^{11}, 2^{12}, 2^{13}$, or 2^{14}. We set the number of sub-samples s as $N^{1-\gamma}$, where $\gamma = \max(0.4, 2q), \ldots, 0.9, 1$. We set the minimal value of γ as $\max(0.4, 2q)$ to ensure that $J_N^2 = O(N^{2q}) \ll n = O(N^\gamma)$. For each setting, we run 200 repetitions.

First, we evaluate the performance of the aggregated estimator, $\overline{g}$, as an estimator for g. We compute the root mean-square-error (RMSE) of $\overline{g}$, under different choices of J_N and s, and different settings of N. The results are summarized in Figure 1. The condition that $J_N^2 \ll n$, which is needed in all the theorems, implies that the larger number of knots we take and the shorter range of s we should consider. In Figure 1, for each selection of the number of knots, we see that the performance of $\overline{g}$ is good and stable during a wide range of s. We also see that the RMSE of $\overline{g}$ deteriorates quickly when $\log(s)/\log(N)$ is approaching $1 - 2q, q \approx \log_N(J_N)$. For example, using 5 knots, $N = 2^{11}, q = \log_N(5) \approx 0.21$ and then $1 - 2q \approx 0.42$; therefore, from the second figure in the bottom row of Figure 1, we see that corresponding RMSE increases a lot when the ratio approaches 0.5. In summary, from Figure 1, we see there is a clear boundary of $\log(s)/\log(N)$: with this boundary,
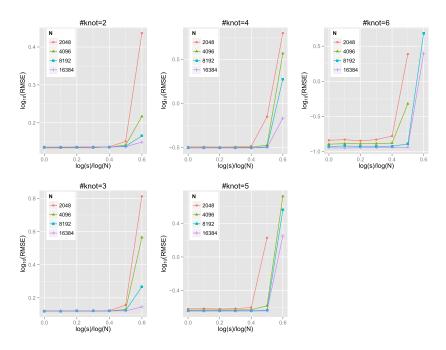
FIG 1. *Root mean-square-errors of the aggregated estimator, $\overline{g}$, under different settings of the number of knots, the number of sub-samples, and the sample size.*

the performance of $\overline{g}$ is very good, while beyond this boundary, the performance is very bad. These findings confirm the theoretical results presented in Theorem 2.2.

Second, we evaluate the performance of the proposed estimators, $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\breve{\boldsymbol{\beta}}^{(j)}$, for estimating $\beta_0^{(j)}$. We consider 95% confidence intervals based on $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\breve{\boldsymbol{\beta}}^{(j)}$ respectively as follows:

$$\mathrm{CI}_1 = \left[\widehat{\boldsymbol{\beta}}^{(j)} \pm \frac{1.96\overline{\sigma}}{\sqrt{n}}\boldsymbol{D}^{-1/2}\right] \quad \text{and} \quad \mathrm{CI}_2 = \left[\breve{\boldsymbol{\beta}}^{(j)} \pm \frac{1.96\overline{\sigma}}{\sqrt{n}}\boldsymbol{A}^{-1/2}\right].$$

For simplicity, we summarize results for the first sub-population in Figures 2–4, where both the coverage probabilities and the interval lengths are displayed, with the results of $\widehat{\beta}^{(1)}$ in solid line with circle and those of $\breve{\beta}^{(1)}$ in dashed line with triangle.

From Figure 2 where $N = 2^{11}$ and Figure 3 where $N = 2^{14}$, we see that within a proper range of s, CI_1 and CI_2 have similar coverage probabilities. We also see that on average, the interval length of CI_2 is shorter than that CI_1. This finding confirm that the asymptotic variance derived in Theorem 2.3 is smaller than that in Theorem 2.1. However, the coverage probability of CI_2 is valid for a shorter range of $\log(s)/\log(N)$, in contrast with that of CI_1. This is finding is consistent with that there are more conditions in Theorem 2.3 than in Theorem 2.1.
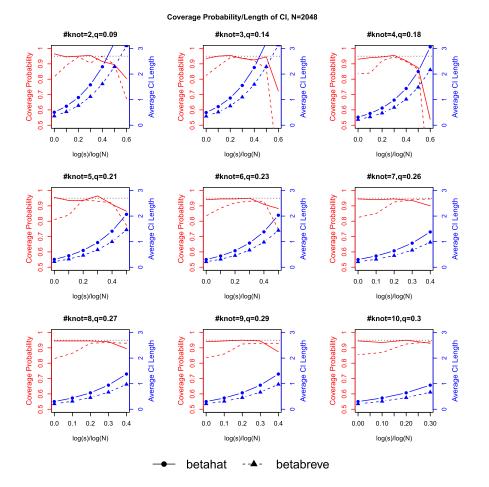
FIG 2. *Coverage probabilities and interval lengths of 95% confidence intervals, CI₁ and CI₂, under different settings of the number of knots and the number of sub-samples, with $N = 2^{11}$.*

To visualize the performance of CI$_2$ more clearly, in Figure 4 we display the coverage probability of CI$_2$ in more detail for different settings of s and N, given different numbers of knots. From Figure 4, we can see that, given the number of knots, a larger N implies a wider valid range for s to achieve a good coverage; given N, a larger number of knots implies a smaller transition point for s.

Third, we evaluate the heterogeneity tests using the following Wald test statistics constructed based on Theorem 3.1:

$$\Psi_1 = I\left\{ \boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big) \notin \sqrt{\frac{2}{n}}\,\overline{\sigma}(\boldsymbol{Q}\widehat{\boldsymbol{D}}^{-1}\boldsymbol{Q}^{\mathrm{T}})^{1/2}C_{\alpha/2} \right\},$$
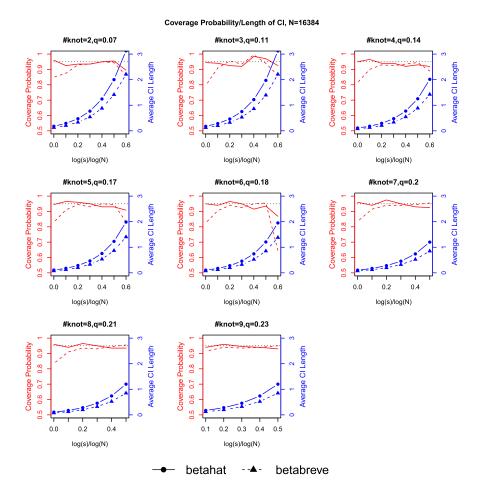
$$\Psi_2 = I\left\{ \boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \breve{\boldsymbol{\beta}}^{(j_2)}\big) \notin \sqrt{\frac{2}{n}}\,\overline{\sigma}(\boldsymbol{Q}\widehat{\boldsymbol{A}}^{-1}\boldsymbol{Q}^{\mathrm{T}})^{1/2}C_{\alpha/2} \right\},$$

Fig 3. *Coverage probabilities and interval lengths of 95% confidence intervals, CI_1 and CI_2, under different settings of the number of knots and the number of sub-samples, with $N = 2^{14}$.*

where $C_{\alpha/2}$ is the upper $\alpha/2$ quantile of a standard normal distribution, and $\widehat{D}$ and $\widehat{A}$ are the sample estimators of D and A, respectively. The results are summarized in Figure 5, where Ψ_1 and Ψ_2 are compared in terms of Type-I error and power, under different settings of s and N. From Panel (a) of Figure 5, we see that both Ψ_1 and Ψ_2 have appropriate type-I error within a wide range of s, but they have inflated type-I error after s passes a transition point. Panels (b)–(d) compare the testing powers under three different alternative hypotheses: $H_a : \beta_0^{(j_1)} - \beta_0^{(j_2)} = \Delta$, where $\Delta = 0.5, 1$ and 1.5, respectively. We see that the power increases as N increase and Δ increases. We also see the power of Ψ_2 is larger than that of Ψ_1 across different settings. These findings confirm the asymptotic results stated in Theorem 3.1.
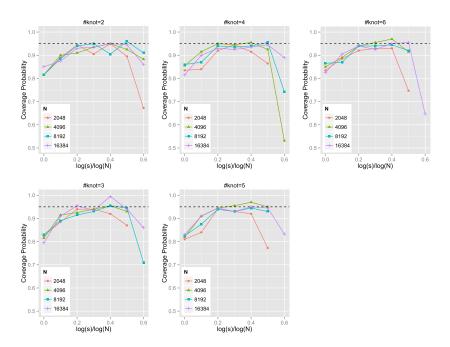
FIG 4. *Coverage probabilities of 95% CI$_2$ confidence intervals under different settings of the number of knots, the number of sub-samples, and the sample size.*

Fourth, we evaluate the more general heterogeneity tests involving all sub-populations using the test statistics constructed based on Theorem 3.2 and Remark 6. The null and alternative hypotheses are formulated as follows:

$$H_0: \ \beta_0^{(1)} = \cdots = \beta_0^{(s)} \quad \text{vs} \quad H_a: \beta_0^{(1)} = \beta_0^{(j)} + \Delta \quad j = 2, \ldots, s,$$

where $\Delta = 0, 0.4, 0.6$ and 1. Note that here $\boldsymbol{Q} = 1$. In the simulations, we set $\beta_0^{(j)} = 1$ for the null hypothesis, and use 500 bootstrapping samples to obtain the upper $\alpha = 0.05$ quantile for $W'_{\mathcal{S}}$. The results are summarized in Figure 6, where the y-axis shows the probability $P(T'_{\mathcal{S}} > c'S)$, under different settings of s and N. From Panel $\Delta = 0$ of Figure 6, we see that the proposed test statistic $T'_{\mathcal{S}}$ has appropriate type-I error around the nominal level $\alpha = 0.05$ within a wide range of s, but they have inflated type-I error after s passes a transition point. Panels (b)–(d) compare the testing powers under three different alternative hypotheses: $H_a: \boldsymbol{\beta}_0^{(1)} - \boldsymbol{\beta}_0^{(j)} = \Delta$, where $2 \leq j \leq s, \Delta = 0.4, 0.6$ and 1, respectively. We see that the power increases as N increase and Δ increases. These findings confirm the asymptotic results stated in Theorem 3.2.

Finally, we evaluate the homogeneity tests for one non-linear function using the test statistic constructed based on Theorem 3.3. The null and alternative
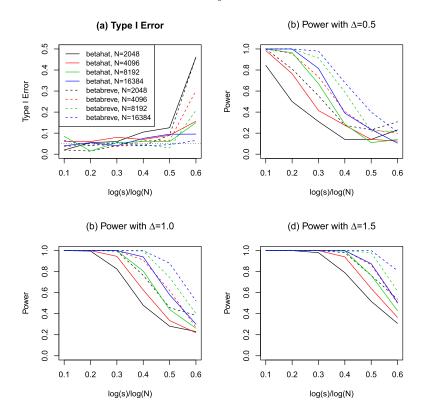
FIG 5. *Type-I error and power of tests* Ψ_1 *and* Ψ_2 *under different settings of the number of sub-samples and the sample size, using 4 knots.*

hypotheses are formulated as follows:

$$H_0 : g_{01}^{(1)}(Z_1) = \cdots = g_{01}^{(s)}(Z_1) \quad \text{vs}$$
$$H_a : g_{01}^{(1)}(Z_1) = g_{01}^{(j)}(Z_1) + \Delta(Z_1 + 1), \quad j \geq 2,$$

where $\Delta = 0, 0.5, 1.0$ and 1.5. In the simulations, we set $g_{01}^{(j)}(Z_1) = 5\sin\{2\pi(Z_1 + 1)\}, j \geq 2$ as used previously for the null hypothesis. The results are summarized in Figure 7, where the y-axis shows the probability $P(-\frac{1}{3\sigma^2} n \cdot \text{LRT}_{nk}^s > \chi_{u_N, 1-\alpha}^2)$ under different settings of s and N and $\chi_{u_N, 1-\alpha}^2$ is the $(1-\alpha)$-th quantile of $\chi_{u_N}^2$. From Panel $\Delta = 0$ of Figure 7, we see that the proposed test statistic $-\frac{1}{3\sigma^2} n \cdot \text{LRT}_{nk}^s$ has appropriate type-I error around the nominal level $\alpha = 0.05$ if s is not too large nor too small $(\log(s)/\log(N) > q/2)$, but they have inflated type-I error after s passes a transition point. As we use a fixed number of knots, the transition point shifts to the right as N increases so that $J_N \ll n^{1/2}$. Panels (b)–(d) compare the testing powers under three different alternative hypotheses with $\Delta = 0.5, 1.0$ and 1.5, respectively. We see that the power increases as N increase and Δ increases. Similar with observations in [12], the powers for
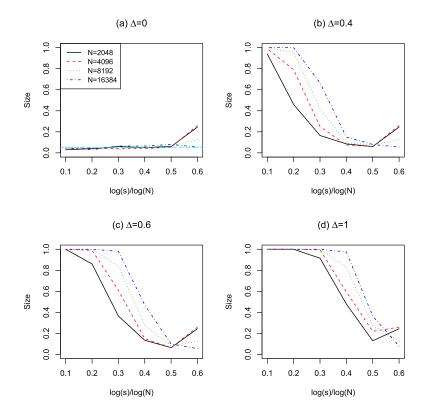
FIG 6. *Probability $P(T'_\mathcal{S} > c'_\mathcal{S})$ under different settings of the number of sub-samples and the sample size, using 4 knots.*

$\Delta > 0$ return back to 1 when s is very large. This can still be explained by highly deviated limiting distribution of the test statistic from both null and alternative hypotheses. Therefore, the homogeneity test does not perform well for very large s.

5. Real data application

We apply the proposed divide-and-conquer strategy for APLMs to the Medicare Provider Utilization and Payment Data (the Physician and Other Supplier Public Use File), with information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. This dataset was prepared by the Centers for Medicare & Medicaid Services (CMS), as part of the Obama Administration efforts to make our healthcare system more transparent, affordable, and accountable. We downloaded the dataset "Medicare Physician and Other Supplier Data CY 2014" from www.CMS.gov with more than nine million records for health care providers from the U.S. or U.S. possessions. We focus on the subset consisting of 50 U.S. states and the District of Columbia (DC), which account for the majority part of the whole dataset.
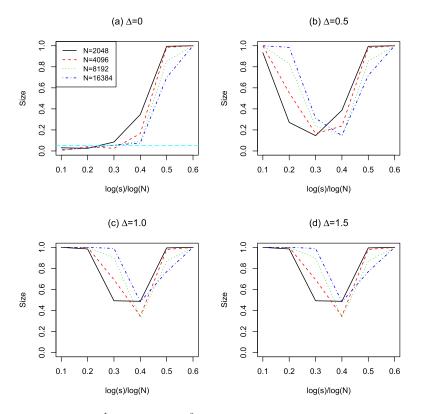
Fig 7. *Probability* $P(-\frac{1}{3\sigma^2} n \cdot \mathrm{LRT}^s_{nk} > \chi^2_{u_N, 1-\alpha})$ *under different settings of the number of sub-samples and the sample size, using 4 knots, where* $\chi^2_{u_N, 1-\alpha}$ *is the* $(1-\alpha)$-*th quantile of* $\chi^2_{u_N}$.

Our goal is to model the outcome variable "average_Medicare_standardized_amt" (average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied) on other covariates, including gender or entity of provider, provider type, Medicare participation status, place of service, HCPCS drug indicator, number of distinct Medicare beneficiaries ("bene_unique_cnt"), number of services provided ("bene_day_srvc_cnt"), and number of distinct Medicare beneficiary/per day services ("line_srvc_cnt"). Detailed explanations of these variables can be found in the official website www.CMS.gov. All covariates except the last three are categorical variables, and particularly the variable for provider type has 91 categories. Because those three quantitative variables are all count data, we take the $\log_{10}$-transformation and rescale each of them to the range $[-1, 1]$ by using the formula $(\boldsymbol{Z} - \min \boldsymbol{Z})/(\max \boldsymbol{Z} - \min \boldsymbol{Z}) \times 2 - 1$. Also, we apply the $\log_{10}$-transformation to the outcome variable, which is skewed to the right. By excluding those records with value 0 for quantitative variables and choosing records with overlapping ranges for last three variables across states, the working dataset has 9,263,068 records,

and the corresponding file size is greater than 2GB. It is hard to apply any complicated model fitting with iterative algorithms on a single PC with limited memory.

Therefore, we turn to the developed divide-and-conquer strategy. It is natural to split the data by location, such as states or counties. According to our theoretical results, the number of sub-populations cannot be too large. The number of counties is more than 3,000 in U.S., while $\sqrt{9,263,068} \approx 3044$. Thus, we split the whole dataset by states and DC, resulting in 51 sub-populations. The number of records for each sub-population varies from 14,809 (Alaska) to 719,970 (California), and the median number is 128,069. It is reasonable to hypothesize that those categorical covariates are heterogeneous because their effects on the average amount that Medicare paid may vary across states. On the other hand, the outcome variable is the standardized payment by removing geographic differences in payment rates for individual services, and all three quantitative covariates are numbers of services and beneficiaries. Then it is reasonable to assume the effects of quantitative covariates are homogeneous.

We choose B-splines with degree of 3 to approximate the non-parametric functions of those three quantitative covariates. Assumption (A4) requires that the number of internal knots should be much smaller than $N^{\frac{1}{4}} \approx 55$. Additionally, we expect these curves are smooth. Thus, we set the number of internal knots as 5. Noting that the sizes of sub-populations are different, rather than a simple average to obtain the aggregated curves, a weighted average is employed by using weights $n_j / \sum_{j=1}^{s} n_j$, where n_j is the size of the jth sub-sample.

Since the effects of those categorical covariates are allowed to be heterogeneous, we use box-plots to summarize the variabilities of their estimates across 51 sub-populations. From Figure 8, which displays the extent of heterogeneity, we can see that only the effect of male versus female has small degree of heterogeneity around 0, and all the other estimates have substantial variabilities. We further test the heterogeneity of the gender effects across 51 sub-populations via the testing procedure proposed in Section 3.1. The bootstrapped (based on 500 bootstrapping samples) critical value is $c'_{\mathcal{S}} = 2.69$ under $\alpha = 1\%$, while the test statistic $W'_{\mathcal{S}} = 10.09$ (p-value is 0). Although the range of this effect seems is small, with such a large sample size, we can easily detect a small heterogeneity across sub-populations. In summary, it implies that the effects of categorical covariates on the average amount that Medicare paid vary a lot across states.

Figure 9 presents the non-parametric estimates of the effects of those three quantitative covariates. The largest value of each quantitative covariate is different across states, so we only plot aggregated curves on the common range from -1 to 0. From panels (a)–(c) of the figure, for each covariate, we can see estimated curves from 51 sub-samples (dashed lines in black color) are almost parallel to each other within a narrow band, while the aggregated curve (solid line in red color) is right in the middle of those sub-sample specific curves. In addition, we are interested in the validity of the aggregation for commonality, noting that the sub-popualtion curves at boundaries vary more than they vary around the middle parts. Then we test the homogeneity of the non-linear com-
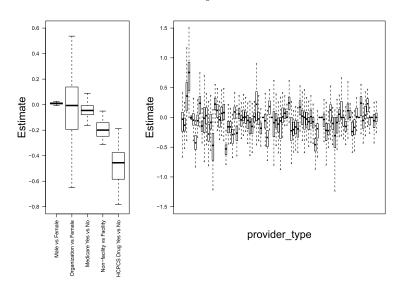
FIG 8. *Box-plots of heterogeneous parameters without aggregation of commonality.across 50 states and the DC: the left panel shows estimates of gender/entity, Medicare participation status, place of service and HCPCS drug status; the right panel shows estimates of 90 provider types versus the reference type.*
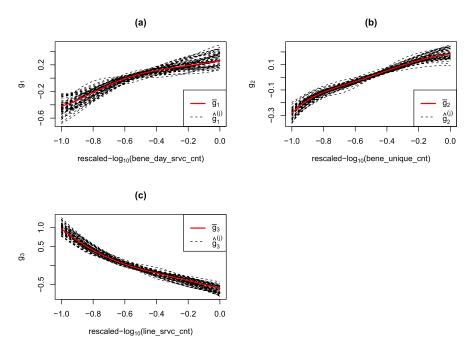

FIG 9. *Estimates of smooth functions based on each sub-population and the aggregation. (a): the estimated curves for "bene_unique_cnt"; (b): the estimated curves for "line_srvc_cnt"; (c): the estimated curves for "line_srvc_cnt".*

poents across 51 sub-populations on the range from -0.8 to -0.1 via the testing procedure proposed in Section 3.2. The test statistic $-\frac{1}{3\sigma^2}n \cdot \mathrm{LRT}^s_{nk} = 480.24$ and $u_N = 500$. Therefore, the resulting $p = 0.73$ means failing to reject the null hypothesis, and aggregation for commonality is a valid strategy.

6. Summary

In this paper, we develop a framework for additive partially linear models for massive heterogeneous data, using the divide-and-conquer strategy. As summarized in [19], the divide-and-conquer strategy is one of the three commonly used strategies for analyzing massive data, with the other two being the sub-sampling strategy and the sequential updating strategy. However, the sub-sampling and sequential updating strategies are only suitable for analyzing homogeneous massive data. We can combine the divide-and-conquer and sub-sampling strategies to analyze heterogeneous data, by dividing the data into homogeneous subgroups and then conducting sub-sampling within each subgroup. We can also combine the divide-and-conquer and sequential updating strategies to analyze heterogeneous data, by dividing the data into homogeneous subgroups and then conducting sequential updating within each subgroup.

The framework developed in this paper extends the partially linear framework proposed in [21]. Their partially linear framework considers only one common feature, using the smoothing-splines technique to fit the non-parametric function based on the general reproducing kernel Hilbert space (RKHS) theory [17]. Although the smoothing-splines technique and the RKHS theory have been well developed in the framework of generalized additive models [7], we find it very hard to extend them to our goal of analyzing massive data with multiple common features. Instead, we adopt polynomial splines for modeling the non-parametric effects of multiple common features simultaneously across all sub-populations while exploring heterogeneity of each sub-population. The proposed methods can be implemented easily and perform well in both simulation studies and the real data application.

Appendix

A.1. Technical lemmas for Section 2.2

Define the centered version of B-spline basis as

$$b^*_{m,k}(z_k) = b_{m,k}(z_k) - \frac{E[b_{m,k}]}{E[b_{1,k}]}b_{1,k}(z_k), \ \ k = 1,\ldots,K, m = -\varrho+1,\ldots,J_N,$$

and the standardized version of B-spline basis as

$$B_{m,k}(z_k) = \frac{b^*_{m,k}(z_k)}{\|b^*_{m,k}\|_{2k}}, \ \ m = -\varrho+1,\ldots,J_N, k = 1,\ldots,K.$$

Then the minimization problem (2.2) is equivalent to the following minimization problem:

$$(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}^{(j)}) = \underset{\boldsymbol{\beta} \in \mathcal{R}^d, \ \boldsymbol{\gamma} \in \mathcal{R}^{K(J_N + \varrho)}}{\operatorname{argmin}} \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} - \boldsymbol{B}(\boldsymbol{Z}_i)^{\mathrm{T}} \boldsymbol{\gamma} \right]^2,$$

where $\boldsymbol{B}(\boldsymbol{z}) = \{B_{m,k}(z_k), m = -\varrho + 1, \ldots, J_N, k = 1, \ldots, K\}^{\mathrm{T}}$. Here, to abuse the notation, we still use $\widehat{\gamma}^{(j)}$. Then $\widehat{g}^{(j)}(\boldsymbol{z}) = \widehat{\gamma}^{\mathrm{T}} \boldsymbol{B}(\boldsymbol{z})$ is a spline estimator of g_0 for the jth sub-population, and the centered spline estimators of a component function is

$$\widehat{g}_k^{(j)}(z_k) = \sum_{m=-\varrho+1}^{J_N} \widehat{\gamma}_{m,k} B_{m,k}(Z_k) - \frac{1}{n} \sum_{i \in G_j} \sum_{m=-\varrho+1}^{J_N} \widehat{\gamma}_{m,k} B_{m,k}(Z_{ik}).$$

In practice, basis $\{b_{m,k}, m = -\varrho + 1, \ldots, J_N, k = 1, \ldots, K\}$ is used for computational implementation, while $\{B_{m,k}\}$ is convenient for asymptotic analysis.

[4] showed that for any function $f \in \mathcal{H}$ and $N \geq 1$, there exists a function $\widetilde{f} \in \mathcal{S}_N$ such that $\|\widetilde{f} - f\|_\infty \leq Ch_N^p$. Thus, for g_0 satisfying Assumption (A1), there exists a $\widetilde{g}^{(j)}(\boldsymbol{z}) = \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{z})\widetilde{\gamma}^j \in \mathcal{G}_n^{(j)}$ s.t. $\|\widetilde{g}^{(j)} - g_0\|_\infty = O(h_N^p)$ and $\widetilde{g}^{(j)}(\boldsymbol{z})$ is the best least-squares projection of $g_0(\boldsymbol{z})$ into the space $\mathcal{G}_n^{(j)}$, implying

$$\langle \widetilde{g}^{(j)}(\boldsymbol{z}) - g_0(\boldsymbol{z}), \boldsymbol{B}(\boldsymbol{z}) \rangle_{jn} = 0, \quad j = 1, \ldots, s. \tag{A.1}$$

Define

$$\widetilde{\boldsymbol{\beta}}^{(j)} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2} \sum_{i \in G_j} \left[Y_i - \widetilde{g}^{(j)}(\boldsymbol{Z}_i) - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} \right]^2,$$

and let $m_{0i}^{(j)} \equiv m_0^{(j)}(\boldsymbol{T}_i) = g_0(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0^{(j)}$, $\widetilde{m}_0^{(j)}(\mathbf{t}) = \widetilde{g}^{(j)}(\boldsymbol{z}) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_0^{(j)}$, and $\widetilde{m}_{0i}^{(j)} \equiv \widetilde{m}_0^{(j)}(\boldsymbol{T}_i) = \widetilde{g}^{(j)}(\boldsymbol{Z}_i) + \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_0^{(j)}$.

Additionally, let $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\beta} \end{pmatrix}$, $\widehat{\boldsymbol{\theta}}^{(j)} = \begin{pmatrix} \widehat{\boldsymbol{\gamma}}^{(j)} \\ \widehat{\boldsymbol{\beta}}^{(j)} \end{pmatrix}$, $\widetilde{\boldsymbol{\theta}}^{(j)} = \begin{pmatrix} \widetilde{\boldsymbol{\gamma}}^{(j)} \\ \widetilde{\boldsymbol{\beta}}^{(j)} \end{pmatrix}$, $\widehat{l}_n^{(j)}(\boldsymbol{\theta}) = l_n^{(j)}(\boldsymbol{\gamma}, \boldsymbol{\beta})$, and

$$\widetilde{m}_i^{(j)} \equiv \widetilde{m}^{(j)}(\boldsymbol{T}_i) = \widetilde{g}^{(j)} + \boldsymbol{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}} = \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z}_i)\widetilde{\boldsymbol{\gamma}}^{(j)} + \boldsymbol{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}^{(j)}.$$

Define

$$V_n^{(j)} \doteq \frac{\partial^2 \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = \frac{1}{n} \sum_{i \in G_j} \left\{ \begin{array}{cc} (\boldsymbol{B}(\boldsymbol{Z}_i))^{\otimes 2} & \boldsymbol{B}(\boldsymbol{Z}_i)\boldsymbol{X}_i^{\mathrm{T}} \\ \boldsymbol{X}_i \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{Z}_i) & \boldsymbol{X}_i^{\otimes 2} \end{array} \right\}.$$

Lemma A.1. *Under Assumptions (A1)–(A4), for each sub-population j,*

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1}\boldsymbol{\Sigma}_1 \boldsymbol{A}^{-1}),$$

where $\boldsymbol{A} = E(\boldsymbol{X}^{\otimes 2})$ and $\boldsymbol{\Sigma}_1 = E(\varepsilon^2 \boldsymbol{X}^{\otimes 2})$.

Proof. Let $\widetilde{\boldsymbol{\delta}}^{(j)} = \sqrt{n}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})$. Then $\widetilde{\boldsymbol{\delta}}^{(j)}$ minimizes

$$\widetilde{l}_n^{(j)}(\boldsymbol{\delta}) = \frac{1}{2} \sum_{i \in G_j} \left[\left(Y_i - \widetilde{m}_{0i}^{(j)} - \frac{1}{\sqrt{n}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\delta} \right)^2 - (Y_i - \widetilde{m}_{0i})^2 \right].$$

Let $\boldsymbol{A}_n^{(j)} = \frac{1}{n} \sum_{i \in G_j} \boldsymbol{X}_i^{\otimes 2}$. By taking derivatives with respect to $\boldsymbol{\delta}$, we obtain

$$\frac{\partial \widetilde{l}_n^{(j)}(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \boldsymbol{A}_n^{(j)} \boldsymbol{\delta} - \frac{1}{\sqrt{n}} \sum_{i \in G_j} (Y_i - \widetilde{m}_{0i}^{(j)}) \boldsymbol{X}_i = \boldsymbol{0},$$

which implies

$$\widetilde{\boldsymbol{\delta}}^{(j)} = \frac{1}{\sqrt{n}} (\boldsymbol{A}_n^{(j)})^{-1} \sum_{i \in G_j} \varepsilon_i \boldsymbol{X}_i + \frac{1}{\sqrt{n}} (\boldsymbol{A}_n^{(j)})^{-1} \sum_{i \in G_j} \left(g_0(\boldsymbol{Z}_i) - \widetilde{g}^{(j)}(\boldsymbol{Z}_i) \right) \boldsymbol{X}_i.$$

With similar arguments with those of Lemma A.1 in [11] and the fact $\|\widetilde{g}^{(j)} - g_0\|_\infty = O(h_N^p)$, the lemma follows. $\qquad\square$

Lemma A.2. *Under Assumptions (A1)–(A4), if $J_N \ll \frac{n}{(\log n)^2}$, there exists a constant C such that $\sup_j \|(V_n^{(j)})^{-1}\|_2 \le C$, a.s.*

Proof. For each sub-population j, Lemma A.2 in [11] showed there exists a constant C such that $\|(V_n^{(j)})^{-1}\|_2 \le C$, a.s., if

$$\sup_{f_1, f_2 \in \mathcal{G}_n^{(j)}} \left| \frac{\langle f_1, f_2 \rangle_{jn} - \langle f_1, f_2 \rangle}{\|f_1\|_2 \|f_2\|_2} \right| = O\left(\frac{\log n}{(nh_N)^{1/2}} \right) = o(1), \quad a.s.,$$

by Lemma A.8 in [18]. Here constant C is taken to be large enough to ensure that the above result holds for all $j = 1, \cdots, s$. The condition $J_N \ll \frac{n}{(\log n)^2}$ implies $O\left(\log n/(nh_N)^{1/2} \right) = o(1)$. Therefore, the lemma is proved. $\qquad\square$

Lemma A.3. *Under Assumptions (A1)–(A4), for each sub-population j, we have*
$$\left\| \widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} \right\| = O_P\left(J_N^{1/2} n^{-1/2} + h_N^p \right).$$

Proof. It follows that

$$\left. \frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(j)}} - \left. \frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}^{(j)}} = \left. \frac{\partial^2 \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}^{(j)}} \left(\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} \right),$$

where $\overline{\boldsymbol{\theta}}^{(j)}$ is between $\widehat{\boldsymbol{\theta}}^{(j)}$ and $\widetilde{\boldsymbol{\theta}}^{(j)}$. Thus, we have

$$\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} = - \left(\left. \frac{\partial^2 \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \right|_{\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}^{(j)}} \right)^{-1} \left. \frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}^{(j)}}.$$

We can write

$$\frac{1}{n}\frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}^{(j)}} = -\frac{1}{n}\sum_{i\in G_j}(Y_i - m_{0i}^{(j)})\begin{pmatrix}\boldsymbol{B}(\boldsymbol{Z}_i)\\ \boldsymbol{X}_i\end{pmatrix}$$
$$+\frac{1}{n}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\begin{pmatrix}\boldsymbol{B}(\boldsymbol{Z}_i)\\ \boldsymbol{X}_i\end{pmatrix}$$
$$+\frac{1}{n}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\begin{pmatrix}\boldsymbol{B}(\boldsymbol{Z}_i)\\ \boldsymbol{X}_i\end{pmatrix}.$$

First, by (A.1), we have $\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{B}(\boldsymbol{Z}_i) = \boldsymbol{0}$. With similar arguments with those of Lemma A.3 in [11], we have

$$\left\|\frac{1}{n}\sum_{i\in G_j}(Y_i - m_{0i}^{(j)})\boldsymbol{B}(\boldsymbol{Z}_i)\right\| = O_P\left(J_N^{1/2}n^{-1/2}\right),$$

$$\left\|\frac{1}{n}\sum_{i\in G_j}(Y_i - m_{0i}^{(j)})\boldsymbol{X}_i\right\| = O_P\left(n^{-1/2}\right),$$

$$\left\|\frac{1}{n}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{X}_i\right\| = O_P\left(h_N^p\right),$$

$$\left\|\frac{1}{n}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{B}(\boldsymbol{Z}_i)\right\| = o_P\left(J_N^{1/2}n^{-1/2}\right),$$

$$\left\|\frac{1}{n}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{X}_i\right\| = o_P\left(n^{-1/2}\right).$$

Therefore, by Lemma A.2, we have

$$\|\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)}\| \le \|(V_n^{(j)})^{-1}\|_2\left\|\frac{1}{n}\frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}^{(j)}}\right\| = O_P\left(J_N^{1/2}n^{-1/2} + h_N^p\right). \quad \square$$

Lemma A.4. *Under Assumptions (A1)–(A4), for each sub-population j, if $J_N \ll \frac{n}{(\log n)^2}$, we have*

$$\left\|\widehat{g}^{(j)} - g_0\right\|_2 = O_P\left(J_N^{1/2}n^{-1/2} + h_N^p\right),$$
$$\left\|\widehat{g}^{(j)} - g_0\right\|_{jn} = O_P\left(J_N^{1/2}n^{-1/2} + h_N^p\right),$$
$$\left\|\widehat{g}_k^{(j)} - g_{0k}\right\|_{2k} = O_P\left(J_N^{1/2}n^{-1/2} + h_N^p\right),$$
$$\left\|\widehat{g}_k^{(j)} - g_{0k}\right\|_{jnk} = O_P\left(J_N^{1/2}n^{-1/2} + h_N^p\right),$$

where $k = 1, \ldots, K$.

Proof. The proof is similar with that of Lemma A.4 in [11] by applying Lemmas A.2 and A.3 and noting that

$$\sup_{f \in \mathcal{S}_N} \frac{\|f\|_{jnk}}{\|f\|_{2k}} = O_P\left(\frac{\log n}{(nh_N)^{1/2}}\right) = o_P(1), \quad k = 1, \ldots, K,$$

which is implied by Lemma A.8 in [18] under condition $J_N \ll n/(\log n)^2$. $\quad\square$

Lemma A.5. *Under Assumptions (A1)–(A4), for each sub-population j, if $n \gg J_N^2$, we have*

$$\frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i \Gamma(\boldsymbol{Z}_i)^{\mathrm{T}} (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) = o_P(n^{-1/2}),$$

$$\frac{1}{n} \sum_{i \in G_j} (\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\widetilde{\boldsymbol{X}}_i = o_P(n^{-1/2}),$$

Proof. The proof is similar with that of Lemma A.5 in [11] by making following revisions. We only show the second equality, and the first one can be proved similarly.

Let $w_1(\boldsymbol{Z}, g) = g(\boldsymbol{Z})\widetilde{\boldsymbol{X}}$, and it follows

$$E\big\|w_1(\boldsymbol{Z}, \widehat{g}^{(j)}) - w_1(\boldsymbol{Z}, g_0)\big\|^2 = E\left\|(\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\widetilde{\boldsymbol{X}}_i\right\|^2 \leq O\big(E\|\widehat{g}^{(j)} - g_0\|_2^2\big).$$

By Lemma A.2 of [8], the logarithm of the ε-bracketing number of the class of functions $\mathcal{A}_1^{(j)}(\delta) = \{w_1(\cdot, \widehat{g}) - w_1(\cdot, g_0) : \widehat{g} \in \mathcal{G}_n^{(j)}, \|\widehat{g} - g_0\|_2 \leq \delta\}$ is $c\{(J_N - \varrho)\log(\delta/\varepsilon) + \log(\delta^{-1})\}$. Thus, the corresponding entropy integral $J_{[]}(\delta, \mathcal{A}_1^{(j)}(\delta), \|\cdot\|_2) \leq c\delta\{(J_N - \varrho)^{1/2} + \log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of [15] and Lemma A.4, $\|\widehat{g}^{(j)} - g_0\|_\infty \leq cJ_N^{1/2}\|\widehat{g}^{(j)} - g_0\|_2 = O_P(J_N n^{-1/2} + J_N^{1/2} h_N^p)$. Let $r_{n,N}^{-1} = J_N^{1/2} n^{-1/2} + h_N^p$, then

$$E\left|\frac{1}{n} \sum_{i \in G_j} \left\{\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\right\}\widetilde{\boldsymbol{X}}_i - E\left\{\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\right\}\widetilde{\boldsymbol{X}}_i\right|$$

$$\leq n^{-1/2} C r_{n,N}^{-1} \left\{(J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N})\right\}$$

$$\times \left[1 + \frac{c r_{n,N}^{-1}\left\{(J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N})\right\}}{r_{n,N}^{-2}\sqrt{n}} C_0\right]$$

$$\leq O(1) n^{-1/2} C r_{n,N}^{-1} \left\{(J_N + \varrho)^{1/2} + \log^{1/2}(r_{n,N})\right\},$$

where the second inequality is based on the fact $r_{n,N} J_N^{1/2}/\sqrt{n} = O(1)$.

Under condition that $n \gg J_N^2$, we have $J_N/\sqrt{n} = o(1)$, implying $J_N n^{-1/2} + J_N^{1/2} h_N^p = o(1)$, and then $r_{n,N}^{-1} J_N^{1/2} = o(1)$. Therefore, the above term is bounded by $o(n^{-1/2})$. $\quad\square$

A.2. Technical lemmas for Section 2.3

Let $\widetilde{g} = \frac{1}{s}\sum_{j=1}^{s}\widetilde{g}^{(j)}$. In order to ensure that $\widetilde{g} \in \mathcal{G}_N$, we re-center the individual estimator $\widetilde{g}^{(j)}(\boldsymbol{z})$ via

$$\widetilde{g}^{(j)}(\boldsymbol{z}) = \sum_{k=1}^{K}\sum_{m=-\varrho}^{J_N}\widetilde{\gamma}_{m,k}b_{m,k}(z_k) - \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{m=-\varrho}^{J_N}\widetilde{\gamma}_{m,k}b_{m,k}(z_{ik}).$$

To abuse the notation, we still denote the centered estimator as $\widetilde{g}^{(j)}(\boldsymbol{z})$. Lemma A.2 shows $\|(V_n^{(j)})^{-1}\|_2 \leq C$ a.s., $j = 1,\ldots,s$, if $n \gg J_N^2$. This property plays a key role in all following proofs. Define $\boldsymbol{u}_i^{(j)} = (V_n^{(j)})^{-1}\begin{pmatrix}\boldsymbol{B}(\boldsymbol{Z}_i)\\\boldsymbol{X}_i\end{pmatrix} = \{u_{im}^{(j)}, m = 1,\ldots,K(J_N + \varrho) + d\}^{\mathrm{T}}$, and it follows $\sum_{i \in G_j}(\boldsymbol{u}_i^{(j)})^{\otimes 2} = n(V_n^{(j)})^{-1}$. Let $\boldsymbol{e}_m$ denote a $(K(J_N + \varrho) + d)$-dim vector with its mth entry as 1 and 0 otherwise, and thus $u_{im}^{(j)} = \boldsymbol{e}_m^{\mathrm{T}}\boldsymbol{u}_i^{(j)}$.

Lemma A.6. *Under Assumptions (A1)–(A4), if $n \gg J_N^2$, we have*

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i\boldsymbol{u}_i^{(j)}\right\| = O_P(J_N^{1/2}N^{-1/2}).$$

Proof. If follows

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i\boldsymbol{u}_i^{(j)}\right\|^2 = \frac{1}{N^2}\sum_{m=1}^{K(J_N+\varrho)+d}\left\{\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i u_{im}^{(j)}\right\}^2.$$

Observing that

$$\frac{1}{N}E\left\{\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i u_{im}^{(j)}\right\}^2 = \frac{1}{N^2}E\left[\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i^2(u_{im}^{(j)})^2\right] = \frac{\sigma^2}{N^2}E\left[\sum_{j=1}^{s}\sum_{i \in G_j}(u_{im}^{(j)})^2\right]$$

$$= \frac{\sigma^2}{N^2}E\left[\sum_{j=1}^{s}n\boldsymbol{e}_m^{\mathrm{T}}(V_n^{(j)})^{-1}\boldsymbol{e}_m\right] \leq \frac{NC\sigma^2}{N^2},$$

where the last inequality is due to the fact $\boldsymbol{e}_m^{\mathrm{T}}(V_n^{(j)})^{-1}\boldsymbol{e}_m \leq \|(V_n^{(j)})^{-1}\|_2^2\|\boldsymbol{e}_m\|^2 \leq C$ a.s.. Thus

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i \in G_j}\varepsilon_i\boldsymbol{u}_i^{(j)}\right\| = O_P(J_N^{1/2}N^{-1/2}). \qquad \square$$

Lemma A.7. *Under Assumptions (A1)–(A4), if $n \gg J_N^2$, we have*

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i \in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{u}_i^{(j)}\right\| = O_P(J_N^{1/2}h_N^p).$$

Proof. Note that

$$\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{u}_i^{(j)}$$

$$=\frac{1}{N}\sum_{j=1}^{s}(V_n^{(j)})^{-1}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\begin{pmatrix}\boldsymbol{B}(\boldsymbol{Z}_i)\\\boldsymbol{X}_i\end{pmatrix}$$

$$=\frac{1}{N}\sum_{j=1}^{s}(V_n^{(j)})^{-1}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\begin{pmatrix}\boldsymbol{0}\\\boldsymbol{X}_i\end{pmatrix},$$

where the last equality follows from (A.1).

Therefore, it follows

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{u}_i^{(j)}\right\|$$

$$\leq\frac{1}{N}\sum_{j=1}^{s}\left\|(V_n^{(j)})^{-1}\right\|_2\sum_{i\in G_j}\left\|(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\begin{pmatrix}\boldsymbol{0}\\\boldsymbol{X}_i\end{pmatrix}\right\|$$

$$\leq\frac{C}{N}\sum_{j=1}^{s}\sum_{i\in G_j}\left\|(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\right\|_\infty\left\|\begin{pmatrix}\boldsymbol{0}\\\boldsymbol{X}_i\end{pmatrix}\right\|_1 = O_P(h_N^p). \qquad \Box$$

Lemma A.8. *Under Assumptions (A1)–(A4), if $n \gg J_N^2$, we have*

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{u}_i^{(j)}\right\| = o_P\left(J_N^{1/2}N^{-1/2}\right).$$

Proof. It follows

$$\left\|\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{u}_i^{(j)}\right\|^2 = \frac{1}{N^2}\sum_{m=1}^{K(J_N+\varrho)+d}\left\{\sum_{j=1}^{s}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})u_{im}^{(j)}\right\}^2.$$

The proof of Lemma A.1 shows

$$\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} = \frac{1}{n}(\boldsymbol{A}_n^{(j)})^{-1}\sum_{i\in G_j}\varepsilon_i\boldsymbol{X}_i + \frac{1}{n}(\boldsymbol{A}_n^{(j)})^{-1}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{X}_i.$$

Then

$$\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})u_{im}^{(j)}$$

$$= \frac{1}{n}\sum_{i_2\in G_j}\varepsilon_{i_2}\boldsymbol{X}_{i_2}^{\mathrm{T}}\left(\sum_{i_1\in G_j}u_{i_1m}^{(j)}(\boldsymbol{A}_n^{(j)})^{-1}\boldsymbol{X}_{i_1}\right)$$

$$+ \frac{1}{n} \sum_{i_2 \in G_j} (\widetilde{g}^{(j)}(\boldsymbol{Z}_{i_2}) - g_0(\boldsymbol{Z}_{i_2})) \boldsymbol{X}_{i_2}^{\mathrm{T}} \left(\sum_{i_1 \in G_j} u_{i_1 m}^{(j)} (\boldsymbol{A}_n^{(j)})^{-1} \boldsymbol{X}_{i_1} \right)$$

$$= \sum_{i \in G_j} \varepsilon_i \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} + \sum_{i \in G_j} (\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)},$$

where $\boldsymbol{v}_m^{(j)} = \sum_{i \in G_j} u_{im}^{(j)} (\boldsymbol{A}_n^{(j)})^{-1} \boldsymbol{X}_i$. Thus, we have

$$\frac{1}{N^2} E \left\{ \sum_{j=1}^{s} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) u_{im}^{(j)} \right\}^2$$

$$= \frac{1}{N^2} E \left\{ \sum_{j=1}^{s} \sum_{i \in G_j} \varepsilon_i \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} + \sum_{j=1}^{s} \sum_{i \in G_j} (\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2$$

$$= \frac{1}{N^2} E \left\{ \sum_{j=1}^{s} \sum_{i \in G_j} \varepsilon_i \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2 + \frac{1}{N^2} E \left\{ \sum_{j=1}^{s} \sum_{i \in G_j} (\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)) \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2$$

$$\leq \frac{\sigma^2}{N^2} E \sum_{j=1}^{s} \sum_{i \in G_j} \left\{ \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2 + \frac{1}{N} \|\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)\|_\infty^2 E \sum_{j=1}^{s} \sum_{i \in G_j} \left\{ \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2.$$

Observing that

$$\sum_{i \in G_j} \left\{ \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right\}^2 = \sum_{i \in G_j} (\boldsymbol{v}_m^{(j)})^{\mathrm{T}} \boldsymbol{X}_i \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} = n (\boldsymbol{v}_m^{(j)})^{\mathrm{T}} \boldsymbol{A}_n^{(j)} \boldsymbol{v}_m^{(j)}$$

$$= \frac{1}{n} \left(\sum_{i \in G_j} u_{im}^{(j)} \boldsymbol{X}_i^{\mathrm{T}} (\boldsymbol{A}_n^{(j)})^{-1} \right) \boldsymbol{A}_n^{(j)} \left(\sum_{i \in G_j} u_{im}^{(j)} \boldsymbol{A}_n^{(j)})^{-1} \boldsymbol{X}_i \right)$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i \in G_j} u_{im}^{(j)} \boldsymbol{X}_i \right)^{\mathrm{T}} (\boldsymbol{A}_n^{(j)})^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i \in G_j} u_{im}^{(j)} \boldsymbol{X}_i \right)$$

Based on the Central Limit Theorem and Slutsky's Theorem, it follows

$$E \left\{ \sum_{i \in G_j} \left(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{v}_m^{(j)} \right)^2 \right\} = O(1).$$

Therefore,

$$\frac{1}{N^2} E \left\{ \sum_{j=1}^{s} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) u_{im}^{(j)} \right\}^2 = O \left(\frac{s}{N^2} + \frac{s h_N^{2p}}{N} \right) = o \left(\frac{1}{N} \right),$$

noting that $h_N^{2p} \ll N^{-1}$. Then, we have

$$\left\| \frac{1}{N} \sum_{j=1}^{s} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \boldsymbol{u}_i^{(j)} \right\| = o_P \left(J_N^{1/2} N^{-1/2} \right). \qquad \square$$

A.3. Proofs of theorems

Proof of Theorem 2.1. The results about $\widehat{g}^{(j)}$ are implied by Lemma A.4 directly. We only need to prove the stated result about $\widehat{\boldsymbol{\beta}}^{(j)}$. Note that the ondition that $J_N^2 \ll n$ implies that $J_N \ll \frac{n}{(\log n)^2}$. Also the condition that $J_N \gg n^{1/(2p)}$ implies that $h_N^p = O(J_N^{-p}) \ll n^{-1/2}$. Therefore, the stated result about $\widehat{\boldsymbol{\beta}}^{(j)}$ can be showed by Lemmas A.1–A.5, following the same argument of proving Theorem 1 in [11]. $\qquad \square$

Proof of Theorem 2.2. We first quantify $\|\widehat{g}^{(j)} - g_0\|_2$. Noting $\|\widetilde{g} - g_0\|_2 \leq \|\widetilde{g} - g_0\|_\infty = O(h_N^p)$ and

$$\frac{1}{s} \sum_{j=1}^{s} \left(\widehat{g}^{(j)}(\boldsymbol{z}) - \widetilde{g}^{(j)}(\boldsymbol{z}) \right) = \frac{1}{s} \boldsymbol{B}^{\mathrm{T}}(\boldsymbol{z}) \sum_{j=1}^{s} (\widehat{\boldsymbol{\gamma}}^{(j)} - \widetilde{\boldsymbol{\gamma}}^{(j)}),$$

we have

$$\left\| \frac{1}{s} \sum_{j=1}^{s} \left(\widehat{g}^{(j)}(\boldsymbol{z}) - \widetilde{g}^{(j)}(\boldsymbol{z}) \right) \right\|_2^2$$

$$= \int_{[0,1]^K} \left[\frac{1}{s} \sum_{j=1}^{s} \left(\widehat{g}^{(j)}(\boldsymbol{z}) - \widetilde{g}^{(j)}(\boldsymbol{z}) \right) \right]^2 f(\boldsymbol{z}) d\boldsymbol{z}$$

$$= \frac{1}{s} \sum_{j=1}^{s} \left(\widehat{\boldsymbol{\gamma}}^{(j)} - \widetilde{\boldsymbol{\gamma}}^{(j)} \right)^{\mathrm{T}} \left[E\boldsymbol{B}(\boldsymbol{z})\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{z}) \right] \sum_{j=1}^{s} \left(\widehat{\boldsymbol{\gamma}}^{(j)} - \widetilde{\boldsymbol{\gamma}}^{(j)} \right)$$

$$\leq \frac{C}{s^2} \left\| \sum_{j=1}^{s} \left(\widehat{\boldsymbol{\gamma}}^{(j)} - \widetilde{\boldsymbol{\gamma}}^{(j)} \right) \right\|^2 \leq \frac{C}{s^2} \left\| \sum_{j=1}^{s} \left(\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} \right) \right\|^2.$$

Then we consider $\left\| \frac{1}{s} \sum_{j=1}^{s} \left(\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} \right) \right\|$. The proof of Lemma A.3 implies that

$$\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)} = (V_n^{(j)})^{-1} \frac{1}{n} \frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}^{(j)}},$$

where $\frac{1}{n} \frac{\partial \widehat{l}_n^{(j)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}^{(j)}}$ is equal to

$$\left\{ -\frac{1}{n} \sum_{i \in G_j} (Y_i - m_{0i}^{(j)}) + \frac{1}{n} \sum_{i \in G_j} (\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i)) + \frac{1}{n} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}} (\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \right\} \begin{pmatrix} \boldsymbol{B}(\boldsymbol{Z}_i) \\ \boldsymbol{X}_i \end{pmatrix}.$$

Then $\frac{1}{s}\sum_{j=1}^{s}\left(\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)}\right)$ is equal to

$$-\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}\varepsilon_i \boldsymbol{u}_i^{(j)} + \frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}(\widetilde{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i))\boldsymbol{u}_i^{(j)}$$
$$+ \frac{1}{N}\sum_{j=1}^{s}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{u}_i^{(j)}.$$

Therefore, combining Lemmas A.6-A.8, we have

$$\frac{1}{s}\left\|\sum_{j=1}^{s}\left(\widehat{\boldsymbol{\theta}}^{(j)} - \widetilde{\boldsymbol{\theta}}^{(j)}\right)\right\| = O_P\left(J_N^{1/2}N^{-1/2} + h_N^p\right).$$

and further we have,

$$\begin{aligned}
\|\overline{g} - g_0\|_2 &= \left\|\frac{1}{s}\sum_{j=1}^{s}\widehat{g}^{(j)} - \frac{1}{s}\sum_{j=1}^{s}\widetilde{g}^{(j)} + \widetilde{g} - g_0\right\|_2 \\
&\leq \left\|\frac{1}{s}\sum_{j=1}^{s}\left(\widehat{g}^{(j)} - \widetilde{g}^{(j)}\right)\right\|_2 + \|\widetilde{g} - g_0\|_2 \\
&= O_P\left(J_N^{1/2}N^{-1/2} + h_N^p\right).
\end{aligned}$$

Next we quantify $\|\overline{g} - g_0\|_N$. Using Lemma A.8 in [18], we have

$$C_N \equiv \sup_{f_1,f_2\in\mathcal{G}_N}\left|\frac{\langle f_1, f_2\rangle_N - \langle f_1, f_2\rangle}{\|f_1\|\|f_2\|}\right| = O\left(\frac{\log N}{(Nh_N)^{1/2}}\right), \quad a.s.$$

Therefore, noting $\overline{g}, \widetilde{g} \in \mathcal{G}_N$, we have

$$\|\overline{g} - g_0\|_N \leq \|\overline{g} - \widetilde{g}\|_N + \|\widetilde{g} - g_0\|_N = O_P\left(J_N^{1/2}N^{-1/2} + h_N^p\right). \qquad \square$$

Proof of Corollary 2.1. For homogenous massive data, $\widehat{\boldsymbol{\beta}}^{(j)}$, $j = 1,\ldots,s$, are i.i.d. random vectors. [10] showed that if $E(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0) = o(N^{-1/2})$, then $\overline{\boldsymbol{\beta}}$ is as efficient as $\widehat{\boldsymbol{\beta}}$, which is defined as via the following minimization using all N observations,

$$(\widehat{\boldsymbol{\beta}}, \widehat{g}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathcal{R}^d,\ g\in\mathcal{G}_N} \frac{1}{2}\sum_{j=1}^{s}\sum_{i\in G_j}\left[Y_i - \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta} - g(\boldsymbol{Z}_i)\right]^2.$$

[11] showed that

$$\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2\boldsymbol{D}^{-1}\right).$$

Therefore it suffices to show that $E(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0) = o(N^{-1/2})$. Following the proof of Theorem 2.1, we have

$$\left(\frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i \Gamma(\boldsymbol{Z}_i)^{\mathrm{T}} \right) (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)$$

$$= \frac{1}{n} \sum_{i \in G_j} \varepsilon_i \widetilde{\boldsymbol{X}}_i - \frac{1}{n} \sum_{i \in G_j} \left(\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \right) \widetilde{\boldsymbol{X}}_i,$$

and it follows

$$\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0 = \left(\frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i \Gamma(\boldsymbol{Z}_i)^{\mathrm{T}} \right)^{-1} \times$$

$$\left[\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \widetilde{\boldsymbol{X}}_i - \frac{1}{n} \sum_{i \in G_j} \left(\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \right) \widetilde{\boldsymbol{X}}_i \right].$$

Under Assumption A3 and the fact that $E(\phi(\boldsymbol{Z})\widetilde{\boldsymbol{X}}) = 0$ for any measurable function ϕ, we can show that

$$0 < c \leq E \left\| \frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i^{\otimes 2} + \frac{1}{n} \sum_{i \in G_j} \widetilde{\boldsymbol{X}}_i \Gamma(\boldsymbol{Z}_i)^{\mathrm{T}} \right\|_2 \leq C,$$

where c and C are some positive constants. Moreover, we have

$$E \left\{ \frac{1}{n} \sum_{i \in G_j} \left(\widehat{g}^{(j)}(\boldsymbol{Z}_i) - g_0(\boldsymbol{Z}_i) \right) \widetilde{\boldsymbol{X}}_i \right\} = E \left(\widehat{g}^{(j)}(\boldsymbol{Z}) - g_0(\boldsymbol{Z}) \right) \widetilde{\boldsymbol{X}}_i$$

$$= O(h_N^p) = o(N^{-1/2}).$$

Therefore, if $n \gg N^{1/2}$, by Cauchy-Schwarz inequality we have $E(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0) = o(N^{-1/2})$. $\square$

Proof of Theorem 2.3. The estimating equation is

$$\sum_{i \in G_j} \boldsymbol{X}_i \left\{ Y_i - \boldsymbol{X}_i^{\mathrm{T}} \breve{\boldsymbol{\beta}}^{(j)} - \overline{g}(\boldsymbol{Z}_i) \right\} = \mathbf{0}.$$

Since $Y_i = \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta}_0^{(j)} + g_0(\boldsymbol{Z}_i) + \varepsilon_i$, we have

$$\sqrt{n} \left(\breve{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} \right)$$

$$= n^{-1/2} \sum_{i \in G_j} \left(\boldsymbol{A}_n^{(j)} \right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i + n^{-1/2} \sum_{i \in G_j} \left(\boldsymbol{A}_n^{(j)} \right)^{-1} \boldsymbol{X}_i (g_0(\boldsymbol{Z}_i) - \overline{g}(\boldsymbol{Z}_i)).$$

Considering the first term on the right hand side of the above equation, we have

$$n^{-1/2} \sum_{i \in G_j} \left(A_n^{(j)} \right)^{-1} X_i^{\mathrm{T}} \varepsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 A^{-1}).$$

Consider the second term on the right hand side. Let $w_2(Z, g) = g(Z)(A_n^{(j)})^{-1} X$. We have

$$E \left\| w_2(Z, \overline{g}) - w_2(Z, g_0) \right\|^2 = E \left\| (\overline{g}(Z_i) - g_0(Z_i)) \left(A_n^{(j)} \right)^{-1} X_i \right\|^2$$
$$\leq O\big(E \|\overline{g} - g_0\|_2^2 \big).$$

By Lemma A.2 of [8], the logarithm of the ε-bracketing number of the class of functions $\mathcal{A}_2(\delta) = \{w_2(\cdot, \overline{g}) - s(\cdot, g_0) : \overline{g} \in \mathcal{G}_N, \|\overline{g} - g_0\|_2 \leq \delta\}$ is $c\{(J_N - \varrho)\log(\delta/\varepsilon) + \log(\delta^{-1})\}$. Thus, the corresponding entropy integral $J_{[\,]}(\delta, \mathcal{A}_2(\delta), \|\cdot\|_2) \leq c\delta\{(J_N - \varrho)^{1/2} + \log^{1/2}(\delta^{-1})\}$. According to Lemma 7 of [15] and Theorem 2.2, $\|\overline{g} - g_0\|_\infty \leq cJ_N^{1/2}\|\overline{g} - g_0\|_2 = O_P(J_N N^{-1/2} + J_N^{1/2} h_N^p)$. Let $r_N^{-1} = J_N^{1/2} N^{-1/2} + h_N^p$, then

$$E \left| \frac{1}{n} \sum_{i \in G_j} \left\{ (\overline{g}(Z_i) - g_0(Z_i)) \left(A_n^{(j)} \right)^{-1} X_i \right\} \right.$$
$$\left. - E \left\{ (\overline{g}(Z_i) - g_0(Z_i)) \left(A_n^{(j)} \right)^{-1} X_i \right\} \right|$$
$$\leq n^{-1/2} C r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\}$$
$$\times \left[1 + \frac{c r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\}}{r_N^{-2} \sqrt{n}} C_0 \right]$$
$$\leq O(s^{1/2}) n^{-1/2} C r_N^{-1} \left\{ (J_N + \varrho)^{1/2} + \log^{1/2}(r_N) \right\}$$
$$= O(n^{-1/2}) \times O \left(J_N n^{-1/2} + s^{1/2} J_N^{1/2} h_N^p \right)$$
$$= O(n^{-1/2}) \times O \left(J_N n^{-1/2} + \left(n^{-1} N^{1+q(1-2p)} \right)^{1/2} \right)$$
$$\leq O(n^{-1/2}) \times O \left(J_N n^{-1/2} + \left(n^{-1} N^{1/(2p)} \right)^{1/2} \right),$$

where the last inequality is due to the condition that $J_N \gg N^{1/(2p)}$. The condition that $J_N^2 \ll n$ implies that $O(J_N n^{-1/2}) = o(1)$ and $n \gg N^{1/(2p)}$ to make sure that the above expectation has an order $o(n^{-1/2})$. Furthermore,

$$E \left\{ (\overline{g}(Z_i) - g_0(Z_i)) \left(A_n^{(j)} \right)^{-1} X_i \right\} \leq O(E\|\overline{g} - g_0\|_\infty) = O(J_N N^{-1/2}).$$

Thus,

$$n^{-1/2} \sum_{i \in G_j} \left(A_n^{(j)} \right)^{-1} X_i (g_0(Z_i) - \overline{g}(Z_i)) = O(J_N s^{-1/2}) + o_P(1) = o_P(1),$$

where the last equality is due to the condition that $J_N^2 \ll s$. Therefore, the theorem is proved. $\square$

Proof of Theorem 3.1. Under the null hypothesis, we have

$$\sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big) = \sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \boldsymbol{\beta}_0^{(j_1)}\big) - \sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_2)} - \boldsymbol{\beta}_0^{(j_2)}\big).$$

By Theorem 2.1, we have

$$\sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_t)} - \boldsymbol{\beta}_0^{(j_t)}\big) \xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}, \sigma^2 \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^{\mathrm{T}}\big),$$

where $t = 1$ or 2. Therefore, $\sqrt{n}\boldsymbol{Q}\big(\widehat{\boldsymbol{\beta}}^{(j_1)} - \widehat{\boldsymbol{\beta}}^{(j_2)}\big) \xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}, 2\sigma^2 \boldsymbol{Q}\boldsymbol{D}^{-1}\boldsymbol{Q}^{\mathrm{T}}\big).$

Consider the second result. According to the proof of Theorem 2.3, we have

$$\sqrt{n}\left(\breve{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\right) = n^{-1/2} \sum_{i \in G_j} \left(\boldsymbol{A}_n^{(j)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i$$

$$+ n^{-1/2} \sum_{i \in G_j} \left(\boldsymbol{A}_n^{(j)}\right)^{-1} \boldsymbol{X}_i(g_0(\boldsymbol{Z}_i) - \overline{g}(\boldsymbol{Z}_i)).$$

Thus, with similar arguments in the proof of Theorem 2.3, we have

$$\sqrt{n}\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \breve{\boldsymbol{\beta}}^{(j_2)}\big) = \sqrt{n}\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_1)} - \boldsymbol{\beta}_0^{(j_1)}\big) - \sqrt{n}\boldsymbol{Q}\big(\breve{\boldsymbol{\beta}}^{(j_2)} - \boldsymbol{\beta}_0^{(j_2)}\big)$$

$$= n^{-1/2} \sum_{i \in G_{j_1}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_1)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i + n^{-1/2} \sum_{i \in G_{j_1}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_1)}\right)^{-1} \boldsymbol{X}_i(g_0(\boldsymbol{Z}_i) - \overline{g}(\boldsymbol{Z}_i))$$

$$- n^{-1/2} \sum_{i \in G_{j_2}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_2)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i - n^{-1/2} \sum_{i \in G_{j_2}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_2)}\right)^{-1} \boldsymbol{X}_i(g_0(\boldsymbol{Z}_i) - \overline{g}(\boldsymbol{Z}_i))$$

$$= n^{-1/2} \sum_{i \in G_{j_1}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_1)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i - n^{-1/2} \sum_{i \in G_{j_2}} \boldsymbol{Q}\left(\boldsymbol{A}_n^{(j_2)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i + o_p(1)$$

$$\xrightarrow{d} \mathcal{N}\big(\boldsymbol{0}, 2\sigma^2 \boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{Q}^{\mathrm{T}}\big).$$

Therefore, the second result is also proved. $\square$

Proof of Theorem 3.2. First of all, based on the proof of Theorem 2.3, we have

$$\sqrt{n}\left(\breve{\boldsymbol{\beta}}^{(j)} - \breve{\boldsymbol{\beta}}_0^{(j)}\right) = n^{-1/2} \sum_{i \in G_j} \left(\widehat{\boldsymbol{A}}_n^{(j)}\right)^{-1} \boldsymbol{X}_i^{\mathrm{T}} \varepsilon_i + R_i.$$

where $R_i = n^{-1/2} \sum_{i \in G_j} \left(\widehat{\boldsymbol{A}}_n^{(j)}\right)^{-1} \boldsymbol{X}_i(g_0(\boldsymbol{Z}_i) - \overline{g}(\boldsymbol{Z}_i))$. Also, we have shown there

$$\max_{j \in \mathcal{S}} \|R_i\|_\infty = O_P(J_N s^{-\frac{1}{2}}).$$

Noting that $s \gg J_N^2 \log(pd)$, it follows

$$\max_{j \in \mathcal{S}} \|R_i\|_\infty = o_P(\log^{-\frac{1}{2}}(pd)).$$

Then, the rest of the proof follows a similar line with the counterpart of Theorem 3.9 in [21] by discussing the asymptotic differences between approximation terms.   □

Proof of Theorem 3.3. The estimates for g_k and g_{-k} depends on B-spline basis expansions. Denote $\boldsymbol{\gamma}_k$ as coefficients for basis $\boldsymbol{B}_k(Z_k)$ of the kth function g_k, and $\boldsymbol{\gamma}_{-k}$ as coefficients for basis $\boldsymbol{B}_{-k}(\boldsymbol{Z}_{-k})$ of g_{-k}. With a little abuse of notations, we write

$$
\begin{aligned}
&L_{nk}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \boldsymbol{\gamma}_{-k}) \\
&= \frac{1}{2n} \sum_{i \in G_j} \left[Y_i - \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta} - \boldsymbol{B}_k^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_k - \boldsymbol{B}_{-k}^{\mathrm{T}}(\boldsymbol{Z}_{i,-k})\boldsymbol{\gamma}_{-k} \right]^2 \\
&= \frac{1}{2n} \sum_{i \in G_j} \left[\varepsilon_i + \boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{\beta}_0^{(j)} - \boldsymbol{\beta}) + \boldsymbol{B}_k^{\mathrm{T}}(Z_{ik})(\widetilde{\boldsymbol{\gamma}}_k^{(j)} - \boldsymbol{\gamma}_k) \right. \\
&\qquad\qquad \left. + (g_{0k}^{(j)}(Z_{ik}) - \widetilde{g}_k^{(j)}(Z_{ik})) + (g_{0,-k}(\boldsymbol{Z}_{i,-k}) - \boldsymbol{B}_{-k}^{\mathrm{T}}(\boldsymbol{Z}_{i,-k})\boldsymbol{\gamma}_{-k}) \right],
\end{aligned}
$$

where $\widetilde{g}_k^{(j)}(Z_{ik}) = \boldsymbol{B}_k^{\mathrm{T}}(Z_{ik})\boldsymbol{\gamma}_k^{(j)}$.

Consider the derivative of $L_{nk}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \boldsymbol{\gamma}_{-k})$ with respect to $\boldsymbol{\gamma}_k$:

$$
\begin{aligned}
\nabla L_{nk}^{(j)}(\boldsymbol{\beta}, \boldsymbol{\gamma}_k, \boldsymbol{\gamma}_{-k}) &= -\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) + \frac{1}{n} \sum_{i \in G_j} \boldsymbol{B}_k(Z_{ik})^{\otimes 2}(\boldsymbol{\gamma}_k - \widetilde{\boldsymbol{\gamma}}_k^{(j)}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} (\widetilde{g}_k^{(j)}(Z_{ik}) - g_{0k}^{(j)}(Z_{ik}))\boldsymbol{B}_k(Z_{ik}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} (\boldsymbol{B}_{-k}^{\mathrm{T}}(\boldsymbol{Z}_{i,-k})\boldsymbol{\gamma}_{-k} - g_{0,-k}(\boldsymbol{Z}_{i,-k}))\boldsymbol{B}_k(Z_{ik}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{B}_k(Z_{ik}).
\end{aligned}
$$

By plugging estimates based on the jth sub-population, it follows that

$$
\begin{aligned}
\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{\boldsymbol{\gamma}}_k^{(j)}, \widehat{\boldsymbol{\gamma}}_{-k}^{(j)}) &= -\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) + \frac{1}{n} \sum_{i \in G_j} \boldsymbol{B}_k(Z_{ik})^{\otimes 2}(\widehat{\boldsymbol{\gamma}}_k^{(j)} - \widetilde{\boldsymbol{\gamma}}_k^{(j)}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} (\widetilde{g}_k^{(j)}(Z_{ik}) - g_{0k}^{(j)}(Z_{ik}))\boldsymbol{B}_k(Z_{ik}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} (\boldsymbol{B}_{-k}^{\mathrm{T}}(\boldsymbol{Z}_{i,-k})\widehat{\boldsymbol{\gamma}}_{-k}^{(j)} - g_{0,-k}(\boldsymbol{Z}_{i,-k}))\boldsymbol{B}_k(Z_{ik}) \\
&\quad + \frac{1}{n} \sum_{i \in G_j} \boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)})\boldsymbol{B}_k(Z_{ik}), \\
\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widetilde{\boldsymbol{\gamma}}_k^{(j)}, \widehat{\boldsymbol{\gamma}}_{-k}^{(j)}) &= -\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \boldsymbol{B}_k(Z_{ik})
\end{aligned}
$$

$$+\frac{1}{n}\sum_{i\in G_j}(\widetilde{g}_k^{(j)}(Z_{ik})-g_{0k}^{(j)}(Z_{ik}))\boldsymbol{B}_k(Z_{ik})$$

$$+\frac{1}{n}\sum_{i\in G_j}(\boldsymbol{B}_{-k}^{\mathrm{T}}(\boldsymbol{Z}_{i,-k})\widehat{\boldsymbol{\gamma}}_{-k}^{(j)}-g_{0,-k}(\boldsymbol{Z}_{i,-k}))\boldsymbol{B}_k(Z_{ik})$$

$$+\frac{1}{n}\sum_{i\in G_j}\boldsymbol{X}_i^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0^{(j)})\boldsymbol{B}_k(Z_{ik}).$$

Then, we obtain

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})=\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})+\frac{1}{n}\sum_{i\in G_j}\boldsymbol{B}_k(Z_{ik})^{\otimes2}(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widetilde{\boldsymbol{\gamma}}_k^{(j)}).$$

Due to the stationary equation to (2.2), we have

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})=-\frac{1}{n}\sum_{i\in G_j}\boldsymbol{B}_k(Z_{ik})^{\otimes2}(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widetilde{\boldsymbol{\gamma}}_k^{(j)}).$$

Let $E\{\boldsymbol{B}_k(Z_{ik})^{\otimes2}\}=\boldsymbol{\Sigma}_k$. According to Lemma A.2, we know $\boldsymbol{\Sigma}_k$'s eigenvalues are bounded and away from 0. Thus,

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})-\nabla L_{nk}^{(j+1)}(\widehat{\boldsymbol{\beta}}^{(j+1)},\widehat{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j+1)})$$
$$=\quad-(\boldsymbol{\Sigma}_k+o_P(1))(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widehat{\boldsymbol{\gamma}}_k^{(j+1)}),$$

where the last equation is due to $\widetilde{\boldsymbol{\gamma}}_k^{(j)}=\widetilde{\boldsymbol{\gamma}}_k^{(j+1)}$ under the null hypothesis.

By Taylor expansion, we have

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})$$
$$=\quad\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})+\frac{1}{n}\sum_{i\in G_j}\boldsymbol{B}_k(Z_{ik})^{\otimes2}(\widehat{\boldsymbol{\gamma}}_k^{(j+1)}-\widetilde{\boldsymbol{\gamma}}_k^{(j+1)})$$
$$=\quad\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})-\nabla L_{nk}^{(j+1)}(\widehat{\boldsymbol{\beta}}^{(j+1)},\widetilde{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j+1)})+o_P(1).$$

Then, applying Taylor expansion again leads to

$$L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})-L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})$$
$$=\quad\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widehat{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})^{\mathrm{T}}(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widehat{\boldsymbol{\gamma}}_k^{(j+1)})$$
$$\quad+\frac{1}{2}(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widehat{\boldsymbol{\gamma}}_k^{(j+1)})^{\mathrm{T}}(\boldsymbol{\Sigma}_k+o_P(1))(\widehat{\boldsymbol{\gamma}}_k^{(j)}-\widehat{\boldsymbol{\gamma}}_k^{(j+1)})$$
$$=\quad-\frac{1}{2}(\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})-\nabla L_{nk}^{(j+1)}(\widehat{\boldsymbol{\beta}}^{(j+1)},\widetilde{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j+1)}))^{\mathrm{T}}(\boldsymbol{\Sigma}_k+o_P(1))$$
$$\quad\times(\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)},\widetilde{\boldsymbol{\gamma}}_k^{(j)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j)})-\nabla L_{nk}^{(j+1)}(\widehat{\boldsymbol{\beta}}^{(j+1)},\widetilde{\boldsymbol{\gamma}}_k^{(j+1)},\widehat{\boldsymbol{\gamma}}_{-k}^{(j+1)})).$$

Based on Lemmas A.1–A.5, we can verify that

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widetilde{\boldsymbol{\gamma}}_k^{(j)}, \widehat{\boldsymbol{\gamma}}_{-k}^{(j)}) = -\frac{1}{n} \sum_{i \in G_j} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) + R_N,$$

where $\|R_N\|_2 = O_P((J_N^{1/2} n^{-1/2} + h_N^p) J_N^{1/2})$. It is straightforward to have:

$$\nabla L_{nk}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)}, \widetilde{\boldsymbol{\gamma}}_k^{(j)}, \widehat{\boldsymbol{\gamma}}_{-k}^{(j)}) - \nabla L_{nk}^{(j+1)}(\widehat{\boldsymbol{\beta}}^{(j+1)}, \widetilde{\boldsymbol{\gamma}}_k^{(j+1)}, \widehat{\boldsymbol{\gamma}}_{-k}^{(j+1)})$$
$$\overset{d}{=} -\frac{1}{n} \sum_{i \in G_j \bigcup G_{j+1}} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) + R_N.$$

Therefore,

$$n\text{LRT}_{nk}^s$$

$$= -n \sum_{j=1}^{s-1} \left\{ \left[\frac{1}{n} \sum_{i \in G_j \bigcup G_{j+1}} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) \right]^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \left[\frac{1}{n} \sum_{i \in G_j \bigcup G_{j+1}} \varepsilon_i \boldsymbol{B}_k(Z_{ik}) \right] \right.$$
$$\left. + O_P((J_N^{1/2} n^{-1/2} + h_N^p) J_N^{1/2} n^{-1/2}) \right\}$$

$$= -\sum_{j=1}^{s-1} \frac{1}{n} \sum_{i \in G_j \bigcup G_{j+1}} \varepsilon_i^2 \boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik})$$

$$- \sum_{j=1}^{s-1} \frac{1}{n} \sum_{(i_1 < i_2) \in G_j \bigcup G_{j+1}} 2\varepsilon_{i_1} \varepsilon_{i_2} \boldsymbol{B}_k(Z_{i_1 k})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{i_2 k})$$

$$+ O_P((J_N^{1/2} n^{-1/2} + h_N^p) J_N n^{1/2}).$$

The rate of the first term for $n\text{LRT}_{nk}^s$ can be found investigating the moment

$$\frac{1}{n^2} \sum_{i \in G_j \bigcup G_{j+1}} E\left\{ \varepsilon_i^2 \boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik}) - E\varepsilon_i^2 \boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik}) \right\}^2$$
$$\leq \frac{2}{n} E\varepsilon_i^4 (\boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik}))^2$$
$$\leq \frac{2C}{n} J_N^2 < \infty,$$

where C is some constant. Also, $E[\boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik})] = J_N$. Then, it follows

$$\sum_{j=1}^{s-1} \frac{1}{n} \sum_{i \in G_j \bigcup G_{j+1}} \varepsilon_i^2 \boldsymbol{B}_k(Z_{ik})^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{B}_k(Z_{ik}) = 2(s-1)\sigma^2 J_N + o_P(sJ_N). \quad \text{(A.2)}$$

Let $W_{i_1 i_2}^{(j)} = 2\varepsilon_{i_1}\varepsilon_{i_2}\boldsymbol{B}_k(Z_{i_1 k})^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{B}_k(Z_{i_2 k})$, and then define

$$W(N) = \sum_{j=1}^{s-1} \sum_{(i_1 < i_2) \in G_j \bigcup G_{j+1}} W_{i_1 i_2}^{(j)}.$$

The next step is to show the asymptotic normality of $n^{-1}W(N)$, which follows a similar line with the proof of Theorem 4.1 in [12] using Proposition 3.2 in [5]. First, it is easy to see $E(W_{i_1 i_2}^{(j)}|Z_{i_1 k}) = 0, \forall i_1 < i_2$. Second, bounds for several moments of $W(N)$ are derived. Define $\sigma(N)^2 = \mathrm{Var}(W(N))$ and following quantities:

$$U_I = \sum_{j=1}^{s-1} \sum_{\substack{(i_1 < i_2) \\ \in G_j \bigcup G_{j+1}}} E[(W_{i_1 i_2}^{(j)})^4],$$

$$U_{II} = \sum_{j=1}^{s-1} \sum_{\substack{(i_1 < i_2 < i_3) \\ \in G_j \bigcup G_{j+1}}} \left\{ E[(W_{i_1 i_2}^{(j)})^2(W_{i_1 i_3}^{(j)})^2] + E[(W_{i_2 i_1}^{(j)})^2(W_{i_2 i_3}^{(j)})^2] \right.$$

$$\left. + E[(W_{i_3 i_1}^{(j)})^2(W_{i_3 i_2}^{(j)})^2] \right\}$$

$$U_{III} = \sum_{j=1}^{s-1} \sum_{\substack{(i_1 < i_2 < i_3 < i_4) \\ \in G_j \bigcup G_{j+1}}} \left\{ E[W_{i_1 i_2}^{(j)} W_{i_1 i_3}^{(j)} W_{i_4 i_2}^{(j)} W_{i_4 i_3}^{(j)}] + E[W_{i_1 i_2}^{(j)} W_{i_1 i_4}^{(j)} W_{i_3 i_2}^{(j)} W_{i_3 i_4}^{(j)}] \right.$$

$$\left. + E[W_{i_1 i_3}^{(j)} W_{i_1 i_4}^{(j)} W_{i_2 i_3}^{(j)} W_{i_2 i_4}^{(j)}] . \right\}$$

By the definition of $\boldsymbol{\Sigma}_k$ and $E(\varepsilon_i^4) < \infty$, we have

$$E[(W_{i_1 i_2}^{(j)})^4] \leq 2^4 E\left[\varepsilon_{i_1}^4 \varepsilon_{i_2}^4 (\boldsymbol{B}_k(Z_{i_1 k})^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{B}_k(Z_{i_2 k}))^4\right] = O(J_N^4)$$

$$E[(W_{i_1 i_2}^{(j)})^2(W_{i_1 i_3}^{(j)})^2] \leq E[(W_{i_1 i_2}^{(j)})^4] = O(J_N^4)$$

$$E[W_{i_1 i_2}^{(j)} W_{i_1 i_3}^{(j)} W_{i_4 i_2}^{(j)} W_{i_4 i_3}^{(j)}] = O(J_N)$$

Then it follows that

$$U_I = O(sn^2 J_N^4), \ U_{II} = O(sn^3 J_N^4), \ U_{III} = O(sn^4 J_N^4).$$

Third,

$$\sigma(N)^4 \sim \left[(s-1)\binom{2n}{2}4\sigma^4 J_N + 2(s-2)\binom{n}{2}4\sigma^4 J_N\right]^2 \sim (12\sigma^4(s-1)n^2 J_N)^2$$

because $E[(W_{i_1 i_2}^{(j)})^2] = 4\sigma^4 J_N$. Since U_I, U_{II} and U_{III} have a smaller order than $\sigma(N)^4$ under the condition $J_N^2 \ll n$, Proposition 3.2 in [5] shows that

$$\frac{1}{\sqrt{12\sigma^4(s-1)n^2 J_N}} W(N) \xrightarrow{d} \mathcal{N}(0,1). \tag{A.3}$$

Last, under the condition $J_n \gg n^{\frac{1}{2p}}$ from Theorem 2.1 and $s^2 \gg J_N$, we know $(J_N^{1/2} n^{-1/2} + h_N^p) J_N n^{1/2} \ll (J_N^{1/2} + n^{-1/2} n^{1/2}) J_N \ll s J_N$. Therefore, by combining (A.2) and (A.3), we have

$$\frac{1}{\sqrt{2\frac{2}{3}(s-1)J_N}} \left(-\frac{1}{3\sigma^2} n \mathrm{LRT}_{nk}^s - \frac{2}{3}(s-1)J_N \right) \xrightarrow{d} \mathcal{N}(0,1). \qquad \square$$

References

[1] AITKIN, M. and RUBIN, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society. Series B* **47** 67–75.

[2] CARROLL, R., RUPPERT, D. and WAND, M. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press. MR1998720

[3] DAVIDIAN, M. and CARROLL, R. J. (1987). Variance function estimation. *Journal of the American Statistical Association* **82** 1079–1091. MR0922172

[4] DE BOOR, C. (1978). *A Practical Guide to Splines* **27**. Springer-Verlag New York. MR0507062

[5] DE JONG, P. (1987). A central limit theorem for generalized quadratic forms. *Probability Theory and Related Fields* **75** 261–277. MR0885466

[6] FANG, Y., LIAN, H., LIANG, H. and RUPPERT, D. (2015). Variance function additive partial linear models. *Electronic Journal of Statistics* **9** 2793–2827. MR3439185

[7] HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models*, 1st ed. *Monographs on statistics and applied probability*. Chapman and Hall, London; New York. MR1082147

[8] HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *The Annals of Statistics* **27** 1536–1563. MR1742499

[9] HUANG, J. Z. et al. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics* **31** 1600–1635. MR2012827

[10] LI, R., LIN, D. K. and LI, B. (2013). Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry* **29** 399–409. MR3117826

[11] LIU, X., WANG, L. and LIANG, H. (2011). Estimation and Variable Selection for Semiparametric Additive Partial Linear Models. *Statistica Sinica* **21** 1225–1248. MR2827522

[12] LU, J., CHENG, G. and LIU, H. (2016). Nonparametric Heterogeneity Testing For Massive Data. *arXiv preprint arXiv:1601.06212*.

[13] OPSOMER, J. D. and RUPPERT, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics* **8** 715–732.

[14] STONE, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics* **13** 689–705. MR0790566

[15] STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics* **14** 590–606. MR0840516

[16] VOLGUSHEV, S., CHAO, S.-K. and CHENG, G. (2017). Distributed inference for quantile regression processes. *arXiv preprint arXiv:1701.06088.*

[17] WAHBA, G. (1990). *Spline models for observational data* **59**. Siam. MR1045442

[18] WANG, L. and YANG, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics* **35** 2474–2503. MR2382655

[19] WANG, C., CHEN, M.-H., SCHIFANO, E., WU, J. and YAN, J. (2015). Statistical Methods and Computing for Big Data. *arXiv preprint arXiv:1502.07989.* MR3553369

[20] XU, G., SHANG, Z. and CHENG, G. (2016). Optimal tuning for divide-and-conquer kernel ridge regression with massive data. *arXiv preprint arXiv:1612.05907.*

[21] ZHAO, T., CHENG, G. and LIU, H. (2016). A partially linear framework for massive heterogeneous data. *The Annals of Statistics* **44** 1400–1437. MR3519928