The Annals of Statistics
2016, Vol. 44, No. 4, 1400–1437
DOI: 10.1214/15-AOS1410

A PARTIALLY LINEAR FRAMEWORK FOR MASSIVE HETEROGENEOUS DATA

BY TIANQI ZHAO*, GUANG CHENG†,1 AND HAN LIU*,2

*Princeton University** *and Purdue University*†

We consider a partially linear framework for modeling massive heterogeneous data. The major goal is to extract common features across all subpopulations while exploring heterogeneity of each subpopulation. In particular, we propose an aggregation type estimator for the commonality parameter that possesses the (nonasymptotic) minimax optimal bound and asymptotic distribution as if there were no heterogeneity. This oracle result holds when the number of subpopulations does not grow too fast. 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. We also test the heterogeneity among a large number of subpopulations. All the above results require to regularize each subestimation as though it had the entire sample. Our general theory applies to the divide-and-conquer approach that is often used to deal with massive homogeneous data. A technical by-product of this paper is statistical inferences for general kernel ridge regression. Thorough numerical results are also provided to back up our theory.

1. Introduction. In this paper, we propose a partially linear regression framework for modeling massive heterogeneous data. Let $\{(Y_i, \mathbf{X}_i, Z_i)\}_{i=1}^N$ be samples from an underlying distribution that may change with N. We assume that there exist s independent subpopulations, and the data from the jth subpopulation follow a partially linear model:

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0^{(j)} + f_0(Z) + \varepsilon, \tag{1.1}$$

where ε has zero mean and known variance σ^2. In the above model, Y depends on $\mathbf{X}$ through a linear function that may vary across all subpopulations, and depends on Z through a nonlinear function that is common to all subpopulations. The possibly different values of $\boldsymbol{\beta}_0^{(j)}$ are viewed as the source of heterogeneity. In reality, the number of subpopulations grows and some subpopulations may have

Received June 2015; revised October 2015.

[1]Supported by NSF CAREER Award DMS-11-51692, DMS-14-18042, Simons Fellowship in Mathematics, Office of Naval Research (ONR N00014-15-1-2331) and a grant from Indiana Clinical and Translational Sciences Institute.

[2]Supported by NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841.

MSC2010 subject classifications. Primary 62G20, 62F25; secondary 62F10, 62F12.

Key words and phrases. Divide-and-conquer method, heterogeneous data, kernel ridge regression, massive data, partially linear model.

extremely high sample sizes. Note that (1.1) is a typical "semi-nonparametric" model [Cheng and Shang (2015)] since we infer both commonality and heterogeneity components throughout the paper.

The model (1.1) is motivated by the following scenario: different labs conduct the same experiment on the relationship between a response variable Y (e.g., heart disease) and a set of predictors $Z, X_1, X_2, \ldots, X_p$. It is known from biological knowledge that the dependence structure between Y and Z (e.g., blood pressure) should be homogeneous for all humans. However, for the other covariates (e.g., certain genes), we allow their (linear) relations with Y to potentially vary in different labs. For example, the genetic functionality of different races might be heterogenous. The linear relation is assumed here for simplicity, and particularly suitable when the covariates are discrete.

Statistical modeling for massive data has attracted a flurry of recent research. For homogeneous data, the statistical studies of the divide-and-conquer method currently focus on either parametric inferences, for example, Li, Lin and Li (2013), Bag of Little Bootstraps [Kleiner et al. (2012)], and parallel MCMC computing [Wang and Dunson (2013)], or nonparametric minimaxity [Zhang, Duchi and Wainwright (2013)]. The other relevant work includes high dimensional linear models with variable selection [Chen and Xie (2012)] and structured perceptron [McDonald, Hall and Mann (2010)]. Heterogenous data are often handled by fitting mixture models [Aitkin and Rubin (1985), Figueiredo and Jain (2002), McLachlan and Peel (2000)], time varying coefficient models [Fan and Zhang (1999), Hastie and Tibshirani (1993)] or multitask regression [Huang and Zhang (2010), Nardi and Rinaldo (2008), Obozinski, Wainwright and Jordan (2008)]. The recent high dimensional work includes Meinshausen and Bühlmann (2015), Städler, Bühlmann and van de Geer (2010). However, as far as we are aware, *semi-nonparametric inference* for massive homogeneous/heterogeneous data still remains untouched.

In this paper, our primary goal is to extract common features across all subpopulations while exploring heterogeneity of each subpopulation. Specifically, we employ a simple aggregation procedure, which averages commonality estimators across all subpopulations, and then construct a plug-in estimator for each heterogeneity parameter based on the combined estimator for commonality. A similar two-stage estimation method was proposed in Li and Liang (2008), but for the purpose of variable selection in $\boldsymbol{\beta}$ based on a single data set. The secondary goal is to apply the divide-and-conquer method to the subpopulation with a huge sample size that is unable to be processed in one single computer. The above purposes are achieved by estimating our statistical model (1.1) with the kernel ridge regression (KRR) method. In the partially linear literature, there also exist other estimation and inference methods (based on a single dataset) such as profile least squares method, partial residual method and backfitting method; see Härdle, Liang and Gao (2000), Ruppert, Wand and Carroll (2003), Yatchew (2003).

The KRR framework is known to be very flexible and well supported by the general reproducing kernel Hilbert space (RKHS) theory [Mendelson (2002), Steinwart et al. (2009), Zhang (2005)]. In particular, partial smoothing spline models [Cheng, Zhang and Shang (2015)] can be viewed as a special case of our general framework. An important technical contribution of this paper is statistical inferences for general kernel ridge regression by extending smoothing spline results developed in Cheng and Shang (2015). This theoretical innovation makes our work go beyond the existing statistical study on the KRR for large datastes, which mainly focus on their nonparametric minimaxity, for example, Bach (2012), Raskutti, Wainwright and Yu (2014), Zhang, Duchi and Wainwright (2013).

Our theoretical studies are mostly concerned with the so-called "oracle rule" for massive data. Specifically, we define the "oracle estimate" for commonality (heterogeneity) as the one computed when all the heterogeneity information are given (the commonality information is given in each subpopulation), that is, $\boldsymbol{\beta}_0^{(j)}$'s are known ($f_0$ is known). We claim that a commonality estimator satisfies the oracle rule if it possesses the same minimax optimality and asymptotic distribution as the "oracle estimate" defined above. A major contribution of this paper is to derive the largest possible diverging rate of s under which our combined estimator for commonality satisfies the oracle rule. In other words, our aggregation procedure is shown to "filter out" the heterogeneity in data when s does not grow too fast with N. On the other hand, we have to set a lower bound on s for our heterogeneity estimate to possess the asymptotic distribution as if the commonality information were available, that is, oracle rule. Our second contribution is to test the heterogeneity among a large number of subpopulations by employing a recent Gaussian approximation theory [Chernozhukov, Chetverikov and Kato (2013)].

In the standard implementation of KRR, we must invert a kernel matrix, which requires costs $O(N^3)$ in time and $O(N^2)$ in memory, respectively; see Saunders, Gammerman and Vovk (1998). This is computationally prohibitive for a large N. Hence, when some subpopulation has a huge sample size, we may apply the divide-and-conquer approach whose statistical analysis directly follows from the above results. In this case, the "oracle estimate" is defined as those computed based on the entire (homogeneous) data in those subpopulations. A rather different goal here is to explore the most computationally efficient way to split the whole sample while performing the best possible statistical inference. Specifically, we derive the largest possible number of splits under which the averaged estimators for both components enjoy the same statistical properties as the oracle estimators.

In both homogeneous and heterogeneous settings above, we note that the upper bounds established for s increase with the smoothness of f_0. Hence, our aggregation procedure favors smoother regression functions in the sense that more subpopulations/splits are allowed in the massive data. On the other hand, we have to admit that our upper and lower bound results for s are only sufficient conditions although empirical results show that our bounds are quite sharp. Another interesting finding

is that even the semi-nonparametric estimation is applied to only one fraction of the entire data; it is nonetheless essential to regularize each subestimation as if it had the entire sample.

In the end, we highlight two key technical challenges: (i) delicate interplay between the parametric and nonparametric components in the *semi-nonparametric estimation*. In particular, we observe a "bias propagation" phenomenon: the bias introduced by the penalization of the nonparametric component propagates to the parametric component, and the resulting parametric bias in turn propagates back to the nonparametric component. To analyze this complicated propagation mechanism, we extend the existing RKHS theory to an enlarged partially linear function space by defining a novel inner product under which the expectation of the Hessian of the objective function becomes identity; see Proposition 2.2. (ii) double asymptotics framework in terms of diverging s and N. In this challenging regime, we develop more refined concentration inequalities in characterizing the concentration property of an averaged empirical process. These very refined theoretical analyses show that an average of s asymptotic linear expansions is still a valid one as $s \wedge N \to \infty$.

The rest of the paper is organized as follows: Section 2 briefly introduces the general RKHS theory and discusses its extension to an enlarged partially linear function space. Section 3 describes our aggregation procedure, and studies the "oracle" property of this procedure from both asymptotic and nonasymptotic perspectives. The efficiency boosting of heterogeneity estimators and heterogenous testing results are also presented in this section. Section 4 applies our general theory to various examples with different smoothness. Section 5 is devoted to the analysis of divide-and-conquer algorithms for homogeneous data. Section 6 presents some numerical experiments. All the technical details are deferred to Section 7 or Online Supplementary [Zhao, Cheng and Liu (2015)].

Notation. Denote $\|\cdot\|_2$ and $\|\cdot\|_\infty$ as the Euclidean L_2 and infinity norm in $\mathbb{R}^p$, respectively. For any function $f : \mathcal{S} \mapsto \mathbb{R}$, let $\|f\|_{\sup} = \sup_{x\in\mathcal{S}} |f(x)|$. We use $\|\cdot\|$ to denote the spectral norm of matrices. For positive sequences a_n and b_n, we write $a_n \lesssim b_n$ ($a_n \gtrsim b_n$) if there exists some universal constant $c > 0$ ($c' > 0$) independent of n such that $a_n \le cb_n$ ($a_n \ge c'b_n$) for all $n \in \mathbb{N}$. We denote $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

2. Preliminaries. In this section, we briefly introduce the general RKHS theory, and then extend it to a partially linear function space. Below is a generic definition of RKHS [Berlinet and Thomas-Agnan (2004)].

DEFINITION 2.1. Denote by $\mathcal{F}(\mathcal{S}, \mathbb{R})$ a vector space of functions from a general set $\mathcal{S}$ to $\mathbb{R}$. We say that $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) on $\mathcal{S}$, provided that:

(i) $\mathcal{H}$ is a vector subspace of $\mathcal{F}(\mathcal{S}, \mathbb{R})$;

(ii) $\mathcal{H}$ is endowed with an inner product, denoted as $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, under which it becomes a Hilbert space;
(iii) for every $y \in \mathcal{S}$, the linear evaluation functional defined by $E_y(f) = f(y)$ is bounded.

If $\mathcal{H}$ is a RKHS, by Riesz representation, we have that for every $y \in \mathcal{S}$, there exists a unique vector, $K_y \in \mathcal{H}$, such that for every $f \in \mathcal{H}$, $f(y) = \langle f, K_y \rangle_{\mathcal{H}}$. The reproducing kernel for $\mathcal{H}$ is defined as $K(x, y) = K_y(x)$.

Denote $U := (\mathbf{X}, Z) \in \mathcal{X} \times \mathcal{Z} \subset \mathbb{R}^p \times \mathbb{R}$, and $\mathbb{P}_U$ as the distribution of U ($\mathbb{P}_X$ and $\mathbb{P}_Z$ are defined similarly). According to Definition 2.1, if $\mathcal{S} = \mathcal{Z}$ and $\mathcal{F}(\mathcal{Z}, \mathbb{R}) = L_2(\mathbb{P}_Z)$, then we can define a RKHS $\mathcal{H} \subset L_2(\mathbb{P}_Z)$ (endowed with a proper inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$), in which the true function f_0 is believed to lie. The corresponding kernel for $\mathcal{H}$ is denoted by K such that for any $z \in \mathcal{Z}$: $f(z) = \langle f, K_z \rangle_{\mathcal{H}}$. By Mercer's theorem, this kernel function has the following unique eigen-decomposition:

$$K(z_1, z_2) = \sum_{\ell=1}^{\infty} \mu_\ell \phi_\ell(z_1) \phi_\ell(z_2),$$

where $\mu_1 \geq \mu_2 \geq \cdots > 0$ are eigenvalues and $\{\phi_\ell\}_{\ell=1}^{\infty}$ are an orthonormal basis in $L_2(\mathbb{P}_Z)$. Mercer's theorem together with the reproducing property implies that $\langle \phi_i, \phi_j \rangle_{\mathcal{H}} = \delta_{ij}/\mu_i$, where δ_{ij} is Kronecker's delta. The smoothness of the functions in RKHS can be characterized by the decaying rate of $\{\mu_\ell\}_{\ell=1}^{\infty}$. Below, we present three different decaying rates together with the corresponding kernel functions.

Finite rank kernel: The kernel has finite rank r if $\mu_\ell = 0$ for all $\ell > r$. For example, the linear kernel $K(\mathbf{z}_1, \mathbf{z}_2) = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle_{\mathbb{R}^d}$ has rank d, and generates a d-dimensional linear function space. The eigenfunctions are given by $\phi_\ell(\mathbf{z}) = z_\ell$ for $\ell = 1, \ldots, d$. The polynomial kernel $K(z_1, z_2) = (1 + z_1 z_2)^d$ has rank $d + 1$, and generates a space of polynomial functions with degree at most d. The eigenfunctions are given by $\phi_\ell = z^{\ell-1}$ for $\ell = 1, \ldots, d + 1$.

Exponentially decaying kernel: The kernel has eigenvalues that satisfy $\mu_\ell \asymp c_1 \exp(-c_2 \ell^p)$ for some $c_1, c_2 > 0$. An example is the Gaussian kernel $K(z_1, z_2) = \exp(-|z_1 - z_2|^2)$. The eigenfunctions are given by Sollich and Williams (2005)

$$\phi_\ell(x) = (\sqrt{5}/4)^{1/4} \big(2^{\ell-2}(\ell - 1)!\big)^{-1/2} e^{-(\sqrt{5}-1)x^2/4} H_{\ell-1}\big((\sqrt{5}/2)^{1/2} x\big), \tag{2.1}$$

for $\ell = 1, 2, \ldots$, where $H_\ell(\cdot)$ is the ℓth Hermite polynomial.

Polynomially decaying kernel: The kernel has eigenvalues that satisfy $\mu_\ell \asymp \ell^{-2\nu}$ for some $\nu \geq 1/2$. Examples include those underlying for Sobolev space and Besov space [Birman and Solomyak (1967)]. In particular, the eigenfunctions of a νth order periodic Sobolev space are trigonometric functions as specified in Section 4.3. The corresponding Sobolev kernels are given in Gu (2013).

In this paper, we consider the following penalized estimation:

$$(\widehat{\boldsymbol{\beta}}^{\dagger}, \widehat{f}^{\dagger}) = \operatorname*{argmin}_{(\boldsymbol{\beta}, f)\in\mathcal{A}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \bigl(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - f(Z_i)\bigr)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{2.2}$$

where $\lambda > 0$ is a regularization parameter and $\mathcal{A}$ is defined as the parameter space $\mathbb{R}^p \times \mathcal{H}$. For simplicity, we do not distinguish $m = (\boldsymbol{\beta}, f) \in \mathcal{A}$ from its associated function

$$m \in \mathcal{M} := \bigl\{ m \mid m(u) = \boldsymbol{\beta}^T\mathbf{x} + f(z), \text{ for } u = (\mathbf{x}, z) \in \mathcal{X}\times\mathcal{Z}, (\boldsymbol{\beta}, f) \in \mathcal{A} \bigr\},$$

throughout the paper. We call $(\widehat{\boldsymbol{\beta}}^{\dagger}, \widehat{f}^{\dagger})$ as partially linear kernel ridge regression (KRR) estimate in comparison with the nonparametric KRR estimate in Shawe-Taylor and Cristianini (2004). In particular, when $\mathcal{H}$ is a νth order Sobolev space endowed with $\langle f, \widetilde{f}\rangle_{\mathcal{H}} := \int_{\mathcal{Z}} f^{(\nu)}(z)\widetilde{f}^{(\nu)}(z)\,dz$, $(\widehat{\boldsymbol{\beta}}^{\dagger}, \widehat{f}^{\dagger})$ becomes the commonly used partial smoothing spline estimate.

We next illustrate that $\mathcal{A}$ can be viewed as a partially linear extension of $\mathcal{H}$ in the sense that it shares some nice reproducing properties as this RKHS $\mathcal{H}$ under the following inner product: for any $m = (\boldsymbol{\beta}, f) \in \mathcal{A}$ and $\widetilde{m} = (\widetilde{\boldsymbol{\beta}}, \widetilde{f}) \in \mathcal{A}$, define

$$\langle m, \widetilde{m}\rangle_{\mathcal{A}} := \langle m, \widetilde{m}\rangle_{L_2(\mathbb{P}_{\mathbf{X},Z})} + \lambda\langle f, \widetilde{f}\rangle_{\mathcal{H}}, \tag{2.3}$$

where $\langle m, \widetilde{m}\rangle_{L_2(\mathbb{P}_{\mathbf{X},Z})} = \mathbb{E}_{\mathbf{X},Z}[(\mathbf{X}^T\boldsymbol{\beta} + f(Z))(\mathbf{X}^T\widetilde{\boldsymbol{\beta}} + \widetilde{f}(Z))]$. Note that m and $\widetilde{m}$ in $\langle m, \widetilde{m}\rangle_{L_2(\mathbb{P}_{\mathbf{X},Z})}$ are both functions in the set $\mathcal{M}$. Similar to the kernel function K_z, we can construct a linear operator $R_u(\cdot) \in \mathcal{A}$ such that $\langle R_u, m\rangle_{\mathcal{A}} = m(u)$ for any $u \in \mathcal{X}\times\mathcal{Z}$. Also, construct another linear operator $P_\lambda : \mathcal{A} \mapsto \mathcal{A}$ such that $\langle P_\lambda m, \widetilde{m}\rangle_{\mathcal{A}} = \lambda\langle f, \widetilde{f}\rangle_{\mathcal{H}}$ for any m and $\widetilde{m}$. See Proposition 2.3 for the construction of R_u and P_λ.

We next present a proposition illustrating the rationale behind the definition of $\langle\cdot,\cdot\rangle_{\mathcal{A}}$. Denote $\otimes$ as the outer product on $\mathcal{A}$. Hence, $\mathbb{E}_U[R_U \otimes R_U] + P_\lambda$ is an operator from $\mathcal{A}$ to $\mathcal{A}$.

PROPOSITION 2.2. $\mathbb{E}_U[R_U \otimes R_U] + P_\lambda = \mathrm{id}$, *where* id *is an identity operator on* $\mathcal{A}$.

PROOF. For any $m = (\boldsymbol{\beta}, f) \in \mathcal{A}$ and $\widetilde{m} = (\widetilde{\boldsymbol{\beta}}, \widetilde{f}) \in \mathcal{A}$, we have

$$\begin{aligned}\bigl\langle\bigl(\mathbb{E}_U[R_U \otimes R_U] + P_\lambda\bigr)m, \widetilde{m}\bigr\rangle_{\mathcal{A}} &= \bigl\langle\mathbb{E}_U[R_U \otimes R_U]m, \widetilde{m}\bigr\rangle_{\mathcal{A}} + \langle P_\lambda m, \widetilde{m}\rangle_{\mathcal{A}} \\ &= \mathbb{E}_U\bigl[m(U)\widetilde{m}(U)\bigr] + \lambda\langle f, \widetilde{f}\rangle_{\mathcal{H}} = \langle m, \widetilde{m}\rangle_{\mathcal{A}}.\end{aligned}$$

Since the choice of $(m, \widetilde{m})$ is arbitrary, we complete our proof. □

As will be seen in the subsequent analysis, for example, in Theorem 3.4, the operator $\mathbb{E}[R_U \otimes R_U] + P_\lambda$ is essentially the expectation of the Hessian of the objective function (w.r.t. Fréchet derivative) minimized in (2.2). Proposition 2.2

shows that the inversion of this Hessian matrix is trivial when the inner product is designed as in (2.3). Due to that, the theoretical analysis of $\widehat{m}^\dagger = (\widehat{\boldsymbol{\beta}}^\dagger, \widehat{f}^\dagger)$ based on the first order optimality condition becomes much more transparent.

To facilitate the construction of R_u and P_λ, we need to endow a new inner product with $\mathcal{H}$:

$$\langle f, \widetilde{f}\rangle_{\mathcal{C}} = \langle f, \widetilde{f}\rangle_{L_2(\mathbb{P}_Z)} + \lambda\langle f, \widetilde{f}\rangle_{\mathcal{H}}, \tag{2.4}$$

for any $f, \widetilde{f} \in \mathcal{H}$. Under (2.4), $\mathcal{H}$ is still a RKHS as the evaluation functional is bounded by Lemma A.1 in the supplemental material [Zhao, Cheng and Liu (2015)]. We denote the new kernel function as $\widetilde{K}(\cdot,\cdot)$, and define a positive definite self-adjoint operator $W_\lambda : \mathcal{H} \mapsto \mathcal{H}$:

$$\langle W_\lambda f, \widetilde{f}\rangle_{\mathcal{C}} = \lambda\langle f, \widetilde{f}\rangle_{\mathcal{H}} \qquad \text{for any } f, \widetilde{f} \in \mathcal{H}', \tag{2.5}$$

whose existence is proven in Lemma A.2 in the supplemental material [Zhao, Cheng and Liu (2015)]. We next define two crucial quantities needed in the construction: $B_k := \mathbb{E}[X_k \mid Z]$ and its Riesz representer $A_k \in \mathcal{H}$ satisfying $\langle A_k, f\rangle_{\mathcal{C}} = \langle B_k, f\rangle_{L_2(\mathbb{P}_Z)}$ for all $f \in \mathcal{H}$. Here, we implicity assume B_k is square integrable. The existence of A_k follows from the boundedness of the linear functional $\mathcal{B}_k f := \langle B_k, f\rangle_{L_2(\mathbb{P}_Z)}$ (by Riesz's representer theorem) as follows:

$$|\mathcal{B}_k f| = \big|\langle B_k, f\rangle_{L_2(\mathbb{P}_Z)}\big| \le \|B_k\|_{L_2(\mathbb{P}_Z)}\|f\|_{L_2(\mathbb{P}_Z)} \le \|B_k\|_{L_2(\mathbb{P}_Z)}\|f\|_{\mathcal{C}}.$$

We are now ready to construct R_u and P_λ based on $\widetilde{K}_z$, W_λ, $\mathbf{B}$ and $\mathbf{A}$ introduced above, where $\mathbf{B} = (B_1, \ldots, B_p)^T$ and $\mathbf{A} = (A_1, \ldots, A_p)^T$. Define $\boldsymbol{\Omega} = \mathbb{E}[(\mathbf{X} - \mathbf{B})(\mathbf{X} - \mathbf{B})^T]$ and $\boldsymbol{\Sigma}_\lambda = \mathbb{E}[\mathbf{B}(Z)(\mathbf{B}(Z) - \mathbf{A}(Z))^T]$.

PROPOSITION 2.3. *For any $u = (\mathbf{x}, z)$, R_u can be expressed as $R_u : u \mapsto (L_u, N_u) \in \mathcal{A}$, where*

$$L_u = (\boldsymbol{\Omega} + \boldsymbol{\Sigma}_\lambda)^{-1}\big(\mathbf{x} - \mathbf{A}(z)\big) \quad \textit{and} \quad N_u = \widetilde{K}_z - \mathbf{A}^T L_u.$$

Moreover, for any $m = (\boldsymbol{\beta}, f) \in \mathcal{A}$, $P_\lambda m$ can be expressed as $P_\lambda m = (L_\lambda f, N_\lambda f) \in \mathcal{A}$, where

$$L_\lambda f = -(\boldsymbol{\Omega} + \boldsymbol{\Sigma}_\lambda)^{-1}\langle \mathbf{B}, W_\lambda f\rangle_{L_2(\mathbb{P}_Z)} \quad \textit{and} \quad N_\lambda f = W_\lambda f - \mathbf{A}^T L_\lambda f.$$

The quantities R_u and P_λ correspond to the variance, that is, $n^{-1}\sum_{i=1}^n R_{U_i}\varepsilon_i$, and bias, that is, $P_\lambda m_0$, in the stochastic expansion of $\widehat{m}^\dagger - m_0$, where $\widehat{m}^\dagger = (\widehat{\boldsymbol{\beta}}^\dagger, \widehat{f}^\dagger)$, $m_0 = (\boldsymbol{\beta}_0, f_0)$; see equation (7.2) in Section 7.1. We remark that the penalized loss function in (2.2) can be written as $(1/n)\sum_{i=1}^n (Y_i - \langle R_{U_i}, m\rangle_{\mathcal{A}})^2 + \langle P_\lambda m, m\rangle_{\mathcal{A}}$. This explains why R_u and P_λ show up in the stochastic expansion, which is derived from the KKT condition of the above loss function and Proposition 2.2. Moreover, $R_u = (L_u, N_u)$ and $P_\lambda m = (L_\lambda f, N_\lambda f)$. Hence, $L_u, L_\lambda f_0$ and $N_u, N_\lambda f_0$ appear in the stochastic expansions of $\widehat{\boldsymbol{\beta}}^\dagger - \boldsymbol{\beta}_0$ and $\widehat{f}^\dagger - f_0$; see Lemma 3.1.

3. Heterogeneous data: Aggregation of commonality. In this section, we start from describing our aggregation procedure and model assumptions in Section 3.1. The main theoretical results are presented in Sections 3.2–3.4 showing that our combined estimate for commonality enjoys the "oracle property." To be more specific, we show that it possesses the same (nonasymptotic) minimax optimal bound (in terms of mean-squared error) and asymptotic distribution as the "oracle estimate" $\widehat{f}_{\mathrm{or}}$ computed when all the heterogeneity information are available:

$$\widehat{f}_{\mathrm{or}} = \operatorname*{argmin}_{f\in\mathcal{H}} \left\{ \frac{1}{N} \sum_{j=1}^{s} \sum_{i\in S_j} \bigl(Y_i - (\boldsymbol{\beta}_0^{(j)})^T \mathbf{X}_i - f(Z_i)\bigr)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}, \tag{3.1}$$

where S_j denotes the index set of all samples from the subpopulation j. The above nice properties hold when the number of subpopulations does not grow too fast and the smoothing parameter is chosen according to the entire sample size N. Based on this combined estimator, we further construct a plug-in estimator for each heterogeneity parameter $\boldsymbol{\beta}_0^{(j)}$, which possesses the asymptotic distribution as if the commonality were known, in Section 3.5. Interestingly, this oracular result holds when the number of subpopulation is not too small. In the end, Section 3.6 tests the possible heterogeneity among a large number of subpopulations.

3.1. *Method and assumptions.* The heterogeneous data setup and averaging procedure are described below:

1. Observe data $(\mathbf{X}_i, Z_i, Y_i)$ with the known labels indicating the subpopulation it belongs to, for $i = 1, \ldots, N$. The size of samples from each subpopulation is assumed to be the same, denoted by n, for simplicity. Hence, $N = n \times s$.
2. On the jth subpopulation, obtain the following penalized estimator:

$$(\widehat{\boldsymbol{\beta}}_{n,\lambda}^{(j)}, \widehat{f}_{n,\lambda}^{(j)}) = \operatorname*{argmin}_{(\boldsymbol{\beta},f)\in\mathcal{A}} \left\{ \frac{1}{n} \sum_{i\in S_j} \bigl(Y_i - \mathbf{X}_i^T\boldsymbol{\beta} - f(Z_i)\bigr)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}. \tag{3.2}$$

3. Obtain the final nonparametric estimate[3] for commonality by averaging

$$\bar{f}_{N,\lambda} = \frac{1}{s} \sum_{j=1}^{s} \widehat{f}_{n,\lambda}^{(j)}. \tag{3.3}$$

We point out that $\widehat{\boldsymbol{\beta}}_{n,\lambda}^{(j)}$ is not our final estimate for heterogeneity. In fact, it can be further improved based on $\bar{f}_{N,\lambda}$; see Section 3.5.

[3]The commonality estimator $\bar{f}_{N,\lambda}$ can be adjusted as a weighted sum $\sum_{j=1}^{s} (n_j/N) \widehat{f}_{n,\lambda}^{(j)}$ if subsample sizes are different. In particular, the divide-and-conquer method can be applied to those subpopulations with huge sample sizes; see Section 5.

For simplicity, we will drop the subscripts (n, λ) and (N, λ) in those notation defined in (3.2) and (3.3) throughout the rest of this paper. Moreover, we make the technical assumption that $s \lesssim N^{\psi}$, although ψ could be very close to 1. The main assumptions of this section are stated below.

ASSUMPTION 3.1 (Regularity condition). (i) ε_i's are i.i.d. sub-Gaussian random variables independent of the designs; (ii) $B_k \in L_2(\mathbb{P}_Z)$ for all k, and $\mathbf{\Omega} := \mathbb{E}[(\mathbf{X} - \mathbf{B}(Z))(\mathbf{X} - \mathbf{B}(Z))^T]$ is positive definite; (iii) $\mathbf{X}_i$'s are uniformly bounded by a constant c_x.

Conditions in Assumption 3.1 are fairly standard in the literature. For example, the positive definiteness of $\mathbf{\Omega}$ is needed for obtaining semiparametric efficient estimation; see Mammen and van de Geer (1997). Note that we do not require the independence between $\mathbf{X}$ and Z throughout the paper.

ASSUMPTION 3.2 (Kernel Condition). We assume that there exist $0 < c_\phi < \infty$ and $0 < c_K < \infty$ such that $\sup_\ell \|\phi_\ell\|_{\sup} \le c_\phi$ and $\sup_z K(z, z) \le c_K$.

Assumption 3.2 is commonly assumed in kernel ridge regression literature [Guo (2002), Lafferty and Lebanon (2005), Zhang, Duchi and Wainwright (2013)]. In the case of finite rank kernel, for example, linear and polynomial kernels, the eigenfunctions are uniformly bounded as long as $\mathcal{Z}$ has finite support. As for the exponentially decaying kernels such as Gaussian kernel, we prove in Section 4.2 that the eigenfunctions given in (2.1) are uniformly bounded by 1.336. Finally, for the polynomially decaying kernels, Proposition 2.2 in Shang and Cheng (2013) showed that the eigenfunctions induced from a νth order Sobolev space (under a proper inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$) are uniformly bounded under mild smoothness conditions for the density of Z.

ASSUMPTION 3.3. For each $k = 1, \ldots, p$, $B_k(\cdot) \in \mathcal{H}$. This is equivalent to

$$\sum_{\ell=1}^{\infty} \mu_\ell^{-1} \langle B_k, \phi_\ell \rangle_{L_2(\mathbb{P}_Z)}^2 < \infty.$$

Assumption 3.3 requires the conditional expectation of X_k given $Z = z$ is as smooth as $f_0(z)$. As can be seen in Section 3.4, this condition is imposed to control the bias of the parametric component, which is caused by penalization on the nonparametric component. We call this interaction the "bias propagation phenomenon" and study it in Section 3.4.

Before laying out our main theoretical results, we define a key quantity used throughout the paper:

$$d(\lambda) := \sum_{\ell=1}^{\infty} \frac{1}{1 + \lambda/\mu_\ell}. \tag{3.4}$$

The quantity $d(\lambda)$ is essentially the "effective dimension," which was introduced in [Zhang (2005)]. For a finite dimensional space, $d(\lambda)$ corresponds to the true dimension, for example, $d(\lambda) \asymp r$ for the finite rank kernel (with rank r). For an infinite-dimensional space, $d(\lambda)$ is jointly determined by the size of that space and the smoothing parameter λ. For example, $d(\lambda) \asymp (-\log\lambda)^{1/p}$ for exponentially decaying kernel, and $d(\lambda) \asymp \lambda^{-1/(2m)}$ for polynomially decaying kernels. More details are provided in Section 4, where the three RKHS examples are carefully discussed.

In the end, we state a technical lemma that is crucially important in the subsequent theoretical derivations. For any function space $\mathcal{F}$, define an entropy integral

$$J(\mathcal{F}, \delta) =: \int_0^\delta \sqrt{\log \mathcal{N}\big(\mathcal{F}, \|\cdot\|_{\sup}, \epsilon\big)}\, d\epsilon,$$

where $\mathcal{N}(\mathcal{F}, \|\cdot\|_{\sup}, \epsilon)$ is an ϵ-covering number of $\mathcal{F}$ w.r.t. supreme norm. Define the following sets of functions: $\mathcal{F}_1 = \{f \mid f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$ for $\mathbf{x} \in \mathcal{X}, \boldsymbol{\beta} \in \mathbb{R}^p, \|f\|_{\sup} \le 1\}$, $\mathcal{F}_2 = \{f \in \mathcal{H} : \|f\|_{\sup} \le 1, \|f\|_{\mathcal{H}} \le d(\lambda)^{-1/2}\lambda^{-1/2}\}$, $\mathcal{F} := \{f = f_1 + f_2 : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2, \|f\|_{\sup} \le 1/2\}$.

LEMMA 3.1. *For any fixed $j = 1, \ldots, s$, we have*

$$\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} = \frac{1}{n}\sum_{i \in S_j} L_{U_i}\varepsilon_i - L_\lambda f_0 - \mathrm{Rem}_\beta^{(j)}, \tag{3.5}$$

and

$$\widehat{f}^{(j)} - f_0 = \frac{1}{n}\sum_{i \in S_j} N_{U_i}\varepsilon_i - N_\lambda f_0 - \mathrm{Rem}_f^{(j)}, \tag{3.6}$$

where $(\mathrm{Rem}_\beta^{(j)}, \mathrm{Rem}_f^{(j)}) \in \mathcal{A}$. *Moreover, suppose Assumptions 3.1 and 3.2 hold, and* $d(\lambda)n^{-1/2}(J(\mathcal{F}, 1) + \log n) = o(1)$, *then we have*

$$\mathbb{E}\big[\|\mathrm{Rem}_\beta^{(j)}\|_2^2\big] \le a(n, \lambda, J), \tag{3.7}$$

and

$$\mathbb{P}\big(\|\mathrm{Rem}_\beta^{(j)}\|_2 \ge b(n, \lambda, J)\big) \lesssim n\exp\big(-c\log^2 n\big), \tag{3.8}$$

where $a(n, \lambda, J) = Cd(\lambda)^2 n^{-1} r_{n,\lambda}^2 (J(\mathcal{F}, 1)^2 + 1) + Cd(\lambda)^2\lambda^{-1} n\exp(-c\log^2 n)$, *and* $b(n, \lambda, J) = Cd(\lambda)n^{-1/2} r_{n,\lambda}(J(\mathcal{F}, 1) + \log n)$, *with* $r_{n,\lambda} = (\log n)^2(d(\lambda)/n)^{1/2} + \lambda^{1/2}$. *The same inequalities also hold for* $\|\mathrm{Rem}_f^{(j)}\|_{\mathcal{C}}$ *under the same set of conditions.*

Equations (3.5) and (3.6) in the above lemma are fundamentally important in deriving the subsequent asymptotic and nonasymptotic results. In particular, by (3.6)

and the definition of $\bar{f}$, we obtain the stochastic expansion

$$\bar{f} - f_0 = \frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i - N_\lambda f_0 - \frac{1}{s}\sum_{j=1}^{s} \mathrm{Rem}_f^{(j)}, \tag{3.9}$$

which is the starting point for deriving the results in Theorems 3.2 and 3.4. These two equations are also of independent interest. For example, they trivially apply to the classical setup where there is only one dataset, that is, $s = 1$. In addition, they can be used in the other model settings where subpopulations do not share the same sample size or are not independent. As far as we know, the (nonasymptotic) moment and probability inequalities (3.7) and (3.8) on the remainder term are new. They are useful in determining a proper growth rate of s such that the aggregated remainder term $(1/s)\sum_{j=1}^{s} \mathrm{Rem}_f^{(j)}$ still vanishes in probability.

3.2. *Nonasymptotic bound for mean-squared error.* The primary goal of this section is to evaluate the estimation quality of the combined estimate from a *nonasymptotic* point of view. Specifically, we derive a finite sample upper bound for the mean-squared error $\mathrm{MSE}(\bar{f}) := \mathbb{E}[\|\bar{f} - f_0\|_{L_2(\mathbb{P}_Z)}^2]$. When s does not grow too fast, we show that $\mathrm{MSE}(\bar{f})$ is of the order $O(d(\lambda)/N + \lambda)$, from which the aggregation effect on f can be clearly seen. If λ is chosen in the order of N, the mean-squared error attains the (unimprovable) optimal minimax rate. As a by-product, we establish a *nonasymptotic* upper bound for the mean-squared error of $\widehat{\boldsymbol{\beta}}^{(j)}$, that is, $\mathrm{MSE}(\widehat{\boldsymbol{\beta}}^{(j)}) := \mathbb{E}[\|\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\|_2^2]$. The results in this section together with Theorem 3.6 in Section 3.4 determine an upper bound of s under which $\bar{f}$ *enjoys the same statistical properties (minimax optimality and asymptotic distribution) as the oracle estimate* $\widehat{f}_{\mathrm{or}}$.

Define $\tau_{\min}(\boldsymbol{\Omega})$ as the minimum eigenvalue of $\boldsymbol{\Omega}$ and $\mathrm{Tr}(K) := \sum_{\ell=1}^{\infty} \mu_\ell$ as the trace of K. Moreover, let $C_1' = 2\tau_{\min}^{-2}(\boldsymbol{\Omega})(c_x^2 p + c_\phi^2 \mathrm{Tr}(K)\sum_{k=1}^{p} \|B_k\|_{\mathcal{H}}^2)$, $C_1 = 2c_\phi^2(1 + C_1' \sum_{k=1}^{p} \|B_k\|_{L_2(\mathbb{P}_Z)}^2)$, $C_2' = \tau_{\min}^{-2}(\boldsymbol{\Omega})\|f_0\|_{\mathcal{H}}^2 \sum_{k=1}^{p} \|B_k\|_{\mathcal{H}}^2$ and $C_2 = 2C_2' \sum_{k=1}^{p} \|B_k\|_{L_2(\mathbb{P}_Z)}^2$.

THEOREM 3.2. *Under Assumptions 3.1–3.3, if* $s = o(Nd(\lambda)^{-2}(J(\mathcal{F}, 1) + \log N)^{-2})$, *then we have*

$$\mathrm{MSE}(\bar{f}) \le C_1\sigma^2 d(\lambda)/N + 2\|f_0\|_{\mathcal{H}}^2\lambda + C_2\lambda^2 + s^{-1}a(n, \lambda, J), \tag{3.10}$$

where $a(n, \lambda, J)$ *is defined in Lemma 3.1.*

Typically, we require an upper bound for s so that the fourth term in the RHS of (3.10) can be dominated by the first two terms, which correspond to variance and bias, respectively. To attain the optimal *bias-variance trade-off*, we choose $\lambda \asymp d(\lambda)/N$. Solving this equation yields the choice of regularization parameter λ,

which varies in different RKHS. The resulting rate of convergence for $\mathrm{MSE}(\bar{f})$ coincides with the minimax optimal rate of the oracle estimate in different RKHS; see Section 4. This can be viewed as a nonasymptotic version of the "oracle property" of $\bar{f}$. In comparison with the nonparametric KRR result in Zhang, Duchi and Wainwright (2013), we realize that adding one parametric component does not affect the finite sample upper bound (3.10).

As a by-product, we obtain a *nonasymptotic* upper bound for $\mathrm{MSE}(\widehat{\boldsymbol{\beta}}^{(j)})$. This result is new, and also of independent interest.

THEOREM 3.3. *Suppose that Assumptions 3.1–3.3 hold. Then we have*

$$\mathrm{MSE}\big(\widehat{\boldsymbol{\beta}}^{(j)}\big) \le C_1' \sigma^2 n^{-1} + C_2' \lambda^2 + a(n, \lambda, J), \tag{3.11}$$

where $a(n, \lambda, J)$ is defined in Lemma 3.1, and C_1' and C_2' are defined before Theorem 3.2.

Again, the first term and second term in the RHS of (3.11) correspond to the variance and bias, respectively. In particular, the second term comes from the bias propagation effect to be discussed in Section 3.4. By choosing $\lambda = o(n^{-1/2})$, we can obtain the optimal rate of $\mathrm{MSE}(\widehat{\boldsymbol{\beta}}^{(j)})$, that is, $O(n^{-1/2})$, but may lose the minimax optimality of $\mathrm{MSE}(\bar{f})$ in most cases.

3.3. *Joint asymptotic distribution.* In this section, we derive a preliminary result on the joint limit distribution of $(\widehat{\boldsymbol{\beta}}^{(j)}, \bar{f}(z_0))$ at any $z_0 \in \mathcal{Z}$. A key issue with this result is that their centering is not at the true value. However, we still choose to present it here since we will observe an interesting phenomenon when removing the bias in Section 3.4.

THEOREM 3.4 (Joint Asymptotics I). *Suppose that Assumptions 3.1 and 3.2 hold, and that as $N \to \infty$, $\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)}/d(\lambda)^{1/2} \to \sigma_{z_0}$, $(W_\lambda \mathbf{A})(z_0)/d(\lambda)^{1/2} \to \boldsymbol{\alpha}_{z_0} \in \mathbb{R}^p$, and $\mathbf{A}(z_0)/d(\lambda)^{1/2} \to -\boldsymbol{\gamma}_{z_0} \in \mathbb{R}^p$. Suppose the following conditions are satisfied*:

$$s = o\big(Nd(\lambda)^{-2}\big(J(\mathcal{F}, 1) + \log N\big)^{-2}\big), \tag{3.12}$$

$$sd(\lambda)/N \log^4 N + \lambda = o\big(d(\lambda)^{-2}\big(J(\mathcal{F}, 1) + \log N\big)^{-2} \log^{-2} N\big). \tag{3.13}$$

Denote $(\boldsymbol{\beta}_0^{(j)*}, f_0^*)$ as $(\mathrm{id} - P_\lambda) m_0^{(j)}$, where $m_0^{(j)} = (\boldsymbol{\beta}_0^{(j)}, f_0)$. We have for any $z_0 \in \mathcal{Z}$ and $j = 1, \ldots, s$:

(i) *if $s \to \infty$ then*

$$\begin{pmatrix} \sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\big) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0^*(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}\right), \tag{3.14}$$

where $\Sigma_{22} = \sigma_{z_0}^2 + 2\boldsymbol{\gamma}_{z_0}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\alpha}_{z_0} + \boldsymbol{\gamma}_{z_0}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\gamma}_{z_0}$;

(ii) *if s is fixed, then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}) \\ \sqrt{N/d(\lambda)}(\bar{f}(z_0) - f_0^*(z_0)) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & s^{-1/2}\boldsymbol{\Sigma}_{21} \\ s^{-1/2}\boldsymbol{\Sigma}_{12} & \Sigma_{22} \end{pmatrix}\right), \tag{3.15}$$

where $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12}^T = \boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}_{z_0} + \boldsymbol{\gamma}_{z_0})$.

Part (i) of Theorem 3.4 says that $\sqrt{n}\widehat{\boldsymbol{\beta}}^{(j)}$ and $\sqrt{N/d(\lambda)}\,\bar{f}(z_0)$ are asymptotically independent as $s \to \infty$. This is not surprising since only samples in one subpopulation (with size n) contribute to the estimation of the heterogeneity component while the entire sample (with size N) to commonality. As $n/N = s^{-1} \to 0$, the former data becomes asymptotically independent of (or asymptotically ignorable to) the latter data. So are these two estimators. The estimation bias $P_\lambda m_0^{(j)}$ can be removed by placing a smoothness condition on B_k, that is, Assumption 3.3. Interestingly, given this additional condition, even when s is fixed, these two estimators can still achieve the asymptotic independence if $d(\lambda) \to \infty$. Please see more details in next section.

The norming $d(\lambda)^{1/2}$ needed in these conditions $\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)}/d(\lambda)^{1/2} \to \sigma_{z_0}$, $(W_\lambda \mathbf{A})(z_0)/d(\lambda)^{1/2} \to \boldsymbol{\alpha}_{z_0} \in \mathbb{R}^p$, and $\mathbf{A}(z_0)/d(\lambda)^{1/2} \to -\boldsymbol{\gamma}_{z_0} \in \mathbb{R}^p$ is due to the following variance calculation:

$$\begin{aligned} &\operatorname{Var}\bigl(\sqrt{N/d(\lambda)}(\bar{f} - f_0^*)\bigr) \\ &\quad \approx \bigl\{\bigl[\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)}/d(\lambda)^{1/2}\bigr]^2 + 2\bigl[\mathbf{A}(z_0)/d(\lambda)^{1/2}\bigr]^T \boldsymbol{\Omega}^{-1}\bigl[W_\lambda \mathbf{A}(z_0)/d(\lambda)^{1/2}\bigr] \\ &\qquad + \bigl[\mathbf{A}(z_0)/d(\lambda)^{1/2}\bigr]^T \boldsymbol{\Omega}^{-1}\bigl[\mathbf{A}(z_0)/d(\lambda)^{1/2}\bigr]\bigr\}, \end{aligned}$$

where the first term is dominating and $\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)} = O(d(\lambda)^{1/2})$. So, by the norming $d(\lambda)^{1/2}$, we obtain the order of $\sqrt{N/d(\lambda)}(\bar{f} - f_0^*)$ as $O_{\mathbb{P}}(1)$.

Our last result in this section is the joint asymptotic distribution of $\{\widehat{\boldsymbol{\beta}}^{(j)}\}_{j=1}^s$ (expressed in a linear contrast form). Denote

$$\underset{\sim}{\widehat{\boldsymbol{\beta}}} = (\widehat{\boldsymbol{\beta}}^{(1)T}, \ldots, \widehat{\boldsymbol{\beta}}^{(s)T})^T \in \mathbb{R}^{ps} \quad \text{and} \quad \underset{\sim}{\boldsymbol{\beta}}_0 = (\boldsymbol{\beta}_0^{(1)T}, \ldots, \boldsymbol{\beta}_0^{(s)T})^T \in \mathbb{R}^{ps}.$$

THEOREM 3.5. *Suppose Assumptions 3.1–3.3 hold. If $\lambda = o(N^{-1/2})$, and s satisfies (3.12) and*

$$s^2 d(\lambda)/N \log^4 N + s\lambda = o\bigl(d(\lambda)^{-2}\bigl(J(\mathcal{F}, 1) + \log N\bigr)^{-2} \log^{-2} N\bigr), \tag{3.16}$$

then for any $\underset{\sim}{\mathbf{u}} = (\mathbf{u}_1^T, \ldots, \mathbf{u}_s^T) \in \mathbb{R}^{ps}$ with $\|\underset{\sim}{\mathbf{u}}\|_2 = 1$, it holds

$$\sqrt{n} V_s^{-1} \underset{\sim}{\mathbf{u}}^T (\underset{\sim}{\widehat{\boldsymbol{\beta}}} - \underset{\sim}{\boldsymbol{\beta}}_0) \rightsquigarrow N(0, \sigma^2),$$

where $V_s^2 = \sum_{j=1}^s \mathbf{u}_j^T \boldsymbol{\Omega}^{-1} \mathbf{u}_j$, as $N \to \infty$.

Note that the upper bound condition on s is slightly different from that in Theorem 3.4.

3.4. *Bias propagation*. In this section, we first analyze the source of estimation bias observed in the joint asymptotics Theorem 3.4. In fact, these analysis leads to a bias propagation phenomenon, which intuitively explains how Assumption 3.3 removes the estimation bias. More importantly, we show that $\bar{f}$ shares exactly the same asymptotic distribution as $\widehat{f}_{\mathrm{or}}$, that is, oracle rule, when s does not grow too fast and λ is chosen in the order of N.

Our study on propagation mechanism is motivated by the following simple observation. Denote $\mathbb{X} \in \mathbb{R}^{n\times p}$ and $\mathbb{Z} \in \mathbb{R}^n$ as the designs based on the samples from the jth subpopulation and let $\boldsymbol{\varepsilon}^{(j)} = [\varepsilon_i]_{i\in S_j} \in \mathbb{R}^n$. The first order optimality condition (w.r.t. $\boldsymbol{\beta}$) gives

$$\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} = \big(\mathbb{X}^T\mathbb{X}\big)^{-1}\mathbb{X}^T\boldsymbol{\varepsilon}^{(j)} - \big(\mathbb{X}^T\mathbb{X}\big)^{-1}\mathbb{X}^T\big(\widehat{f}^{(j)}(\mathbb{Z}) - f_0(\mathbb{Z})\big), \tag{3.17}$$

where $f_0(\mathbb{Z})$ is a n-dimensional vector with entries $f_0(Z_i)$ for $i \in S_j$ and $\widehat{f}^{(j)}(\mathbb{Z})$ is defined similarly. Hence, the estimation bias of $\widehat{\boldsymbol{\beta}}^{(j)}$ inherits from that of $\widehat{f}^{(j)}$. A more complete picture on the propagation mechanism can be seen by decomposing the total bias $P_\lambda m_0^{(j)}$ into two parts:

$$\text{parametric bias:}\quad L_\lambda f_0 = -(\boldsymbol{\Omega} + \boldsymbol{\Sigma}_\lambda)^{-1}\langle \mathbf{B}, W_\lambda f_0\rangle_{L_2(\mathbb{P}_Z)}, \tag{3.18}$$

$$\text{nonparametric bias:}\quad N_\lambda f_0 = W_\lambda f_0 - \mathbf{A}^T L_\lambda f_0 \tag{3.19}$$

according to Proposition 2.3. The first term in (3.19) explains the bias introduced by penalization; see (2.5). This bias propagates to the parametric component through $\mathbf{B}$, as illustrated in (3.18). The parametric bias $L_\lambda f_0$ propagates back to the nonparametric component through the second term of (3.19). Therefore, by strengthening $B_k \in L_2(\mathbb{P}_Z)$ to $B_k \in \mathcal{H}$, that is, Assumption 3.3, it can be shown that the order of $L_\lambda f_0$ in (3.18) reduces to that of λ. And then we can remove $L_\lambda f_0$ asymptotically by choosing a sufficiently small λ. In this case, the nonparametric bias becomes $W_\lambda f_0$.

We summarize the above discussions in the following theorem.

THEOREM 3.6 (Joint Asymptotics II). *Suppose Assumption 3.3 and the conditions in Theorem 3.4 hold. If we choose* $\lambda = o(\sqrt{d(\lambda)/N} \wedge n^{-1/2})$, *then*:

(i) *if* $s \to \infty$ *then*

$$\begin{pmatrix} \sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\big) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0(z_0) - W_\lambda f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2\begin{pmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^* \end{pmatrix}\right), \tag{3.20}$$

where $\Sigma_{22}^* = \sigma_{z_0}^2 + \boldsymbol{\gamma}_{z_0}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}_{z_0}$;

(ii) *if s is fixed, then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \\ \sqrt{N/d(\lambda)}(\bar{f}(z_0) - f_0(z_0) - W_\lambda f_0(z_0)) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & s^{-1/2}\boldsymbol{\Sigma}_{21}^* \\ s^{-1/2}\boldsymbol{\Sigma}_{12}^* & \Sigma_{22}^* \end{pmatrix}\right), \tag{3.21}$$

where $\boldsymbol{\Sigma}_{12}^* = \boldsymbol{\Sigma}_{12}^{*T} = \boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}_{z_0}$ and Σ_{22}^* is the same as in (i).

Moreover, if $d(\lambda) \to \infty$, then $\boldsymbol{\Sigma}_{12}^* = \boldsymbol{\Sigma}_{21}^* = \mathbf{0}$ and $\Sigma_{22}^* = \sigma_{z_0}^2$ in (i) *and* (ii).

The nonparametric estimation bias $W_\lambda f_0(z_0)$ can be further removed by performing undersmoothing, a standard procedure in nonparametric inference; see, for example, Shang and Cheng (2013). We will illustrate this point in Section 4.

By examining the proof for case (ii) of Theorem 3.6 (and taking $s = 1$), we know that the oracle estimate $\widehat{f}_{\mathrm{or}}$ defined in (3.1) attains the same asymptotic distribution as that of $\bar{f}$ in (3.20) when s grows at a proper rate. Therefore, we claim that our combined estimate $\bar{f}$ satisfies the desirable oracle property.

In Section 4, we apply Theorem 3.6 to several examples, and find that even though the minimization (3.2) is based only on one fraction of the entire sample, it is nonetheless essential to regularize each subestimation as if it had the entire sample. In other words, λ should be chosen in the order of N. Similar phenomenon also arises in analyzing minimax optimality of each subestimation; see Section 3.2.

When $s = 1$, our model reduces to the standard partially linear model. The joint distribution of parametric and nonparametric estimators is shown in Part (ii) of Theorem 3.6, where their asymptotic covariance is derived as $\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}_{z_0}$ with $\boldsymbol{\gamma}_{z_0} = \lim_{N\to\infty} -\mathbf{A}(z_0)/d(\lambda)^{1/2}$. In Section 7.5, we show that $\mathbf{A}(z_0)$ is bounded for any $z_0 \in \mathcal{Z}$ (uniformly over λ). Hence, the asymptotic correlation disappears, that is, $\boldsymbol{\gamma}_{z_0} = \mathbf{0}$, as $d(\lambda) \to \infty$. This corresponds to the exponentially decaying kernel and polynomially decaying kernel. In fact, this finding generalizes the *joint asymptotics phenomenon* recently discovered for partial smoothing spline models; see Cheng and Shang (2015). However, when $d(\lambda)$ is finite, for example, finite rank kernel, the asymptotic correlation remains. This is not surprising since the semiparametric estimation in this case essentially reduces to a parametric one.

REMARK 3.1. Theorem 3.6 implies that $\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \rightsquigarrow N(\mathbf{0}, \sigma^2\boldsymbol{\Omega}^{-1})$ when $\lambda = o(n^{-1/2})$. When the error ε follows a Gaussian distribution, it is well known that $\widehat{\boldsymbol{\beta}}^{(j)}$ achieves the semiparametric efficiency bound [Kosorok (2008)]. Hence, the semiparametric efficient estimate can be obtained by applying the kernel ridge method. However, we can further improve its estimation efficiency to a parametric level by taking advantage of $\bar{f}$ (built on the whole samples). This represents an important feature of massive data: strength-borrowing.

REMARK 3.2. We can also construct a simultaneous confidence band for f_0 based on the stochastic expansion of $\bar{f}$ and strong approximation techniques [Bickel and Rosenblatt (1973)]. Specifically, we start from (3.9) that implies

$$\left\| \bar{f} - f_0^* - \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i N_{U_i} \right\|_{\sup} = \left\| \frac{1}{s} \sum_{j=1}^{s} \mathrm{Rem}_f^{(j)} \right\|_{\sup}. \tag{3.22}$$

Similar to the pointwise case, we can show that the remainder term on the RHS of (3.22) is $o_P(1)$ once s does not grow too fast. Hence, the distribution of $\sup_z |\bar{f}(z) - f_0^*(z)|$ can be approximated by that of $\sup_z |N^{-1} \sum_{i=1}^{N} \varepsilon_i N_{U_i}(z)|$. We next apply strong approximation techniques to prove that $N^{-1} \sum_{i=1}^{N} \varepsilon_i N_{U_i}$ can be further approximated by a proper Gaussian process. This would yield a simultaneous confidence band. More rigorous arguments can be adapted from the proof of Theorem 5.1 in Shang and Cheng (2013).

3.5. *Efficiency boosting*: *From semiparametric level to parametric level.* The previous sections show that the combined estimate $\bar{f}$ achieves the "oracle property" in both asymptotic and nonasymptotic senses when s does not grow too fast and λ is chosen according to the entire sample size. In this section, we further employ $\bar{f}$ to boost the estimation efficiency of $\widehat{\boldsymbol{\beta}}^{(j)}$ from semiparametric level to parametric level. This leads to our final estimate for heterogeneity, that is, $\check{\boldsymbol{\beta}}^{(j)}$ defined in (3.23). More importantly, $\check{\boldsymbol{\beta}}^{(j)}$ possesses the limit distribution as if the commonality in each subpopulation were known, and hence satisfies the "oracle rule." This interesting efficiency boosting phenomenon will be empirically verified in Section 6. A similar two-stage estimation method was proposed in Li and Liang (2008), but for the purpose of variable selection in $\boldsymbol{\beta}$ based on a single data set.

Specifically, we define the following improved estimator for $\boldsymbol{\beta}_0$:

$$\check{\boldsymbol{\beta}}^{(j)} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i \in S_j} \big(Y_i - \mathbf{X}_i^T \boldsymbol{\beta} - \bar{f}(Z_i)\big)^2. \tag{3.23}$$

Theorem 3.7 below shows that $\check{\boldsymbol{\beta}}^{(j)}$ achieves the parametric efficiency bound as if the nonparametric component f were known. This is not surprising given that the nonparametric estimate $\bar{f}$ now possesses a faster convergence rate after aggregation. What is truly interesting is that we need to set a lower bound for s, that is, (3.24), which slows down the convergence rate of $\check{\boldsymbol{\beta}}^{(j)}$, that is, $\sqrt{n}$, such that $\bar{f}$ can be treated as if it were known. Note that the homogeneous data setting is trivially excluded in this case.

THEOREM 3.7. *Suppose Assumptions* 3.1 *and* 3.2 *hold. If s satisfies conditions* (3.12), (3.13) *and*

$$s^{-1} = o\big(d(\lambda)^{-2} \log^{-4} N\big), \tag{3.24}$$

and we choose $\lambda = o(d(\lambda)/N)$, *then we have*

$$\sqrt{n}(\check{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \rightsquigarrow N(0, \sigma^2 \boldsymbol{\Sigma}^{-1}),$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{X}\mathbf{X}^T]$.

Note that $\mathbf{X}$ and Z are not assumed to be independent. Hence, the parametric efficiency bound $\boldsymbol{\Sigma}^{-1}$ is not larger than the semiparametric efficiency bound $\boldsymbol{\Omega}^{-1}$. The intuition for this lower bound of s is that the total sample size N should grow much faster than the subsample size n, so that the nonparametric estimator $\bar{f}$ converges faster than the parametric estimator $\check{\boldsymbol{\beta}}^{(j)}$. In this case, the error of estimating f_0 is negligible so that $\check{\boldsymbol{\beta}}^{(j)}$ behaves asymptotically as if f were known, resulting in *parametric* efficiency.

3.6. *Testing for heterogeneity.* The heterogeneity across different sub-populations is a crucial feature of massive data. However, there is still some chance that some subpopulations may share the same underlying distribution. In this section, we consider testing for the heterogeneity among subpopulations. We start from a simple pairwise testing, and then extend it to a more challenging simultaneous testing that can be applied to a large number of subpopulations.

Consider a general class of pairwise heterogeneity testing:

$$H_0 : Q(\boldsymbol{\beta}_0^{(j)} - \boldsymbol{\beta}_0^{(k)}) = \mathbf{0} \qquad \text{for } j \neq k, \tag{3.25}$$

where $Q = (Q_1^T, \ldots, Q_q^T)^T$ is a $q \times p$ matrix with $q \leq p$. The general formulation (3.25) can test either the whole vector or one fraction of $\boldsymbol{\beta}_0^{(j)}$ is equal to that of $\boldsymbol{\beta}_0^{(k)}$. A test statistic can be constructed based on either $\widehat{\boldsymbol{\beta}}$ or its improved version $\check{\boldsymbol{\beta}}$. Let $C_\alpha \subset \mathbb{R}^q$ be a confidence region satisfying $\mathbb{P}(\mathbf{b} \in C_\alpha) = 1 - \alpha$ for any $\mathbf{b} \sim N(0, I_q)$. Specifically, we have the following α-level Wald tests:

$$\Psi_1 = I\{Q(\widehat{\boldsymbol{\beta}}^{(j)} - \widehat{\boldsymbol{\beta}}^{(k)}) \notin \sqrt{2/n}\sigma(Q\boldsymbol{\Omega}^{-1}Q^T)^{1/2}C_\alpha\},$$

$$\Psi_2 = I\{Q(\check{\boldsymbol{\beta}}^{(j)} - \check{\boldsymbol{\beta}}^{(k)}) \notin \sqrt{2/n}\sigma(Q\boldsymbol{\Sigma}^{-1}Q^T)^{1/2}C_\alpha\}.$$

The consistency of the above tests are guaranteed by Theorem 3.8 below. In addition, we note that the power of the latter test is larger than the former; see the analysis below Theorem 3.8. The price we need to pay for this larger power is to require a lower bound on s.

THEOREM 3.8. *Suppose that the conditions in Theorem 3.6 are satisfied. Under the null hypothesis specified in* (3.25), *we have*

$$\sqrt{n}Q(\widehat{\boldsymbol{\beta}}^{(j)} - \widehat{\boldsymbol{\beta}}^{(k)}) \rightsquigarrow N(\mathbf{0}, 2\sigma^2 Q\boldsymbol{\Omega}^{-1}Q^T).$$

Moreover, under the conditions in Theorem 3.7, we have

$$\sqrt{n}Q(\check{\boldsymbol{\beta}}^{(j)} - \check{\boldsymbol{\beta}}^{(k)}) \rightsquigarrow N(\mathbf{0}, 2\sigma^2 Q\boldsymbol{\Sigma}^{-1}Q^T),$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\mathbf{X}\mathbf{X}^T]$.

The larger power of Ψ_2 is due to the smaller asymptotic variance of $\check{\boldsymbol{\beta}}^{(j)}$, and can be deduced from the following power function. For simplicity, we consider $H_0 : \beta_{01}^{(j)} - \beta_{01}^{(k)} = 0$, that is, $Q = (1, 0, 0 \ldots, 0)$. In this case, we have $\Psi_1 = I\{|\widehat{\beta}_1^{(j)} - \widehat{\beta}_1^{(k)}| > \sqrt{2}\sigma[\boldsymbol{\Omega}^{-1}]_{11}^{1/2} z_{\alpha/2}/\sqrt{n}\}$, and $\Psi_2 = I\{|\check{\beta}_1^{(j)} - \check{\beta}_1^{(k)}| > \sqrt{2}\sigma[\boldsymbol{\Sigma}^{-1}]_{11}^{1/2} z_{\alpha/2}/\sqrt{n}\}$. The (asymptotic) power function under the alternative that $\beta_{01}^{(j)} - \beta_{01}^{(k)} = \beta^*$ for some nonzero β^* is

$$\text{Power}(\beta^*) = 1 - \mathbb{P}\left(W \in \left[-\frac{\beta^*\sqrt{n}}{\sigma^*} \pm z_{\alpha/2}\right]\right),$$

where $W \sim N(0, 1)$ and σ^* is $\sqrt{2}\sigma[\boldsymbol{\Omega}^{-1}]_{11}^{1/2}$ for Ψ_1 and $\sqrt{2}\sigma[\boldsymbol{\Sigma}^{-1}]_{11}^{1/2}$ for Ψ_2. Hence, a smaller σ^* gives rise to a larger power, and Ψ_2 is more powerful than Ψ_1. Please see Section 6 for empirical support for this power comparison.

We next consider the problem of heterogeneous testing for a large number of subpopulations:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \qquad \text{for all } j \in \mathcal{G}, \tag{3.26}$$

where $\mathcal{G} \subset \{1, 2, \ldots, s\}$, versus the alternative:

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \qquad \text{for some } j \in \mathcal{G}. \tag{3.27}$$

The above $\widetilde{\boldsymbol{\beta}}^{(j)}$'s are pre-specified for each $j \in \mathcal{G}$. If all $\widetilde{\boldsymbol{\beta}}^{(j)}$'s are the same, then it becomes a type of heterogeneity test for the group of subpopulations indexed by $\mathcal{G}$. Here, we allow $|\mathcal{G}|$ to be as large as s, and thus it can increase with n. Let $\widehat{\boldsymbol{\Sigma}}^{(j)}$ be the sample covariance matrix of $\mathbf{X}$ for the jth subpopulation, that is, $n^{-1}\sum_{i \in S_j} \mathbf{X}_i\mathbf{X}_i^T$. Define the test statistic

$$T_{\mathcal{G}} := \max_{j \in \mathcal{G}, 1 \le k \le p} \sqrt{n}(\check{\boldsymbol{\beta}}_k^{(j)} - \widetilde{\boldsymbol{\beta}}_k^{(j)}).$$

We approximate the distribution of the above test statistic using multiplier bootstrap. Define the following quantity:

$$W_{\mathcal{G}} := \max_{j \in \mathcal{G}, 1 \le k \le p} \frac{1}{\sqrt{n}} \sum_{i \in S_j} (\widehat{\boldsymbol{\Sigma}}^{(j)})_k^{-1} \mathbf{X}_i e_i,$$

where e_i's are i.i.d. $N(0, \sigma^2)$ independent of the data and $(\widehat{\boldsymbol{\Sigma}}^{(j)})_k^{-1}$ is the kth row of $(\widehat{\boldsymbol{\Sigma}}^{(j)})^{-1}$. Let $c_{\mathcal{G}}(\alpha) = \inf\{t \in \mathbb{R} : \mathbb{P}(W_{\mathcal{G}} \le t \mid \mathbb{X}) \ge 1 - \alpha\}$. We employ the recent Gaussian approximation and multiplier bootstrap theory [Chernozhukov, Chetverikov and Kato (2013)] to obtain the following theorem.

THEOREM 3.9. *Suppose Assumptions 3.1 and 3.2 hold. In addition, suppose* (3.12) *and* (3.13) *in Theorem 3.4 hold. For any* $\mathcal{G} \subset \{1, 2, \ldots, s\}$ *with* $|\mathcal{G}| = d$, *if* (i) $s \gtrsim d(\lambda)^2 \log(pd) \log^4 N$, (ii) $(\log(pdn))^7/n \le C_1 n^{-c_1}$ *for some constants* $c_1, C_1 > 0$, *and* (iii) $p^2 \log(pd)/\sqrt{n} = o(1)$, *then under* H_0 *and choosing* $\lambda = o(d(\lambda)/N)$, *we have*

$$\sup_{\alpha \in (0,1)} \left| \mathbb{P}\big(T_{\mathcal{G}} > c_{\mathcal{G}}(\alpha)\big) - \alpha \right| = o(1).$$

REMARK 3.3. We can perform heterogeneity testing even without specifying $\widetilde{\boldsymbol{\beta}}^{(j)}$'s. This can be done by simply reformulating the null hypothesis as follows (for simplicity we set $\mathcal{G} = [s]$): $H_0 : \alpha^{(j)} = 0$ for $j \in [s-1]$, where $\alpha^{(j)} = \boldsymbol{\beta}^{(j)} - \boldsymbol{\beta}^{(j+1)}$ for $j = 1, \ldots, s-1$. The test statistic is $T'_{\mathcal{G}} = \max_{1 \le j \le s-1} \max_{1 \le k \le p} \alpha_k^{(j)}$. The bootstrap quantity is defined as

$$W'_{\mathcal{G}} := \max_{1 \le j \le s-1, 1 \le k \le p} \frac{1}{\sqrt{n}} \sum_{i \in S_j} \big(\widehat{\boldsymbol{\Sigma}}^{(j)}\big)_k^{-1} \mathbf{X}_i e_i - \frac{1}{\sqrt{n}} \sum_{i \in S_{j+1}} \big(\widehat{\boldsymbol{\Sigma}}^{(j+1)}\big)_k^{-1} \mathbf{X}_i e_i.$$

The proof is similar to that of Theorem 3.9 and is omitted.

4. Examples. In this section, we consider three specific classes of RKHS with different smoothness, characterized by the decaying rate of the eigenvalues: finite rank, exponential decay and polynomial decay. In particular, we give explicit upper bounds for s under which the combined estimate enjoys the oracle property, and also explicit lower bounds for obtaining efficiency boosting studied in Section 3.5. Interestingly, we find that the upper bound for s increases for RKHS with faster decaying eigenvalues. Hence, our aggregation procedure favors smoother regression functions in the sense that more subpopulations are allowed to be included in the observations. The choice of λ is also explicitly characterized in terms of the entire sample size and the decaying rate of eigenvalues. In all three examples, the undersmoothing is implicitly assumed for removing the nonparametric estimation bias. Our bounds on s and λ here are not the most general ones, but are those that can easily deliver theoretical insights.

4.1. *Example I*: *Finite rank kernel.* The RKHS with finite rank kernels includes linear functions, polynomial functions, and, more generally, functional classes with finite dictionaries. In this case, the effective dimension is simply proportional to the rank r. Hence, $d(\lambda) \asymp r$. Combining this fact with Theorem 3.6, we get the following corollary for finite rank kernels.

COROLLARY 4.1. *Suppose Assumptions 3.1–3.3 hold and* $s \to \infty$. *For any* $z_0 \in \mathcal{Z}$, *if* $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(N^2 \log^{-12} N)$ *and* $s = o(\frac{N}{\sqrt{\log \lambda^{-1}} \log^6 N})$, *then*

$$\begin{pmatrix} \sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\big) \\ \sqrt{N}\big(\bar{f}(z_0) - f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^* \end{pmatrix}\right),$$

where $\Sigma_{22}^* = \sum_{\ell=1}^r \phi_\ell(z_0)^2 + \boldsymbol{\gamma}_{z_0}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\gamma}_{z_0}$ *and* $\boldsymbol{\gamma}_{z_0} = \sum_{\ell=1}^r \langle \mathbf{B}, \phi_\ell \rangle_{L_2(\mathbb{P}_Z)} \phi_\ell(z_0)$.

From the above corollary, we can easily tell that the upper bound for s can be as large as $o(N \log^{-7} N)$ by choosing a sufficiently large λ. Hence, s can be chosen nearly as large as N. As for the lower bound of s for boosting the efficiency, we have $s \gtrsim r^2 \log^4 N$ by plugging $d(\lambda) \asymp r$ into (3.24). This lower bound is clearly smaller than the upper bound. Hence, the efficiency boosting is feasible.

Corollary 4.2 below specifies conditions on s and λ under which $\bar{f}$ achieves the nonparametric minimaxity.

COROLLARY 4.2. *Suppose that Assumptions 3.1–3.3 hold. When* $\lambda = r/N$ *and* $s = o(N \log^{-5} N)$, *we have*

$$\mathbb{E}\big[\|\bar{f} - f_0\|_{L_2(\mathbb{P}_Z)}^2\big] \le Cr/N,$$

for some constant C.

4.2. *Example II*: *Exponential decay kernel*. We next consider the RKHS for which the kernel has exponentially decaying eigenvalues, that is, $\mu_\ell = \exp(-\alpha \ell^p)$ for some $\alpha > 0$. In this case, we have $d(\lambda) \asymp (\log \lambda^{-1})^{1/p}$ by explicit calculations.

COROLLARY 4.3. *Suppose Assumptions 3.1–3.3 hold, and for any* $z_0 \in \mathcal{Z}$, $f_0 \in \mathcal{H}$ *satisfies* $\sum_{\ell=1}^\infty |\phi_\ell(z_0) \langle f_0, \phi_\ell \rangle_{\mathcal{H}}| < \infty$. *If* $\lambda = o(N^{-1/2} \log^{1/(2p)} N \wedge n^{-1/2})$, $\log(\lambda^{-1}) = o(N^{p/(p+4)} \log^{-6p/(p+4)} N)$ *and* $s = o(\frac{N}{\log^6 N \log^{(p+4)/p}(\lambda^{-1})})$, *then*

$$\begin{pmatrix} \sqrt{n}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_{z_0}^2 \end{pmatrix}\right),$$

where $\sigma_{z_0}^2 = \lim_{\lambda \to 0} d(\lambda)^{-1} \sum_{\ell=1}^\infty \frac{\phi_\ell^2(z_0)}{(1+\lambda/\mu_\ell)^2}$.

Corollary 4.3 implies the shrinking rate of the confidence interval for $f_0(z_0)$ as $\sqrt{d(\lambda)/N}$. This motivates us to choose λ [equivalently $d(\lambda)$] as large as possible (as small as possible). Plugging such a λ into the upper bound of s yields $s = o(N \log^{-(7p+4)/p} N)$. For example, when $p = 1$ ($p = 2$), the upper bound is $s = o(N \log^{-11} N)$($s = o(N \log^{-9} N)$). Note that this upper bound for s only differs from that for the finite rank kernel up to some logarithmic term. This is mainly because RKHS with exponentially decaying eigenvalues has an effective dimension $(\log N)^{1/p}$ (for the above λ). Again, by (3.24) we get the lower bound of $s \gtrsim (\log \lambda^{-1})^{2/p} \log^2 N$. When $\lambda \asymp N^{-1/2} \log^{1/(2p)} N \wedge n^{-1/2}$, it is approximately $s \gtrsim \log^{(4p+2)/p} N$.

As a concrete example, we consider the Gaussian kernel $K(z_1, z_2) = \exp(-|z_1 - z_2|^2/2)$. The eigenfunctions are given in (2.1), and the eigenvalues

are exponentially decaying, as $\mu_\ell = \eta^{2\ell+1}$, where $\eta = (\sqrt{5}-1)/2$. According to Krasikov (2004), we can get that

$$c_\phi = \sup_{\ell\in\mathbb{N}} \|\phi_\ell\|_{\sup} \le \frac{2e^{15/8}(\sqrt{5}/4)^{1/4}}{3\sqrt{2\pi}2^{1/6}} \le 1.336.$$

Thus, Assumption 3.2 is satisfied. We next give an upper bound of $\sigma_{z_0}^2$ in Corollary 4.3 as follows:

$$\begin{aligned}\sigma_{z_0}^2 &\le \lim_{N\to\infty} \sigma^2 c_\phi^2 h \sum_{\ell=0}^{\infty} \bigl(1+\lambda\eta\exp\bigl(-2(\log\eta)\ell\bigr)\bigr)^{-2} = (1/2)c_\phi^2\sigma^2\log^{-1}(1/\eta)\\ &\le 4.27\sigma^2,\end{aligned}$$

where equality follows from Lemma C.1 in Appendix C [Zhao, Cheng and Liu (2015)] for the case $t=2$. Hence, a (conservative) $100(1-\alpha)\%$ confidence interval for $f_0(z_0)$ is given by $\bar{f}(z_0) \pm 1.3106\sigma z_{\alpha/2}\sqrt{d(\lambda)/N}$.

COROLLARY 4.4. *Suppose that Assumptions 3.1–3.3 hold. By choosing* $\lambda = (\log N)^{1/p}/N$ *and* $s = o(N\log^{-(5p+3)/p} N)$, *we have*

$$\mathbb{E}\bigl[\|\bar{f} - f_0\|_{L_2(\mathbb{P}_Z)}^2\bigr] \le C(\log N)^{1/p}/N.$$

We know that the above rate is minimax optimal according to Zhang, Duchi and Wainwright (2013). Note that the upper bound for s required here is similar to that for obtaining the joint limiting distribution in Corollary 4.3.

4.3. *Example III*: *Polynomial decay kernel*. We now consider the RKHS for which the kernel has polynomially decaying eigenvalues, that is, $\mu_\ell = c\ell^{-2\nu}$ for some $\nu > 1/2$. Hence, we can explicitly calculate that $d(\lambda) = \lambda^{-1/(2\nu)}$. The resulting penalized estimate is called as "partial smoothing spline" in the literature; see Gu (2013), Wang (2011).

COROLLARY 4.5. *Suppose Assumptions 3.1–3.3 hold, and* $\sum_{\ell=1}^{\infty} |\phi_\ell(z_0)\langle f_0, \phi_\ell\rangle_{\mathcal{H}}| < \infty$ *for any* $z_0 \in \mathcal{Z}$ *and* $f_0 \in \mathcal{H}$. *For any* $\nu > 1+\sqrt{3}/2 \approx 1.866$, *if* $\lambda \asymp N^{-d}$ *for some* $\frac{2\nu}{4\nu+1} < d < \frac{4\nu^2}{10\nu-1}$, $\lambda = o(n^{-1/2})$ *and* $s = o(\lambda^{(10\nu-1)/(4\nu^2)} N\log^{-6} N)$, *then*

$$\begin{pmatrix} \sqrt{n}\bigl(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\bigr) \\ \sqrt{N/d(\lambda)}\bigl(\bar{f}(z_0) - f_0(z_0)\bigr)\end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2\begin{pmatrix}\boldsymbol{\Omega}^{-1} & \mathbf{0}\\ \mathbf{0} & \sigma_{z_0}^2\end{pmatrix}\right),$$

where $\sigma_{z_0}^2 = \lim_{\lambda\to 0} d(\lambda)^{-1}\sum_{\ell=1}^{\infty} \frac{\phi_\ell^2(z_0)}{(1+\lambda/\mu_\ell)^2}$.

Similarly, we choose $\lambda \asymp N^{-2\nu/(4\nu+1)} \wedge n^{-1/2}$ to get the fastest shrinking rate of the confidence interval. Plugging the above λ into the upper bound for s, we get

$$s = o\big(N^{(8\nu^2-8\nu+1)/(2\nu(4\nu+1))} \log^{-6} N \wedge N(\log N)^{-(48\nu^2)/(8\nu^2+10\nu+1)}\big).$$

When N is large, the above bound reduces to $s = o(N^{(8\nu^2-8\nu+1)/(2\nu(4\nu+1))} \times \log^{-6} N)$. We notice that the upper bound for s increases as ν increases, indicating that the aggregation procedure favors smoother functions. As an example, for the case that $\nu = 2$, we have the upper bound for $s = o(N^{17/36} \log^{-6} N) \approx o(N^{0.47} \log^{-6} N)$. Again, we obtain the lower bound $s \gtrsim \lambda^{-1/\nu} \log^4 N$ by plugging $d(\lambda) \asymp \lambda^{-1/(2\nu)}$ into (3.24). When $\lambda \asymp N^{-2\nu/(4\nu+2)}$, we get $s \gtrsim N^{1/(4\nu+1)} \log^2 N$. For $\nu = 2$, this is approximately $s \gtrsim N^{0.22} \log^4 N$.

As a concrete example, we consider the periodic Sobolev space $H_0^\nu[0,1]$ with the following eigenfunctions:

$$\phi_\ell(x) = \begin{cases} 1, & \ell = 0, \\ \sqrt{2}\cos(\ell\pi x), & \ell = 2k \text{ for } k = 1, 2, \ldots, \\ \sqrt{2}\sin\big((\ell+1)\pi x\big), & \ell = 2k-1 \text{ for } k = 1, 2, \ldots, \end{cases} \tag{4.1}$$

and eigenvalues

$$\mu_\ell = \begin{cases} \infty, & \ell = 0, \\ (\ell\pi)^{-2\nu}, & \ell = 2k, \text{ for } k = 1, 2, \ldots, \\ \big((\ell+1)\pi\big)^{-2\nu}, & \ell = 2k-1, \text{ for } k = 1, 2, \ldots. \end{cases} \tag{4.2}$$

Hence, Assumption 3.2 trivially holds. Under the above eigen-system, the following lemma gives an explicit expression of $\sigma_{z_0}^2$.

LEMMA 4.1. *Under the eigen-system defined by* (4.1) *and* (4.2), *we can explicitly calculate*

$$\sigma_{z_0}^2 = \lim_{\lambda\to 0} d(\lambda)^{-1} \sum_{\ell=1}^{\infty} \frac{\phi_\ell^2(z_0)}{(1+\lambda/\mu_\ell)^2} = \int_0^\infty \frac{1}{(1+x^{2\nu})^2}\,dx = \frac{\pi}{2\nu\sin(\pi/(2\nu))}.$$

Therefore, by Corollary 4.5, we have that when $\lambda \asymp N^{-2\nu/(4\nu+1)}$ and $s = o(N^{(8\nu^2-8\nu+1)/(2\nu(4\nu+1))} \log^{-6} N)$,

$$\begin{pmatrix} \sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\big) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^2 \begin{pmatrix} \boldsymbol{\Omega}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_{z_0}^2 \end{pmatrix}\right), \tag{4.3}$$

where $\sigma_{z_0}^2$ is given in Lemma 4.1. When $\nu = 2$, $\lambda \asymp N^{-4/9}$ and the upper bound for $s = o(N^{17/36} \log^{-6} N)$.

COROLLARY 4.6. *Suppose that Assumptions* 3.1–3.3 *hold. If we choose* $\lambda = N^{-2\nu/(2\nu+1)}$, *and* $s = o(N^{(4\nu^2-4\nu+1)/(4\nu^2+2\nu)} \log^{-4} N)$, *the combined estimator achieves optimal rate of convergence, that is*,

$$\mathbb{E}\big[\|\bar{f} - f_0\|_{L_2(\mathbb{P}_Z)}^2\big] \le C N^{-2\nu/(2\nu+1)}. \tag{4.4}$$

The above rate is known to be minimax optimal for the class of functions in consideration [Stone (1985)].

5. Application to homogeneous data: Divide-and-conquer approach. In this section, we apply the divide-and-conquer approach, which is commonly used to deal with massive homogeneous data, to some subpopulations that have huge sample sizes. A general goal of this section is to explore the most computationally efficient way to split the sample in those subpopulations while preserving the best possible statistical inference. Specifically, we want to derive the largest possible number of splits under which the averaged estimators for both components enjoy the same statistical performances as the "oracle" estimator that is computed based on the entire sample. Without loss of generality, we assume the entire sample to be homogeneous by setting all $\boldsymbol{\beta}_0^{(j)}$'s to be equal throughout this section. It is worth mentioning that Li, Lin and Li (2013) have done an earlier and interesting work on parametric or nonparametric models.

The divide-and-conquer method *randomly* splits the massive data into s mutually exclusive subsamples. For simplicity, we assume all the subsamples share the same sample size, denoted as n. Hence, $N = n \times s$. With a bit abuse of notation, we define the divide-and-conquer estimators as $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\widehat{f}^{(j)}$ when they are based on the jth subsample. Thus, the averaged estimator is defined as

$$\bar{\boldsymbol{\beta}} = (1/s)\sum_{j=1}^{s} \widehat{\boldsymbol{\beta}}^{(j)} \quad \text{and} \quad \bar{f}(\cdot) = (1/s)\sum_{j=1}^{s} \widehat{f}^{(j)}(\cdot).$$

Comparing to the oracle estimator, the aggregation procedure reduces the computational complexity in terms of the entire sample size N to the subsample size N/s. In the case of kernel ridge regression, the complexity is $O(N^3)$, while our aggregation procedure (run in one single machine) reduces it to $O(N^3/s^2)$. Propositions 5.1 below state conditions under which the divide-and-conquer estimators maintain the same statistical properties as oracle estimate, that is, so-called oracle property.

PROPOSITION 5.1. *Suppose that the conditions in Theorem 3.6 hold. If we choose* $\lambda = o(N^{-1/2})$, *then*

$$\begin{pmatrix} \sqrt{N}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0(z_0) - W_\lambda f_0(z_0)\big) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \begin{pmatrix} \sigma^2\boldsymbol{\Omega}^{-1} & \boldsymbol{\Sigma}_{12}^* \\ \boldsymbol{\Sigma}_{21}^* & \Sigma_{22}^* \end{pmatrix}\right),$$

where $\boldsymbol{\Sigma}_{12}^* = \boldsymbol{\Sigma}_{21}^{*T} = \sigma^2\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}_{z_0}$ *and* $\Sigma_{22}^* = \sigma^2(\sigma_{z_0}^2 + \boldsymbol{\gamma}_{z_0}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}_{z_0})$. *Moreover, if* $d(\lambda) \to \infty$, *then* $\boldsymbol{\gamma}_{z_0} = \mathbf{0}$. *In this case,* $\boldsymbol{\Sigma}_{12}^* = \boldsymbol{\Sigma}_{21}^{*T} = \mathbf{0}$ *and* $\Sigma_{22}^* = \sigma^2\sigma_{z_0}^2$.

The conclusion of Proposition 5.1 holds no matter s is fixed or diverges (once the condition for s in Theorem 3.6 are satisfied). In view of Proposition 5.1, we note that the above joint asymptotic distribution is exactly the same as that for the oracle estimate, that is, $s = 1$.

REMARK 5.1. We can also derive the minimax rate of $\mathrm{MSE}(\bar{f})$, which is exactly the same as that in Theorem 3.2, based on similar proof techniques.

6. Numerical experiment. In this section, we empirically examine the impact of the number of subpopulations on the statistical inference built on $(\widehat{\boldsymbol{\beta}}^{(j)}, \bar{f})$. As will be seen, the simulation results strongly support our general theory.

Specifically, we consider the partial smoothing spline models in Section 4.3. In the simulation setup, we let $\varepsilon \sim N(0, 1)$, $p = 1$ and $\nu = 2$ (cubic spline). Moreover, $Z \sim \mathrm{Uniform}(-1, 1)$ and $X = (W + Z)/2$, where $W \sim \mathrm{Uniform}(-1, 1)$, such that X and Z are dependent. It is easy to show that $\boldsymbol{\Omega} = E[(X - E[X \mid Z])^2] = 1/12$ and $\boldsymbol{\Sigma} = E[X^2] = 1/6$. To design the heterogenous data setting, we let $\boldsymbol{\beta}_0^{(j)} = j$ for $j = 1, 2, \ldots, s$ on the jth subpopulation. The nonparametric function $f_0(z)$, which is common across all subpopulations, is assumed to be $0.6b_{30,17}(z) + 0.4b_{3,11}$, where b_{a_1,a_2} is the density function for $\mathrm{Beta}(a_1, a_2)$.

We start from the 95% predictive interval [at (x_0, z_0)] implied by the joint asymptotic distribution (4.3):

$$\Big[\widehat{Y}^{(j)} \pm 1.96\sigma\sqrt{x_0^T\boldsymbol{\Omega}^{-1}x_0/n + \sigma_{z_0}^2/\big(N\lambda^{1/(2\nu)}\big) + 1}\Big],$$

where $\widehat{Y}^{(j)} = x_0^T\widehat{\boldsymbol{\beta}}^{(j)} + \bar{f}(z_0)$ is the predicted response. The unknown error variance σ is estimated by $(\widehat{\sigma}^{(j)})^2 = n^{-1}\sum_{i\in S_j}(Y_i - X_i^T\widehat{\boldsymbol{\beta}}^{(j)} - \widehat{f}^{(j)}(Z_i))^2/(n - Tr(A(\lambda)))$, where $A(\lambda)$ denotes the smoothing matrix, followed by an aggregation $\bar{\sigma}^2 = 1/s\sum_{j=1}^{s}(\widehat{\sigma}^{(j)})^2$. In the simulations, we fix $x_0 = 0.5$ and choose $z_0 = 0.25, 0.5, 0.75$ and 0.95. The coverage probability is calculated based on 200 repetitions. As for N and s, we set $N = 256, 528, 1024, 2048, 4096$, and choose $s = 2^0, 2^1, \ldots, 2^{t-3}$ when $N = 2^t$. The simulation results are summarized in Figure 1. We notice an interesting phase transition from Figure 1: when $s \le s^*$ where $s^* \approx N^{0.45}$, the coverage probability is approximately 95%; when $s \ge s^*$, the coverage probability drastically decreases. This empirical observation is strongly supported by our theory developed in Section 4.3 where $s^* \approx N^{0.42}\log^{-6} N$ for $\nu = 2$.

We next compute the mean-squared errors of $\bar{f}$ under different choices of N and s in Figure 2. It is demonstrated that the increasing trends of MSE as s increases are very similar for different N. More importantly, all the MSE curves suddenly blow up when $s \approx N^{0.4}$. This is also close to our theoretical result that the transition point is around $N^{0.45}\log^{-6} N$.

We next empirically verify the efficiency boosting theory developed in Section 3.5. Based on $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\check{\boldsymbol{\beta}}^{(j)}$, we construct the following two types of 95% confidence intervals for $\boldsymbol{\beta}_0^{(j)}$:

$$\mathrm{CI}_1 = [\widehat{\boldsymbol{\beta}}^{(j)} \pm 1.96\boldsymbol{\Omega}^{-1/2}n^{-1/2}\bar{\sigma}],$$

and

$$\mathrm{CI}_2 = [\check{\boldsymbol{\beta}}^{(j)} \pm 1.96\boldsymbol{\Sigma}^{-1/2}n^{-1/2}\bar{\sigma}].$$

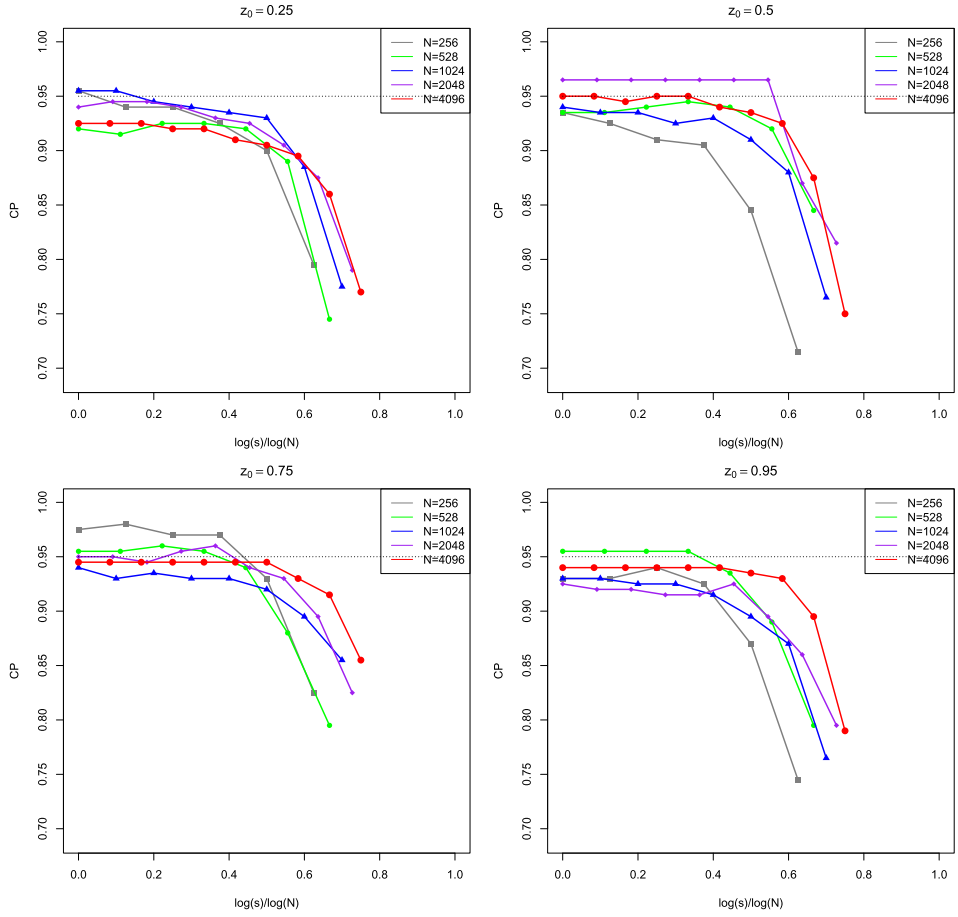

FIG. 1. *Coverage probability of* 95% *predictive interval with different choices of* s *and* N.

Obviously, CI_2 is shorter than CI_1. However, Theorem 3.7 shows that CI_2 is valid only when s satisfies both a upper bound and a lower bound. This theoretical condition is empirically verified in Figure 3 which exhibits the validity range of CI_2 in terms of s. In Figure 4, we further compare CI_2 and CI_1 in terms of their coverage probabilities and lengths. This figure shows that when s is in a proper range, the coverage probabilities of CI_1 and CI_2 are similar, while CI_2 is significantly shorter.

Finally, we consider the heterogeneity testing. In Figure 5, we compare tests Ψ_1 and Ψ_2 under different choices of N and $s \geq 2$. Specifically, Figure 5(i) compares the nominal levels, while Figure 5(ii)–(iv) compare the powers under various alternative hypotheses $H_1 : \boldsymbol{\beta}_0^{(j)} - \boldsymbol{\beta}_0^{(k)} = \Delta$, where $\Delta = 0.5, 1, 1.5$. It is clearly seen that both tests are consistent, and their powers increase as Δ or N increases. In addition, we observe that Ψ_2 has uniformly larger powers than Ψ_1.

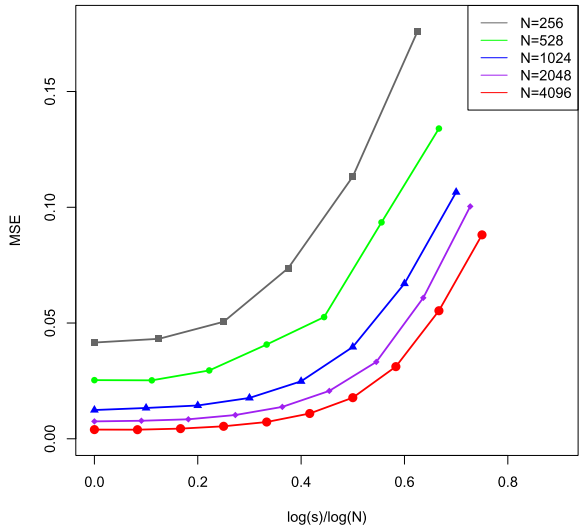

FIG. 2. *Mean-square errors of* $\bar{f}$ *under different choices of* N *and* s.

7. Proof of main results. In this section, we present main proofs of Lemma 3.1 and Theorems 3.2, 3.4 and 3.6 in the main text.

7.1. *Proof of Lemma* 3.1. We start from analyzing the minimization problem (3.2) on each subpopulation. Recall $m = (\boldsymbol{\beta}, f)$ and $U = (\mathbf{X}, Z)$. The objective

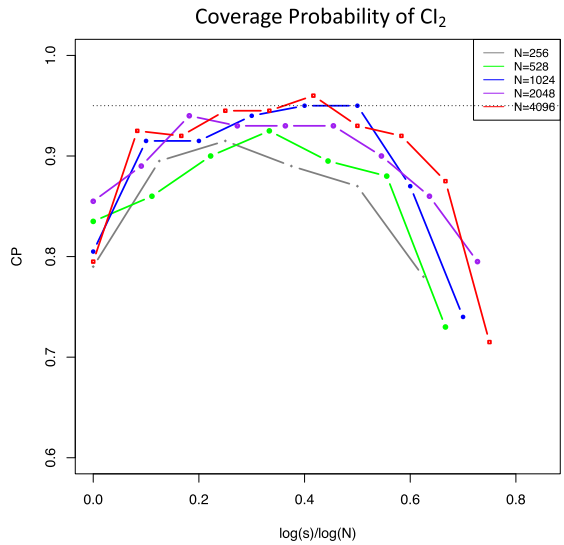

FIG. 3. *Coverage probability of* 95% *confidence interval based on* $\check{\boldsymbol{\beta}}^{(j)}$.

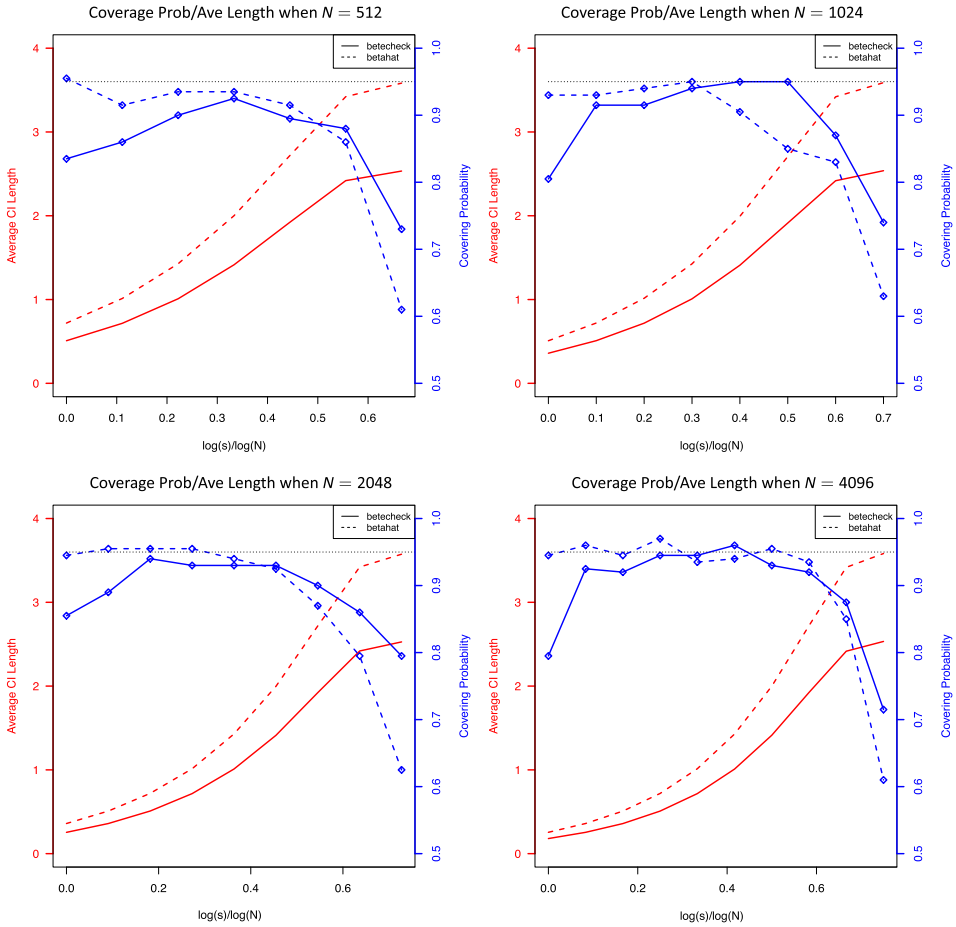

FIG. 4. *Coverage probabilities and average lengths of* 95% *for two types of confidence intervals. In the above figures, dashed lines represent* CI_1*, which is constructed based on* $\widehat{\boldsymbol{\beta}}^{(j)}$*, and solid lines represent* CI_2*, which is constructed based on* $\check{\boldsymbol{\beta}}^{(j)}$.

function can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{i\in S_j}\bigl(Y_i-\mathbf{X}_i^T\boldsymbol{\beta}-f(Z_i)\bigr)^2+\lambda\|f\|_{\mathcal{H}}^2\\
&\quad=\frac{1}{n}\sum_{i\in S_j}\bigl(Y_i-m(U_i)\bigr)^2+\langle P_\lambda m,m\rangle_{\mathcal{A}}\\
&\quad=\frac{1}{n}\sum_{i\in S_j}\bigl(Y_i-\langle R_{U_i},m\rangle_{\mathcal{A}}\bigr)^2+\langle P_\lambda m,m\rangle_{\mathcal{A}}.
\end{aligned}$$

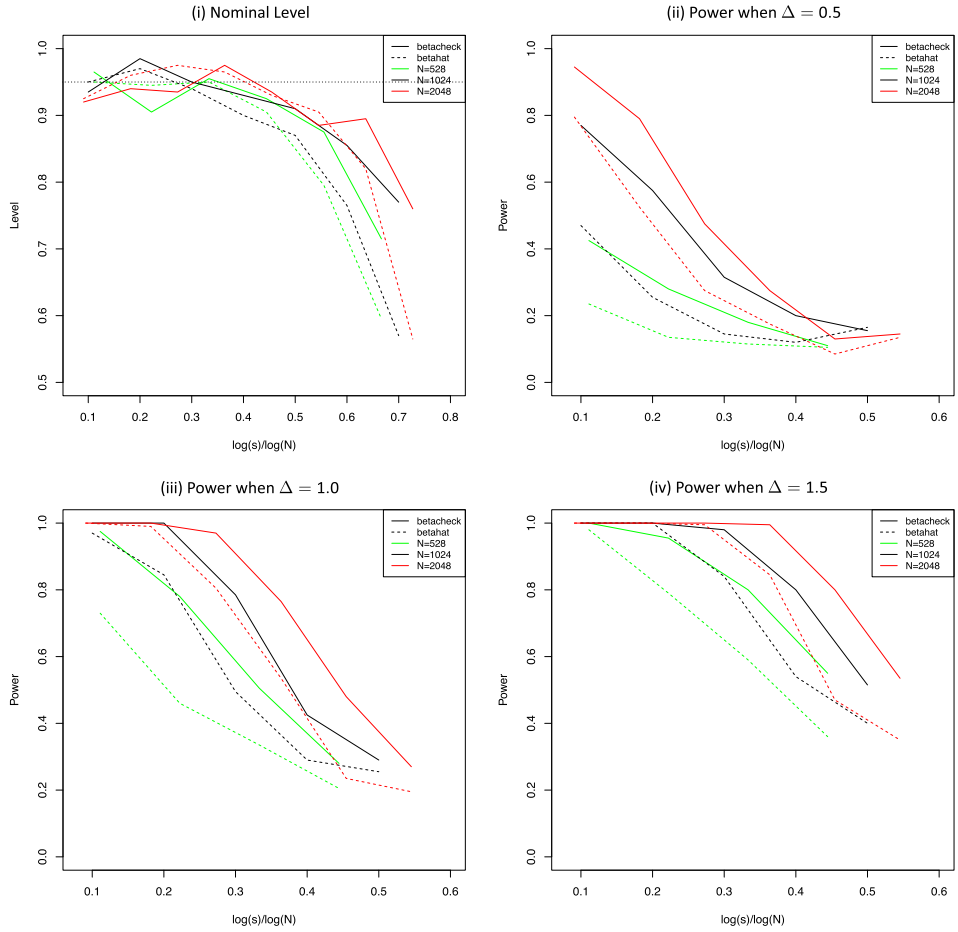

FIG. 5. (i) *Nominal level of heterogeneity tests* Ψ_1 *and* Ψ_2; (ii)–(iv) *Power of heterogeneity tests* Ψ_1 *and* Ψ_2 *when* $\Delta = 0.5, 1.0, 1.5$. *In the above figures, dashed lines represent* Ψ_1, *which is constructed based on* $\check{\boldsymbol{\beta}}$, *and solid lines represent* Ψ_2, *which is constructed based on* $\widehat{\boldsymbol{\beta}}$.

The first-order optimality condition (w.r.t. Fréchet derivative) gives

$$\frac{1}{n}\sum_{i\in S_j} R_{U_i}\big(\widehat{m}^{(j)}(U_i) - Y_i\big) + P_\lambda \widehat{m}^{(j)} = 0,$$

where $\widehat{m}^{(j)} = (\widehat{\boldsymbol{\beta}}^{(j)}, \widehat{f}^{(j)})$. This implies that

$$-\frac{1}{n}\sum_{i\in S_j} R_{U_i}\varepsilon_i + \frac{1}{n}\sum_{i\in S_j} R_{U_i}\big(\widehat{m}^{(j)}(U_i) - m_0^{(j)}(U_i)\big) + P_\lambda \widehat{m}^{(j)} = 0,$$

where $m_0^{(j)} = (\boldsymbol{\beta}_0^{(j)}, f_0)$. Define $\Delta m^{(j)} := \widehat{m}^{(j)} - m_0^{(j)}$. Adding $\mathbb{E}_U[R_U \Delta m^{(j)}(U)]$ on both sides of the above equation, we have

$$\begin{aligned}
&\mathbb{E}_U\big[R_U \Delta m^{(j)}(U)\big] + P_\lambda \Delta m^{(j)} \\
&\quad = \frac{1}{n}\sum_{i \in S_j} R_{U_i}\varepsilon_i - P_\lambda m_0^{(j)} \\
&\qquad - \frac{1}{n}\sum_{i \in S_j}\big(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\big[R_U \Delta m^{(j)}(U)\big]\big).
\end{aligned} \tag{7.1}$$

The LHS of (7.1) can be rewritten as

$$\begin{aligned}
\mathbb{E}_U\big[R_U \Delta m^{(j)}(U)\big] + P_\lambda \Delta m^{(j)} &= \mathbb{E}_U\big[R_U \langle R_U, \Delta m^{(j)}\rangle_{\mathcal{A}}\big] + P_\lambda \Delta m^{(j)} \\
&= \big(\mathbb{E}_U[R_U \otimes R_U] + P_\lambda\big)\Delta m^{(j)} = \Delta m^{(j)},
\end{aligned}$$

where the last equality follows from proposition 2.2. Then (7.1) becomes

$$\begin{aligned}
\widehat{m}^{(j)} - m_0^{(j)} &= \frac{1}{n}\sum_{i \in S_j} R_{U_i}\varepsilon_i - P_\lambda m_0^{(j)} \\
&\quad - \frac{1}{n}\sum_{i \in S_j}\big(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\big[R_U \Delta m^{(j)}(U)\big]\big).
\end{aligned} \tag{7.2}$$

We denote the last term in the RHS of (7.2) as

$$\mathrm{Rem}^{(j)} := \frac{1}{n}\sum_{i \in S_j}\big(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\big[R_U \Delta m^{(j)}(U)\big]\big).$$

Recall that $R_u = (L_u, N_u)$ and $P_\lambda m_0^{(j)} = (L_\lambda f_0, N_\lambda f_0)$. Thus, the above remainder term decomposes into two components:

$$\mathrm{Rem}_\beta^{(j)} := \frac{1}{n}\sum_{i \in S_j}\big(L_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\big[L_U \Delta m^{(j)}(U)\big]\big),$$

$$\mathrm{Rem}_f^{(j)} := \frac{1}{n}\sum_{i \in S_j}\big(N_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\big[N_U \Delta m^{(j)}(U)\big]\big).$$

Therefore, (7.2) can be rewritten into equations (3.5) and (3.6) for all $j = 1, \ldots, s$. This completes the proof of the first part of Lemma 3.1. Taking average of (3.6) for all j over s, and by definition of $\bar{f}$, we have

$$\bar{f} - f_0 = \frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i - N_\lambda f_0 - \frac{1}{s}\sum_{j=1}^{s}\mathrm{Rem}_f^{(j)}, \tag{7.3}$$

where we used $1/N\sum_{i=1}^{N} N_{U_i}\varepsilon_i = 1/s\sum_{j=1}^{s} 1/n\sum_{i\in S_j} N_{U_i}\varepsilon_i$. Equations (3.5) and (7.3) are the basic equalities to derive the finite sample rate of convergence and joint limit distribution of $\widehat{\boldsymbol{\beta}}^{(j)}$ and $\bar{f}$. To this end, we need to control the remainder terms in the above two equalities, which is the second part of Lemma 3.1. We delegate the proofs to the following two lemmas.

LEMMA 7.1. *Suppose the conditions in Lemma 3.1 hold. We have for all* $j = 1, \ldots, s$

$$\mathbb{E}[\|\mathrm{Rem}_{\beta}^{(j)}\|_2^2] \le a(n, \lambda, J),$$

for sufficiently large n, *where* $a(n, \lambda, J)$ *is as defined in Lemma 3.1. Moreover, the inequality also holds for* $\mathbb{E}[\|\mathrm{Rem}_{f}^{(j)}\|_{\mathcal{C}}^2]$.

LEMMA 7.2. *Suppose the conditions in Lemma 3.1 hold. We have the following two sets of results that control the remainder terms*:

(i) *For all* $j = 1, \ldots, s$, *it holds that*

$$\mathbb{P}\big(\|\mathrm{Rem}_{\beta}^{(j)}\|_2 \ge b(n, \lambda, J)\big) \lesssim n\exp\big(-c\log^2 n\big), \tag{7.4}$$

where $b(n, \lambda, J)$ *is as defined in Lemma 3.1.*

(ii) *In addition, we have*

$$\left\|\frac{1}{s}\sum_{j=1}^{s}\mathrm{Rem}_{\beta}^{(j)}\right\|_2 = o_P\big(s^{-1/2}b(n, \lambda, J)\log N\big). \tag{7.5}$$

Furthermore, (7.4) *and* (7.5) *also hold if* $\|\mathrm{Rem}_{\beta}^{(j)}\|_2$ *and* $\|s^{-1}\sum_{j=1}^{s}\mathrm{Rem}_{\beta}^{(j)}\|_2$ *are replaced by* $\|\mathrm{Rem}_{f}^{(j)}\|_{\mathcal{C}}$ *and* $\|s^{-1}\sum_{j=1}^{s}\mathrm{Rem}_{f}^{(j)}\|_{\mathcal{C}}$.

By the above two lemmas, we complete the second part of Lemma 3.1.

7.2. *Proof of Theorem 3.2.* By (7.3), it follows that

$$\begin{aligned}\mathbb{E}[\|\bar{f} - f_0\|_{\mathcal{C}}^2] \le{}& 3\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i\right\|_{\mathcal{C}}^2\right] + 3\|N_\lambda f_0\|_{\mathcal{C}}^2 \\ &+ 3\mathbb{E}\left[\left\|\frac{1}{s}\sum_{j=1}^{s}\mathrm{Rem}_{f}^{(j)}\right\|_{\mathcal{C}}^2\right].\end{aligned} \tag{7.6}$$

By Lemma A.5 in the supplemental material [Zhao, Cheng and Liu (2015)] and the fact that each $N_{U_i}\varepsilon_i$ is i.i.d., it follows that

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i\right\|_{\mathcal{C}}^2\right] = \frac{1}{N}\mathbb{E}[\|N_U\varepsilon\|_{\mathcal{C}}^2] \le C_1\sigma^2\frac{d(\lambda)}{N}, \tag{7.7}$$

and

$$\|N_\lambda f_0\|_{\mathcal{C}}^2 \le 2\|f_0\|_{\mathcal{H}}^2\lambda + C_2\lambda^2, \tag{7.8}$$

where C_1 and C_2 are constants specified in Lemma A.5. As for the third term in (7.6), we have by independence across subpopulations that

$$\mathbb{E}\left[\left\|\frac{1}{s}\sum_{j=1}^{s}\mathrm{Rem}_f^{(j)}\right\|_{\mathcal{C}}^2\right] = \frac{1}{s^2}\sum_{j=1}^{s}\mathbb{E}\big[\|\mathrm{Rem}_f^{(j)}\|_{\mathcal{C}}^2\big]. \tag{7.9}$$

Combining (7.6)–(7.9) and Lemma 7.1, and by the fact that $\|\bar f - f_0\|_{L_2(\mathbb{P}_Z)}^2 \le \|\bar f - f_0\|_{\mathcal{C}}^2$, we complete the proof of Theorem 3.2.

7.3. *Proof of Theorem 3.4.* Recall that $m_0^{(j)*} = (\boldsymbol{\beta}_0^{(j)*}, f_0^*) = (\mathrm{id} - P_\lambda)m_0^{(j)}$ where $m_0^{(j)} = (\boldsymbol{\beta}_0^{(j)}, f_0)$. This implies that $\boldsymbol{\beta}_0^{(j)*} = \boldsymbol{\beta}_0^{(j)} - L_\lambda f_0$ and $f_0^* = f_0 - N_\lambda f_0$. By (3.5) and (7.3), for arbitrary $\mathbf{x}$ and z_0,

$$\begin{aligned}
&(\mathbf{x}^T, 1)\begin{pmatrix}\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}) \\ \sqrt{N/d(\lambda)}(\bar f(z_0) - f_0^*(z_0))\end{pmatrix} \\
&\quad= \sqrt{n}\mathbf{x}^T(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}) + \sqrt{N/d(\lambda)}(\bar f_{N,\lambda}(z_0) - f_0^*(z_0)) \\
&\quad= \underbrace{\frac{1}{\sqrt{n}}\sum_{i\in S_j}\mathbf{x}^T L_{U_i}\varepsilon_i + \frac{1}{\sqrt{N}}\sum_{i=1}^{N} d(\lambda)^{-1/2}N_{U_i}(z_0)\varepsilon_i}_{\text{(I)}} \\
&\qquad+ \underbrace{\sqrt{n}\mathbf{x}^T\,\mathrm{Rem}_\beta^{(j)} + \sqrt{N/d(\lambda)}s^{-1}\sum_{j=1}^{s}\mathrm{Rem}_f^{(j)}(z_0)}_{\text{(II)}}.
\end{aligned}$$

In what follows, we will show that the main term (I) is asymptotically normal and the remainder term (II) is of order $o_P(1)$. Given that $\mathbf{x}$ is arbitrary, we apply Wold device to complete the proof of joint asymptotic normality.

Asymptotic normality of (I): We present the result for showing asymptotic normality of (I) in the following lemma and defer its proof to supplemental material [Zhao, Cheng and Liu (2015)].

LEMMA 7.3. *Suppose Assumptions 3.1, 3.2 hold and that* $\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)}/d(\lambda)^{1/2} \to \sigma_{z_0}$, $(W_\lambda\mathbf{A})(z_0)/d(\lambda)^{1/2} \to \boldsymbol{\alpha}_{z_0} \in \mathbb{R}^p$, *and* $\mathbf{A}(z_0)/d(\lambda)^{1/2} \to -\boldsymbol{\gamma}_{z_0} \in \mathbb{R}^p$ *as* $N \to \infty$. *We have*:

(i) *if* $s \to \infty$, *then*

$$(I) \rightsquigarrow N\big(0, \sigma^2(\mathbf{x}^T\boldsymbol{\Omega}^{-1}\mathbf{x} + \Sigma_{22})\big). \tag{7.10}$$

(ii) *if s is fixed, then*

$$(I) \rightsquigarrow N\big(0, \sigma^2\big(\mathbf{x}^T \boldsymbol{\Omega}^{-1}\mathbf{x} + \Sigma_{22} + 2s^{-1/2}\mathbf{x}^T \boldsymbol{\Sigma}_{12}\big)\big). \tag{7.11}$$

Control of the remainder term (II): We now turn to bound the remainder term (II). We can show that if (3.12) holds, then $d(\lambda)n^{-1/2}(J(\mathcal{F}, 1) + \log n) = o(1)$. Hence, by Lemma 7.2, we have

$$\begin{aligned}\sqrt{n}|\mathbf{x}^T \operatorname{Rem}_\beta^{(j)}| &\le \sqrt{n}\|\mathbf{x}\|_2 \|\operatorname{Rem}_\beta^{(j)}\|_2 = o_P\big(n^{1/2} b(n, \lambda, J)\big) \\ &= o_P\big(\sqrt{N} s^{-1/2} b(n, \lambda, J)\big),\end{aligned} \tag{7.12}$$

where we used the boundedness of $\mathbf{x}$. Also,

$$\begin{aligned}\sqrt{N/d(\lambda)}\left|s^{-1}\sum_{j=1}^{s} \operatorname{Rem}_f^{(j)}(z_0)\right| &\le \sqrt{N/d(\lambda)}\|\widetilde{K}_{z_0}\|_{\mathcal{C}}\left\|s^{-1}\sum_{j=1}^{s}\operatorname{Rem}_f^{(j)}\right\|_{\mathcal{C}} \\ &\lesssim \sqrt{N}\left\|s^{-1}\sum_{j=1}^{s}\operatorname{Rem}_f^{(j)}\right\|_{\mathcal{C}} \\ &= o_P\big(\sqrt{N} s^{-1/2} b(n, \lambda, J)\log N\big),\end{aligned} \tag{7.13}$$

where the second inequality follows from Lemma A.4 in the supplemental material [Zhao, Cheng and Liu (2015)]. Therefore, by (7.12) and (7.13), we have

$$(\mathrm{II}) = o_P\big(\sqrt{N} s^{-1/2} b(n, \lambda, J)\log N\big). \tag{7.14}$$

Now by definition of $b(n, \lambda, J)$ and condition (3.13), we have $(\mathrm{II}) = o_P(1)$.

Combining (7.10) and (7.14), it follows that if $s \to \infty$, then

$$\big(\mathbf{x}^T, 1\big)\begin{pmatrix}\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\big) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0^*(z_0)\big)\end{pmatrix} \rightsquigarrow N\big(0, \sigma^2\big(\mathbf{x}^T \boldsymbol{\Omega}^{-1}\mathbf{x} + \Sigma_{22}\big)\big).$$

Combining (7.11) and (7.14), it follows that if s is fixed, then

$$\begin{aligned}&\big(\mathbf{x}^T, 1\big)\begin{pmatrix}\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\big) \\ \sqrt{N/d(\lambda)}\big(\bar{f}(z_0) - f_0^*(z_0)\big)\end{pmatrix} \\ &\quad\rightsquigarrow N\big(0, \sigma^2\big(\mathbf{x}^T \boldsymbol{\Omega}^{-1}\mathbf{x} + \Sigma_{22} + 2s^{-1/2}\mathbf{x}^T \boldsymbol{\Sigma}_{12}\big)\big).\end{aligned}$$

By the arbitrariness of $\mathbf{x}$, we reach the conclusion of the theorem using Wold device.

7.4. *Proof of Lemma 7.2*: *Controlling the remainder term.*

(i) Recall that $\operatorname{Rem}^{(j)} = (\operatorname{Rem}_\beta^{(j)}, \operatorname{Rem}_f^{(j)}) \in \mathcal{A}$. We first derive the bound of $\|\operatorname{Rem}^{(j)}\|_{\mathcal{A}}$. Recall

$$\operatorname{Rem}^{(j)} = \frac{1}{n}\sum_{i \in S_j} \Delta m^{(j)}(U_i) R_{U_i} - \mathbb{E}_U\big[\Delta m^{(j)}(U) R_U\big].$$

Let $Z_n(m) = c_r^{-1} d(\lambda)^{-1/2} n^{-1/2} \sum_{i \in S_j} \{m(U_i) R_{U_i} - \mathbb{E}[m(U) R_U]\}$, where c_r is the constant specified in Lemma A.4. Note that $Z_n(m)$ is implicitly related to j but we omit the superscript of (j). We have $\mathrm{Rem}^{(j)} = c_r^{-1} \sqrt{n/d(\lambda)} Z_n(\Delta m^{(j)})$. We apply Lemma F.1 in the supplemental material [Zhao, Cheng and Liu (2015)] to obtain an exponential inequality for $\sup_{m \in \mathcal{F}} \|Z_n(m)\|_{\mathcal{A}}$. The first step is to show that $Z_n(m)$ is a sub-Gaussian process by Lemma G.1 in the supplemental material [Zhao, Cheng and Liu (2015)]. Let $g(U_i, m) = c_r^{-1} \sqrt{n/d(\lambda)} (m(U_i) R_{U_i} - \mathbb{E}[m(U) R_U])$. Now for any m_1 and m_2,

$$\begin{aligned}
&\big\|g(U_i, m_1) - g(U_i, m_2)\big\|_{\mathcal{A}} \\
&\quad = c_r^{-1} \sqrt{n/d(\lambda)} \big\{ \big\| \big(m_1(U_i) - m_2(U_i)\big) R_{U_i} \big\|_{\mathcal{A}} \\
&\qquad + \big\| \mathbb{E}\big[\big(m_1(U) - m_2(U)\big) R_U\big] \big\|_{\mathcal{A}} \big\} \\
&\quad \le 2\sqrt{n} \|m_1 - m_2\|_{\sup},
\end{aligned}$$

where we used the fact that $\|R_u\|_{\mathcal{A}} \le c_r d(\lambda)^{1/2}$ by Lemma A.4. Note that $Z_n(m) = \frac{1}{n} \sum_{i \in S_j} g(U_i, m)$. Therefore, by Lemma G.1, we have for any $t > 0$,

$$\begin{aligned}
&\mathbb{P}\big(\big\|Z_n(m_1) - Z_n(m_2)\big\|_{\mathcal{A}} \ge t\big) \\
&\quad = \mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^{n} \big\{ g(U_i, m_1) - g(U_i, m_2) \big\} \right\|_{\mathcal{A}} \ge t \right) \\
&\quad \le 2 \exp\left(- \frac{t^2}{8 \|m_1 - m_2\|_{\sup}^2} \right).
\end{aligned} \tag{7.15}$$

Then by Lemma G.1, we have

$$\mathbb{P}\Big(\sup_{m \in \mathcal{F}} \big\|Z_n(m)\big\|_{\mathcal{A}} \ge C J\big(\mathcal{F}, \mathrm{diam}(\mathcal{F})\big) + x \Big) \le C \exp\left(\frac{-x^2}{C\, \mathrm{diam}(\mathcal{F})^2} \right), \tag{7.16}$$

where $\mathrm{diam}(\mathcal{F}) = \sup_{m_1, m_2 \in \mathcal{F}} \|m_1 - m_2\|_{\sup}$.

Define $q_{n,\lambda} = c_r r_{n,\lambda} d(\lambda)^{1/2}$ and $\widetilde{m} = q_{n,\lambda}^{-1} \Delta m^{(j)}/2$. Again we do not specify its relationship with j. Define the event $\mathcal{E} = \{\|\Delta m^{(j)}\|_{\mathcal{A}} \le r_{n,\lambda}\}$. On the event $\mathcal{E}$, we have

$$\|\widetilde{m}\|_{\sup} \le c_r d(\lambda)^{1/2} (2q_{n,\lambda})^{-1} \big\|\Delta m^{(j)}\big\|_{\mathcal{A}} \le 1/2,$$

where we used the fact that $\|\widetilde{m}\|_{\sup} \le c_r d(\lambda)^{1/2} \|\widetilde{m}\|_{\mathcal{A}}$ by Lemma A.4. This implies $|\mathbf{x}^T \widetilde{\boldsymbol{\beta}} + \widetilde{f}(z)| \le 1/2$ for any $(\mathbf{x}, z)$. Letting $\mathbf{x} = 0$, one gets $\|\widetilde{f}\|_{\sup} \le 1/2$, which further implies $|\mathbf{x}^T \widetilde{\boldsymbol{\beta}}| \le 1$ for all $\mathbf{x}$ by triangular inequality. Moreover, on the even $\mathcal{E}$ we have

$$\|\widetilde{f}\|_{\mathcal{H}} \le \lambda^{-1/2} \|\widetilde{m}\|_{\mathcal{A}} \le \lambda^{-1/2} / (2q_{n,\lambda}) \big\|\Delta m^{(j)}\big\|_{\mathcal{A}} \le c_r^{-1} d(\lambda)^{-1/2} \lambda^{-1/2}$$

by the definition of $\|\cdot\|_{\mathcal{A}}$. Hence, we have shown that $\mathcal{E} \subset \{\widetilde{m} \in \mathcal{F}\}$. Combining this fact with (7.16), and noting that $\operatorname{diam}(\mathcal{F}) \le 1$, we have

$$\mathbb{P}\big(\{\|Z_n(\widetilde{m})\|_{\mathcal{A}} \ge C J(\mathcal{F}, 1) + x\} \cap \mathcal{E}\big) \le C \exp(-x^2/C), \tag{7.17}$$

by Lemma F.1. Using the definition of $\widetilde{m}$, and the relationship that $\mathrm{Rem}^{(j)} = c_r^{-1}\sqrt{n/d(\lambda)}Z_n(\Delta m^{(j)})$, we calculate that

$$Z_n(\widetilde{m}) = (1/2)d(\lambda)^{-1/2}n^{1/2}q_{n,\lambda}^{-1}\,\mathrm{Rem}^{(j)} = (1/2)c_r^{-1}d(\lambda)^{-1}n^{1/2}r_{n,\lambda}^{-1}\,\mathrm{Rem}^{(j)}.$$

Plugging the above form of $Z_n(\widetilde{m})$ into (7.17) and letting $x = \log n$ in (7.17), we have

$$\mathbb{P}\big(\{\|\mathrm{Rem}^{(j)}\|_{\mathcal{A}} \ge b(n, \lambda, J)\} \cap \mathcal{E}\big) \le C \exp(-\log^2 n/C), \tag{7.18}$$

where we used the definition that $b(n, \lambda, J) = Cd(\lambda)n^{-1/2}r_{n,\lambda}(J(\mathcal{F}, 1) + \log n)$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}\big(\|\mathrm{Rem}^{(j)}\|_{\mathcal{A}} \ge b(n, \lambda, J)\big)\\
&\quad\le \mathbb{P}\big(\{\|\mathrm{Rem}^{(j)}\|_{\mathcal{A}} \ge b(n, \lambda, J)\} \cap \mathcal{E}\big) + \mathbb{P}(\mathcal{E}^c)\\
&\quad\le C \exp(-\log^2 n/C) + \mathbb{P}(\mathcal{E}^c).
\end{aligned} \tag{7.19}$$

We have the following lemma that controls $\mathbb{P}(\mathcal{E}^c)$.

LEMMA 7.4. *Suppose the conditions in Lemma 3.1 hold. There exist a constant c such that*

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}\big(\|\Delta m^{(j)}\|_{\mathcal{A}} \ge r_{n,\lambda}\big) \lesssim n \exp(-c \log^2 n),$$

for all $j = 1, \ldots, s$.

By Lemma 7.4 and (7.19), we have

$$\mathbb{P}\big(\|\mathrm{Rem}^{(j)}\|_{\mathcal{A}} \ge b(n, \lambda, J)\big) \lesssim n \exp(-c \log^2 n). \tag{7.20}$$

We can apply similar arguments as above to bound $\|\mathrm{Rem}_f^{(j)}\|_{\mathcal{C}}$, by changing $\omega(\mathcal{F}, 1)$ to $\omega(\mathcal{F}_2, 1)$, which is dominated by $\omega(\mathcal{F}, 1)$. The bound of $\|\mathrm{Rem}_\beta^{(j)}\|_2$ then follows from triangular inequality.

(ii) We will use an Azuma-type inequality in Hilbert space to control the averaging remainder term $s^{-1}\sum_{j=1}^s \mathrm{Rem}^{(j)}$, as all $\mathrm{Rem}^{(j)}$ are independent and have zero mean. Define the event $\mathcal{A}_j = \{\|\mathrm{Rem}^{(j)}\|_{\mathcal{A}} \le b(n, \lambda, J)\}$. By Lemma G.1, we have

$$\begin{aligned}
&\mathbb{P}\left(\left\{\bigcap_j \mathcal{A}_j\right\} \cap \left\{\left\|s^{-1}\sum_{j=1}^s \mathrm{Rem}^{(j)}\right\|_{\mathcal{A}} > s^{-1/2} b(n, \lambda, J) \log N\right\}\right)\\
&\quad\le 2 \exp(-\log^2 N/2).
\end{aligned} \tag{7.21}$$

Moreover, by (7.20),

$$\mathbb{P}(\mathcal{A}_j^c) \lesssim n\exp(-c\log^2 n). \tag{7.22}$$

Hence, it follows that

$$\begin{aligned}
&\mathbb{P}\left(\left\|s^{-1}\sum_{j=1}^{s}\mathrm{Rem}^{(j)}\right\|_{\mathcal{A}} > s^{-1/2}b(n,\lambda,J)\log N\right)\\
&\quad\le \mathbb{P}\left(\left\{\bigcap_{j=1}^{s}\mathcal{A}_j\right\}\cap\left\{\left\|s^{-1}\sum_{j=1}^{s}\mathrm{Rem}^{(j)}\right\|_{\mathcal{A}} > s^{-1/2}b(n,\lambda,J)\log N\right\}\right)\\
&\qquad+\mathbb{P}\left(\bigcup_j \mathcal{A}_j^c\right)\\
&\quad\lesssim 2\exp(-\log^2 N/2) + ns\exp(-c\log^2 n) \lesssim N\exp(-c\log^2 n),
\end{aligned}$$

where the second inequality follows from (7.21), (7.22) and union bound. By our technical assumption that $s \lesssim N^{\psi}$ (stated before Assumption 3.1), we have $N\exp(-c\log^2 n) \asymp N\exp(-c'\log^2 N) \to 0$ as $N\to\infty$. This completes the proof of Part (ii).

Applying similar arguments as in (i), we get the similar inequalities for $\|1/s\sum_{j=1}^{s}\mathrm{Rem}_{\beta}^{(j)}\|_2$ and $\|1/s\sum_{j=1}^{s}\mathrm{Rem}_{f}^{(j)}\|_{\mathcal{C}}$.

7.5. *Proof of Theorem* 3.6. In view of Theorem 3.6, we first prove

$$\begin{pmatrix}\sqrt{n}(\boldsymbol{\beta}_0^{(j)*} - \boldsymbol{\beta}_0^{(j)})\\ \sqrt{N/d(\lambda)}(f_0^*(z_0) - f_0(z_0) - W_\lambda f_0(z_0))\end{pmatrix} \to \mathbf{0} \tag{7.23}$$

for both (i) and (ii). By Proposition 2.3, we have

$$\begin{pmatrix}\boldsymbol{\beta}_0^{(j)*} - \boldsymbol{\beta}_0^{(j)}\\ f_0^*(z_0) - f_0(z_0)\end{pmatrix} = \begin{pmatrix}L_\lambda f_0\\ W_\lambda f_0(z_0) + \mathbf{A}(z_0)^T L_\lambda f_0\end{pmatrix}. \tag{7.24}$$

By Lemma A.5, it follows that under Assumption 3.3, $\|L_\lambda f_0\|_2 \lesssim \lambda$. Now we turn to $f_0^*(z_0) - f_0(z_0)$. Observe that

$$\mathbf{A}(z) = \langle \mathbf{A}, \widetilde{K}_z\rangle_{\mathcal{C}} = \langle \mathbf{B}, \widetilde{K}_z\rangle_{L_2(\mathbb{P}_Z)} = \sum_{\ell=1}^{\infty}\frac{\langle \mathbf{B}, \phi_\ell\rangle_{L_2(\mathbb{P}_Z)}}{1+\lambda/\mu_\ell}\phi_\ell(z). \tag{7.25}$$

Applying Cauchy–Schwarz, we obtain

$$A_k(z_0)^2 \le \left(\sum_{\ell=1}^{\infty}\frac{\langle B_k, \phi_\ell\rangle_{L_2(\mathbb{P}_Z)}^2}{\mu_\ell}\phi_\ell^2(z_0)\right)\left(\sum_{\ell=1}^{\infty}\frac{\mu_\ell}{(1+\lambda/\mu_\ell)^2}\right) \le c_\phi^2\|B_k\|_{\mathcal{H}}^2\,\mathrm{Tr}(K),$$

where the last inequality follows from the uniform boundedness of ϕ_ℓ. Hence, we have that $A_k(z_0)$ is uniformly bounded, which implies $\mathbf{A}(z_0)^T L_\lambda f_0 \le$

$\|\mathbf{A}(z_0)\|_2 \|L_\lambda f_0\|_2 \lesssim \lambda$. Therefore, if we choose $\lambda = o(\sqrt{d(\lambda)/N} \wedge n^{-1/2})$, then we get (7.23), which eliminates the estimation bias for $\boldsymbol{\beta}_0^{(j)}$.

Now we consider the asymptotic variance for cases (i) and (ii). It suffices to show that $\boldsymbol{\alpha}_{z_0} = \mathbf{0}$ under Assumption 3.3. Recall that

$$\boldsymbol{\alpha}_{z_0} = \lim_{N\to\infty} d(\lambda)^{-1/2} W_\lambda \mathbf{A}(z_0).$$

By Lemma A.2 in the supplemental material [Zhao, Cheng and Liu (2015)] and (7.25), we have

$$\begin{aligned} W_\lambda A_k(z_0) &= \sum_{\ell=1}^{\infty} \frac{\langle B_k, \phi_\ell\rangle_{L_2(\mathbb{P}_Z)}}{1+\lambda/\mu_\ell} \frac{\lambda}{\lambda+\mu_\ell} \phi_\ell(z_0) \\ &\le \left(\sum_{\ell=1}^{\infty} \frac{\langle B_k, \phi_\ell\rangle^2_{L_2(\mathbb{P}_Z)}}{\mu_\ell} \phi_\ell^2(z_0)\right)\left(\sum_{\ell=1}^{\infty} \frac{\mu_\ell}{(1+\lambda/\mu_\ell)^2}\right) \\ &\le c_\phi^2 \|B_k\|_{\mathcal{H}}^2 \operatorname{Tr}(K). \end{aligned}$$

Hence, by dominated conference theorem, as $\lambda \to 0$ we have $W_\lambda A_k(z_0) \to 0$. As $d(\lambda)^{-1} = O(1)$, it follows that $\boldsymbol{\alpha}_{z_0} = \lim_{N\to\infty} d(\lambda)^{-1/2} W_\lambda \mathbf{A}(z_0) = \mathbf{0}$.

When $d(\lambda) \to \infty$, we have $\boldsymbol{\gamma}_{z_0} = -\lim_{N\to\infty} \mathbf{A}(z_0)/d(\lambda)^{1/2} = \mathbf{0}$, as $A_k(z_0)$ is uniformly bounded. Hence, $\boldsymbol{\Sigma}_{12}^* = \boldsymbol{\Sigma}_{21}^* = \mathbf{0}$ and $\Sigma_{22}^* = \sigma^2 \sigma_{z_0}^2$.

Acknowledgements. We thank Co-Editor Runze Li, an Associate Editor and two referees for helpful comments that lead to important improvements on the paper.

Guang Cheng was on sabbatical at Princeton while part of this work was carried out; he would like to thank the Princeton ORFE department for its hospitality.

SUPPLEMENTARY MATERIAL

Supplement to "A partially linear framework for massive heterogeneous data" (DOI: 10.1214/15-AOS1410SUPP; .pdf). We provide the detailed proofs in the supplement.

REFERENCES

AITKIN, M. and RUBIN, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **47** 67–75.

BACH, F. (2012). Sharp analysis of low-rank kernel matrix approximations. Preprint. Available at arXiv:1208.2015.

BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907

BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. MR0348906

BIRMAN, M. S. and SOLOMYAK, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes w_p^α. *Mat. Sb.* **115** 331–355.

CHEN, X. and XIE, M. (2012). A split-and-conquer approach for analysis of extraordinarily large data, Technical Report 2012-01, Dept. Statistics, Rutgers Univ., Piscataway, NJ.

CHENG, G. and SHANG, Z. (2015). Joint asymptotics for semi-nonparametric regression models with partially linear structure. *Ann. Statist.* **43** 1351–1390. MR3346706

CHENG, G., ZHANG, H. H. and SHANG, Z. (2015). Sparse and efficient estimation for partial spline models with increasing dimension. *Ann. Inst. Statist. Math.* **67** 93–127. MR3297860

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

FAN, J. and ZHANG, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518. MR1742497

FIGUEIREDO, M. A. and JAIN, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.

GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. Springer, New York. MR3025869

GUO, W. (2002). Inference in smoothing spline analysis of variance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 887–898. MR1979393

HÄRDLE, W., LIANG, H. and GAO, J. (2000). *Partially Linear Models*. Physica, Heidelberg. MR1787637

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. Roy. Statist. Soc. Ser. B* **55** 757–796. MR1229881

HUANG, J. and ZHANG, T. (2010). The benefit of group sparsity. *Ann. Statist.* **38** 1978–2004. MR2676881

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. (2012). The big data bootstrap. Preprint. Available at arXiv:1206.6415.

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York. MR2724368

KRASIKOV, I. (2004). New bounds on the Hermite polynomials. *East J. Approx.* **10** 355–362. MR2076893

LAFFERTY, J. and LEBANON, G. (2005). Diffusion kernels on statistical manifolds. *J. Mach. Learn. Res.* **6** 129–163. MR2249817

LI, R. and LIANG, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36** 261–286. MR2387971

LI, R., LIN, D. K. J. and LI, B. (2013). Statistical inference in massive data sets. *Appl. Stoch. Models Bus. Ind.* **29** 399–409. MR3117826

MAMMEN, E. and VAN DE GEER, S. (1997). Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* **25** 1014–1035. MR1447739

MCDONALD, R., HALL, K. and MANN, G. (2010). Distributed training strategies for the structured perceptron. In *Human Language Technologies*: *The* 2010 *Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, CA.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

MEINSHAUSEN, N. and BÜHLMANN, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879

MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 29–43. Springer, Berlin. MR2040403

NARDI, Y. and RINALDO, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605–633. MR2426104

OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Union support recovery in high-dimensional multivariate regression. In 46*th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, Allerton House, UIUC, IL.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2014). Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *J. Mach. Learn. Res.* **15** 335–366. MR3190843

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720

SAUNDERS, C., GAMMERMAN, A. and VOVK, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the* 15*th International Conference on Machine Learning* (*ICML*-1998). Morgan Kaufmann, San Mateo, CA.

SHANG, Z. and CHENG, G. (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist.* **41** 2608–2638. MR3161439

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge.

SOLLICH, P. and WILLIAMS, C. K. (2005). Understanding Gaussian process regression using the equivalent kernel. In *Deterministic and Statistical Methods in Machine Learning* 211–228. Springer, Berlin.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

STEINWART, I., HUSH, D. R., SCOVEL, C. et al. (2009). Optimal rates for regularized least squares regression. In *Conference on Learning Theory*. Montreal, Canada.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

WANG, Y. (2011). *Smoothing Splines*: *Methods and Applications*. CRC Press, Boca Raton, FL. MR2814838

WANG, X. and DUNSON, D. B. (2013). Parallel mcmc via weierstrass sampler. Preprint. Available at arXiv:1312.4605.

YATCHEW, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge Univ. Press, Cambridge.

ZHANG, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17** 2077–2098. MR2175849

ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*. Princeton, NJ.

ZHAO, T., CHENG, G. and LIU, H. (2016). Supplement to "A partially linear framework for massive heterogeneous data." DOI:10.1214/15-AOS1410SUPP.

T. ZHAO
H. LIU
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: tianqi@princeton.edu
hanliu@princeton.edu

G. CHENG
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47906
USA
E-MAIL: chengg@purdue.edu