Published in final edited form as:
Ann Stat. 2016 August ; 44(4): 1400–1437. doi:10.1214/15-AOS1410.

A PARTIALLY LINEAR FRAMEWORK FOR MASSIVE HETEROGENEOUS DATA

Tianqi Zhao[‡], **Guang Cheng**[§] **[Associate Professor]**, and **Han Liu**[‡]

[‡]Department of operations research, and financial engineering, Princeton University, Princeton, New Jersey 08544, USA

[§]Department of Statistics, Purdue University, West Lafayette, IN 47906, USA

Abstract

We consider a partially linear framework for modelling massive heterogeneous data. The major goal is to extract common features across all sub-populations while exploring heterogeneity of each sub-population. In particular, we propose an aggregation type estimator for the commonality parameter that possesses the (non-asymptotic) minimax optimal bound and asymptotic distribution as if there were no heterogeneity. This oracular result holds when the number of sub-populations 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 sub-populations. All the above results require to regularize each sub-estimation as though it had the entire sample size. 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 the statistical inferences for the general kernel ridge regression. Thorough numerical results are also provided to back up our theory.

Keywords

partially linear model; heterogenous data; massive data; reproducing kernel Hilbert space; joint asymptotics; mean square error; bias propagation

1. Introduction

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

tianqi@princeton.edu
hanliu@princeton.edu
chengg@purdue.edu

AMS 2000 subject classifications: Primary 62G20, 62F25; secondary 62F10, 62F12

SUPPLEMENTARY MATERIAL
Supplementary material for: A Partially Linear Framework for Massive Heterogenous Data (DOI: To Be Assigned; .pdf). We provide the detailed proof in the supplement.

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

where ε has zero mean and known variance σ^2. In the above model, Y depends on $\boldsymbol{X}$ through a linear function that may vary across all sub-populations, and depends on Z through a nonlinear function that is common to all sub-populations. The possibly different values of $\boldsymbol{\beta}_0^{(j)}$ are viewed as the source of heterogeneity. We also point out that the sample sizes in some sub-populations could be extremely high in reality. Note that (1.1) is a typical "semi-nonparametric" model (Cheng and Shang, 2013) since we want to 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 $Z, X_1, X_2,...,X_p$. It is known from biological knowledge that the dependence structure between Y and Z (e.g., blood pressure) should be homogeneous for all human. 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 modelling 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, e.g., Bag of Little Bootstraps (Kleiner et al., 2012), and parallel MCMC computing (Wang and Dunson, 2013), or nonparametric minimaxity (Zhang et al., 2013). The other relevant work includes high dimensional linear models with variable selection (Chen and Xie, 2012) and structured perceptron (McDonald et al., 2010). Heterogenous data are often handled by fitting mixture models (Aitkin and Rubin, 1985; McLachlan and Peel, 2004; Figueiredo and Jain, 2002), time varying coefficient models (Hastie and Tibshirani, 1993; Fan and Zhang, 1999) or multitask regression (Huang and Zhang, 2010; Nardi and Rinaldo, 2008; Obozinski et al., 2008). The recent high dimensional work includes Stäadler et al. (2010); Meinshausen and Bühlmann (2014). 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 sub-populations while exploring heterogeneity of each sub-population. Specifically, we employ a simple aggregation procedure, which averages commonality estimators across all sub-populations, and then construct a plug-in estimator for each heterogeneity parameter based on the combined estimator for commonality. The secondary goal is to apply the divide-and-conquer method to the sub-population having 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. 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, the partial smoothing spline model (Wahba, 1990) can be viewed a special case. An important technical contribution of this paper is that a (point-wise) limit distribution of the KRR estimate is established by generalizing the smoothing spline inference results in Cheng and Shang (2013). This theoretical innovation makes our work go beyond the existing statistical study on the KRR estimation in large datastes, which mainly focus on their nonparametric minimaxity, e.g., Zhang et al. (2013); Bach (2012); Raskutti et al. (2014).

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), i.e., $\beta_0^{(j)}$'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 can "filter out" the heterogeneity in data when s does not grow too fast with N. Interestingly, 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, i.e., oracle rule. Our second contribution is to test the heterogeneity among a large number of sub-populations by employing the recent Gaussian approximation theory (Chernozhukov et al., 2013). The above results directly apply to the divide-and-conquer approach that deals with the sub-population with a huge sample size. In this case, the "oracle estimate" is defined as those computed based on the entire (homogeneous) data in those sub-populations. 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 setting 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 sub-populations/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 sub-estimation as if it had the entire sample.

In the end, we highlight two key technical challenges: (i) nontrivial interaction 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. (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 aggregated empirical processes. These delicate theoretical analysis 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 non-asymptotic 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.

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 $\mathscr{F}\left(\mathscr{S},\right)$ a vector space of functions from a general set $\mathscr{S}$ to . We say that $\mathscr{H}^*$ is a reproducing kernel Hilbert space (RKHS) on $\mathscr{S}$, provided that:

- **(i)** $\mathscr{H}^*$ is a vector subspace of $\mathscr{F}\left(\mathscr{S},\right)$;
- **(ii)** $\mathscr{H}^*$ is endowed with an inner product, denoted as $\langle\cdot,\cdot\rangle_{\mathscr{H}^*}$, under which it becomes a Hilbert space;
- **(iii)** for every $y \in \mathscr{S}$, the linear evaluation functional defined by $E_y(f) = f(y)$ is bounded.

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

Denote $U := (\boldsymbol{X}, Z) \in \mathscr{X} \times \mathscr{Z} \subset {}^p \times$, 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 $\mathscr{S}=\mathscr{Z}$ and $\mathscr{F}\left(\mathscr{Z},\right)=L_2\left(\mathbb{P}_Z\right)$, then we can define a RKHS $\mathscr{H} \subset L_2\left(\mathbb{P}_Z\right)$ (endowed with a proper inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$), in which the true function f_0 is believed to lie. The corresponding kernel for $\mathscr{H}$ is denoted by K such that for any $z \in \mathscr{Z}$: $f\left(z\right)=\langle f, K_z\rangle_{\mathscr{H}}$. By Mercer theorem, this kernel function has the following unique eigen-decomposition:

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

where $\mu_1 \geq \mu_2 \geq ... > 0$ are eigenvalues and $\{\phi_\ell\}_{\ell=1}^{\infty}$ are an orthonormal basis in $L_2(\mathbb{P}_Z)$. Let $\{\theta_\ell\}_{\ell=1}^{\infty}$ be the Fourier coefficient of f_0 under the basis $\{\phi_\ell\}_{\ell=1}^{\infty}$ (given the $L_2(\mathbb{P}_Z)$-inner product $\langle\cdot,\cdot\rangle_{L_2(\mathbb{P}_Z)}$). Mercer theorem together with the reproducing property implies that $\langle\phi_i, \phi_j\rangle_{\mathscr{H}} = \delta_{ij}/\mu_i$, where δ_{ij} is the 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 $K(\boldsymbol{z}_1, \boldsymbol{z}_2) = \langle\boldsymbol{z}_1, \boldsymbol{z}_2\rangle_{\mathbb{R}^d}$ has rank d, and generates a d-dimensional linear function space. The eigenfunctions are given by $\phi_\ell(\boldsymbol{z}) = z_\ell$ 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 $\mu_\ell \asymp c_1 \quad exp(-c_2 \ell^p)$ 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)

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

for $\ell = 1, 2, \cdots$, where $H_\ell(\cdot)$ is the $\ell\text{- th}$ 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 Solomjak, 1967). In particular, the eigenfunctions of a ν-th 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:

$$\left(\hat{\boldsymbol{\beta}}^{\dagger}, \hat{f}^{\dagger}\right) = \underset{(\boldsymbol{\beta}, \mathbf{f}) \in \mathscr{A}}{argmin} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \mathbf{f}(\mathbf{Z_i})\right)^2 + \lambda\|f\|_{\mathscr{H}}^2\right\}, \qquad (2.2)$$

where $\lambda > 0$ is a regularization parameter and $\mathscr{A}$ is defined as the parameter space $^p \times \mathscr{H}$. For simplicity, we do not distinguish $m = (\boldsymbol{\beta}, \mathbf{f}) \in \mathscr{A}$ from its associated function $m \in \mathscr{M} := \left\{\boldsymbol{x}^T\boldsymbol{\beta} + \mathbf{f}(\mathbf{z}) : (\boldsymbol{\beta}, \mathbf{f}) \in \mathscr{A} \quad \text{and} \quad (\boldsymbol{x}, \mathbf{z}) \in \mathscr{X} \times \mathscr{Z}\right\}$ throughout the paper. We call $\left(\hat{\boldsymbol{\beta}}^{\dagger}, \hat{f}^{\dagger}\right)$ as partially linear kernel ridge regression (KRR) estimate in comparison with the nonparametric KRR estimate in Shawe-Taylor and Cristianini (2004). In particular, when

$\mathscr{H}$ is a ν-th order Sobolev space endowed with $\left\langle f, \widetilde{f} \right\rangle_{\mathscr{H}} := \int_{\mathscr{Z}} f^{(\nu)}(z)\, \widetilde{f}^{(\nu)}(z)\, dz$, $\left(\hat{\boldsymbol{\beta}}^{\dagger}, \hat{f}^{\dagger}\right)$ becomes the commonly used partial smoothing spline estimate.

We next illustrate that $\mathscr{A}$ can be viewed as a partially linear extension of $\mathscr{H}$ in the sense that it shares some nice reproducing properties as this RKHS $\mathscr{H}$ under the following inner product:

$$\langle m, \widetilde{m} \rangle_{\mathscr{A}} = \left\langle \boldsymbol{X}^{T}\boldsymbol{\beta} + \mathbf{f}(\mathbf{Z}), \boldsymbol{X}^{\mathbf{T}}\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{f}}(\mathbf{Z}) \right\rangle_{L_2(\mathbb{P}_U)} + \lambda \left\langle f, \widetilde{f} \right\rangle_{\mathscr{H}}, \qquad (2.3)$$

where $m = (\boldsymbol{\beta}, \mathbf{f}) \in \mathscr{A}$ and $\widetilde{m} = \left(\widetilde{\boldsymbol{\beta}}, \widetilde{f}\right) \in \mathscr{A}$. Similar as the kernel function K_Z, we can construct a linear operator $R_u(\cdot) \in \mathscr{A}$ such that $\langle R_u, m \rangle_{\mathscr{A}} = m(u)$ for any $u \in \mathscr{X} \times \mathscr{Z}$. Also, construct another linear operator $P_{\lambda}: \mathscr{A} \mapsto \mathscr{A}$ such that $\langle P_{\lambda} m, \widetilde{m} \rangle_{\mathscr{A}} = \lambda \left\langle f, \widetilde{f} \right\rangle_{\mathscr{H}}$ 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 rational behind the definition of $\langle \cdot, \cdot \rangle_{\mathscr{A}}$. Denote $\otimes$ as the outer product on $\mathscr{A}$. Hence, $\mathbb{E}_U [R_U \otimes R_U] + P_{\lambda}$ is an operator from $\mathscr{A}$ to $\mathscr{A}$.

Proposition 2.2

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

Proof—For any $m(\boldsymbol{\beta}, \mathbf{f}) \in \mathscr{A}$ and $\widetilde{m} = \left(\widetilde{\boldsymbol{\beta}}, \widetilde{f}\right) \in \mathscr{A}$, we have

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

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

As will be seen in the subsequent analysis, e.g., in Theorem 3.3, the operator $[R_U \otimes R_U] + P_{\lambda}$ 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 $\hat{m} = \left(\hat{\boldsymbol{\beta}}, \hat{f}\right)$ 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 $\mathscr{H}$:

$$\left\langle f, \widetilde{f} \right\rangle_{\mathscr{C}} = \left\langle f, \widetilde{f} \right\rangle_{L_2(\mathbb{P}_Z)} + \lambda \left\langle f, \widetilde{f} \right\rangle_{\mathscr{H}}, \qquad (2.4)$$

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

$$\left\langle W_\lambda f, \widetilde{f} \right\rangle_{\mathscr{C}} = \lambda \left\langle f, \widetilde{f} \right\rangle_{\mathscr{H}} \quad \text{for any} \quad f, \widetilde{f} \in \mathscr{H}', \tag{2.5}$$

whose existence is proven in Lemma A.2. Since $\|f\|^2_{L_2(\mathbb{P}_Z)} = \sum_{i=1}^{\infty} \theta_i^2$ and $\|f\|^2_{\mathscr{H}} = \sum_{i=1}^{\infty} \theta_i^2 / \mu_i$, we now have $\|f\|^2_{\mathscr{C}} = \sum_{i=1}^{\infty} \theta_i^2 (1 + \lambda / \mu_i)$ by (2.4). We next define two crucial quantities needed in the construction: $B_k := [X_k | Z]$ and its Riesz representer $A_k \in \mathscr{H}$ satisfying $\langle A_k, f \rangle_{\mathscr{C}} = \langle B_k, f \rangle_{L_2(\mathbb{P}_Z)}$ for all $f \in \mathscr{H}$. Here, we implicity assume B_k is square integrable. The existence of A_k follows from the boundedness of the linear functional $\mathscr{B}_k f := \langle B_k, f \rangle_{L_2(\mathbb{P}_Z)}$ (by Riesz's representer theorem) as follows:

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

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

Proposition 2.3

For any $u = (\boldsymbol{x}, z)$, R_u can be expressed as $R_u : u \mapsto (L_u, N_u) \in \mathscr{A}$, where

$$L_u = (\Omega + \Sigma_\lambda)^{-1} (\boldsymbol{x} - \boldsymbol{A}(z)) \quad \text{and} \quad N_u = \widetilde{K}_z - \boldsymbol{A}^T L_u,$$

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

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

Notation—Denote $\|\cdot\|_2$ and $\|\cdot\|_\infty$ as the Euclidean L_2 and infinity norm in p, respectively. For any function $f : \mathscr{Z} \mapsto$, let $\|f\|_{sup} = sup_{z \in \mathscr{Z}} |f(z)|$. 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 constant $c > 0$ ($c' > 0$) independent of n such that $a_n \le cb_n$ ($a_n \ge c' b_n$) for all $n \in$. We denote $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We define h as the inverse of $\sum_{\ell=1}^{\infty} 1 (1 + \lambda / \mu_\ell)$, which is known to be the "effective dimension" of a kernel K $(\text{w.r.t.} \quad L_2(\mathbb{P}_Z))$ (Zhang, 2005). For any function space $\mathscr{F}$, define

$$\omega\left(\mathscr{F},\delta\right)=\int_0^\delta \sqrt{log\, \mathscr{N}\left(\mathscr{F},\|\cdot\|_{sup},\epsilon\right)}d\epsilon,$$

where $\mathscr{N}\left(\mathscr{F},\|\cdot\|_{sup},\epsilon\right)$ is an $\in$-covering number of $\mathscr{F}$ w.r.t. supreme norm. Define the following sets of functions: $\mathscr{F}_1=\left\{f\left(\boldsymbol{x}\right)=\boldsymbol{x}^T\beta{:}\|f\|_{sup}\le 1 \quad \text{for} \quad \text{all} \quad \boldsymbol{x}\in\mathscr{X},\beta\in{}^{p}\right\}$, $\mathscr{F}_2=\left\{f\in\mathscr{H}{:}\|f\|_{sup}\le 1, \|f\|_{\mathscr{H}}\le h^{1/2}\lambda^{-1/2}\right\}$, $\mathscr{F}:=\left\{f=f_1+f_2{:}f_1\in\mathscr{F}_1, f_2\in\mathscr{F}_2, \|f\|_{sup}\le 1/2\right\}$.

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 (non-asymptotic) minimax optimal bound (in terms of mean-squared error) and asymptotic distribution as the "oracle estimate" $\hat{f}_{or}$ computed when all the heterogeneity information are available:

$$\hat{f}_{or}=\underset{f\in\mathscr{H}}{argmin}\left\{\frac{1}{N}\sum_{j=1}^{s}\sum_{i\in L_j}\left(Y_i-\left(\boldsymbol{\beta}_0^{(j)}\right)^T\boldsymbol{X}_i-f\left(Z_i\right)\right)^2+\lambda\|f\|_{\mathscr{H}}^2\right\}. \quad (3.1)$$

The above nice properties hold when the number of sub-populations 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 sub-population is not too small. In the end, Section 3.6 tests the possible heterogeneity among a large number of sub-populations.

3.1. Method and Assumptions

The heterogeneous data setup and averaging procedure are described below:

1. Obverse data ($\boldsymbol{X}_i$, Z_i, Y_i) with the known label $L_i \in$ {1, 2,...,*s* indicating the sub-population it belongs to, for i = 1,...,*N*. The size of samples from each sub-population is assumed to be the same, denoted by *n*, for simplicity. Hence, $N = n \times s$.

2. On the *j*-th sub-population, obtain the following penalized estimator:

$$\left(\hat{\boldsymbol{\beta}}_{n,\lambda}^{(j)},\hat{f}_{n,\lambda}^{(j)}\right)=\underset{(\boldsymbol{\beta},\mathbf{f})\in\mathscr{A}}{argmin}\left\{\frac{1}{n}\sum_{i\in L_j}\left(\boldsymbol{X}_i^T\boldsymbol{\beta}+\mathbf{f}\left(\mathbf{Z_i}\right)-\mathbf{Y_i}\right)^2+\lambda\|f\|_{\mathscr{H}}^2\right\}. \quad (3.2)$$

3. Obtain the final nonparametric estimate[1] for commonality by averaging:

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

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

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. 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 $\Omega := \left[(\boldsymbol{X}-\boldsymbol{B}(Z))(\boldsymbol{X}-\boldsymbol{B}(Z))^T\right]$ is positive definite; (iii) 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 Ω is needed for obtaining semiparametric efficient estimation; see Mammen and van de Geer (1997). Note that we do not require the independence between $\boldsymbol{X}$ 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 $sup_\ell \|\phi_\ell\|_{sup} \leq c_\phi$ and $\sup_z K(z, z) \leq c_K$.

Assumption 3.2 is commonly assumed in kernel ridge regression literature (Zhang et al., 2013; Lafferty and Lebanon, 2005; Guo, 2002). In the case of finite rank kernel, e.g., 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. Lastly, for the polynomially decaying kernels, Proposition 2.2 in Shang and Cheng (2013) showed that the eigenfunctions induced from a ν-th order Sobolev space (under a proper inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}}$) are uniformly bounded under mild smoothness conditions for the density of Z.

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

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

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

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 the nonparametric component. We call this interaction as the "bias propagation phenomenon", and study it in Section 3.4.

3.2. Non-Asymptotic Bound for Mean-Squared Error

The primary goal of this section is to evaluate the estimation quality of the combined estimate from a *non-asymptotic* point of view. Specifically, we derive a finite sample upper bound for the mean-squared error $MSE\left(\bar{f}\right) := \left[\| \bar{f} - f_0 \|^2_{L_2(\mathbb{P}_Z)}\right]$. When s does not grow too fast, we show that MSE($\bar{f}$) is of the order $O\left((Nh)^{-1} + \lambda\right)$, 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 (un-improvable) optimal minimax rate. As a by-product, we establish a *non-asymptotic* upper bound for the mean-squared error of $\hat{\beta}^{(j)}$, i.e., $MSE\left(\hat{\beta}^{(j)}\right) := \left[\|\hat{\beta}^{(j)} - \beta_0^{(j)}\|_2^2\right]$. The results in this section together with Theorem 3.4 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 $\hat{f}_{or}$.

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

Theorem 3.1

Suppose that Assumptions 3.1–3.3 hold. If $s = o\left(Nh^2(\omega(\mathscr{F},1) + log\, N)^{-2}\right)$, then we have

$$MSE\left(\bar{f}\right) \le C_1 \sigma^2 (Nh)^{-1} + 2\|f_0\|_{\mathscr{H}}^2 \lambda + C_2 \lambda^2 + s^{-1} a(n, s, h, \lambda, \omega), \tag{3.4}$$

where $a(n, s, h, \lambda, \omega) = Ch^{-1} n^{-1} r_{n,s}^2 \left(\omega(\mathscr{F}, 1)^2 + 1\right) + Ch^{-2} \lambda^{-1} n\, exp\left(-c\, log^2 N\right)$, $r_{n,s} = (nh)^{-1/2} \log^2 N + \lambda$ and C is some generic constant.

Typically, we require an upper bound for s so that the third term in the R.H.S. of (3.4) can be dominated by the first two terms, which correspond to variance and bias, respectively. Hence, we choose $\lambda \asymp (Nh)^{-1}$ to attain the optimal *bias-variance trade-off*. The resulting rate coincides with the minimax optimal rate of the oracle estimate in different RKHS; see Section 4. This can be viewed as a non-asymptotic version of the "oracle property" of $\bar{f}$. In comparison with the nonparametric KRR result in Zhang et al. (2013), we realize that adding one parametric component does not affect the finite sample upper bound (3.4).

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

Theorem 3.2—Suppose that Assumptions 3.1 – 3.3 hold. Then we have

$$MSE\left(\hat{\boldsymbol{\beta}}^{(j)}\right) \le C_1'\sigma^2 n^{-1} + C_2'\lambda^2 + a\left(n, s, h, \lambda, \omega\right), \qquad (3.5)$$

where $a(n, s, h, \lambda, \omega)$ is defined in Theorem 3.1.

Again, the first term and second term in the R.H.S. of (3.5) 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 $MSE\left(\hat{\boldsymbol{\beta}}^{(j)}\right)$, i.e., $O(n^{-1/2})$, but may lose the minimax optimality of 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 $\left(\hat{\boldsymbol{\beta}}^{(j)}, \bar{f}\left(z_0\right)\right)$ at any $z_0 \in \mathscr{Z}$. 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 3.3 (Joint Asymptotics I)

Suppose that Assumptions 3.1 and 3.2 hold, and that as $N \to \infty, h\|\widetilde{K}_{z_0}\|^2_{L_2(\mathbb{P}_Z)} \to \sigma^2_{z_0}$, $h^{1/2}\left(W_\lambda \boldsymbol{A}\right)\left(z_0\right) \to \alpha_{z_0} \in {}^p$, and $h^{1/2}\boldsymbol{A}\left(z_0\right) \to -\gamma_{z_0} \in {}^p$. Suppose the following conditions are satisfied

$$s = o\left(Nh^2(\omega\left(\mathscr{F}, 1\right) + log\, N)^{-2}\right), \qquad (3.6)$$

$$s(Nh)^{-1}\, log^4\, N + \lambda = o\left(h^2(\omega\left(\mathscr{F}, 1\right) + log\, N)^{-2}\, log^{-2}\left(N\right)\right). \qquad (3.7)$$

Denote $\left(\beta_0^{(j)*}, f_0^*\right)$ as $\left(id - P_\lambda\right) m_0^{(j)}$, where $m_0^{(j)} = \left(\beta_0^{(j)}, f_0\right)$. We have for any $z_0 \in Z$ and j = 1,...,s,

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

$$\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\right. \\ \sqrt{Nh}\left(\bar{f}\ (z_0) - f_0^*\ (z_0)\right) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}}\begin{pmatrix} \Omega^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}\right), \tag{3.8}$$

where $\Sigma_{22}=\sigma_{z_0}^2+2\gamma_{z_0}^T\Omega^{-1}\alpha_{z_0}+\gamma_{z_0}^T\Omega^{-1}\gamma_{z_0}$;

(ii) if s is fixed, then

$$\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\right) \\ \sqrt{Nh}\left(\bar{f}\ (z_0) - f_0^*\ (z_0)\right) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}}\begin{pmatrix} \Omega^{-1} & s^{-1/2}\Sigma_{21} \\ s^{-1/2}\Sigma_{12} & \Sigma_{22} \end{pmatrix}\right), \tag{3.9}$$

where $\Sigma_{12}=\Sigma_{12}^T=\Omega^{-1}\ (\alpha_{z_0}+\gamma_{z_0})$.

Part (i) of Theorem 3.3 says that $\sqrt{n}\hat{\beta}^{(j)}$ and $\sqrt{Nh}\ \bar{f}\ (z_0)$ are asymptotically independent as $s \to \infty$. This is not surprising since only samples in one sub-population (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, i.e., Assumption 3.3. Interestingly, given this additional condition, even when s is fixed, these two estimators can still achieve the asymptotic independence if $h \to 0$. Please see more details in next section.

3.4. Bias Propagation

In this section, we first analyze the source of estimation bias observed in the joint asymptotics Theorem 3.3. 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 $\hat{f}_{or}$, i.e., 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 $\in {}^{n\times p}$ and $\in {}^{n}$ as the designs based on the samples from the jth sub-population and let $\varepsilon^{(j)}=[\epsilon_i]_{i\in L_j} \in {}^{n}$. The first order optimality condition (w.r.t. $\boldsymbol{\beta}$) gives

$$\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}=\left({}^T\right)^{-1}{}^T\boldsymbol{\varepsilon}^{(j)} - \left({}^T\right)^{-1}{}^T\left(\hat{f}^{(j)}\ () - f_0\ ()\right), \tag{3.10}$$

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

$$\text{parametric bias:} L_\lambda f_0 = -\left(\Omega + \Sigma_\lambda\right)^{-1} \langle \boldsymbol{B}, W_\lambda f_0 \rangle_{L_2\left(\mathbb{P}_Z\right)}, \quad (3.11)$$

$$\text{nonparametric bias:} N_\lambda f_0 = W_\lambda f_0 - \boldsymbol{A}^T L_\lambda f_0 \quad (3.12)$$

according to Proposition 2.3. The first term in (3.12) explains the bias introduced by penalization; see (2.5). This bias propagates to the parametric component through $\boldsymbol{B}$, as illustrated in (3.11). The parametric bias $L_\lambda f_0$ propagates back to the nonparametric component through the second term of (3.12). Therefore, by strengthening $B_k \in L_2\left(\mathbb{P}_Z\right)$ to $B_k \in \mathscr{H}$, i.e., Assumption 3.3, it can be shown that the order of $L_\lambda f_0$ in (3.11) 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 3.4. (Joint Asymptotics II)—Suppose Assumption 3.3 and the conditions in Theorem 3.3 hold. If we choose $\lambda = o\left((Nh)^{-1/2} \wedge n^{-1/2}\right)$, then

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

$$\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\right) \\ \sqrt{Nh}\left(\bar{f}\left(z_0\right) - f_0\left(z_0\right) - W_\lambda f_0\left(z_0\right)\right) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}}\begin{pmatrix} \Omega^{-1} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22}^* \end{pmatrix}\right), \quad (3.13)$$

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

(ii) if s is fixed, then

$$\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\right) \\ \sqrt{Nh}\left(\bar{f}\left(z_0\right) - f_0\left(z_0\right) - W_\lambda f_0\left(z_0\right)\right) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}}\begin{pmatrix} \Omega^{-1} & s^{-1/2}\Sigma_{21}^* \\ s^{-1/2}\Sigma_{12}^* & \Sigma_{22}^* \end{pmatrix}\right), \quad (3.14)$$

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

Moreover, if $h \rightarrow 0$, then $\Sigma_{12}^{*}=\Sigma_{21}^{*}=\mathbf{0}$ 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, e.g., Shang and Cheng (2013). We will illustrate this point in Section 4.

By examining the proof for case (ii) of Theorem 3.4 (and taking $s = 1$), we know that the oracle estimate $\bar{f}_{or}$ defined in (3.1) attains the same asymptotic distribution as that of $\bar{f}$ in (3.13) 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 3.4 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 sub-estimation 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 sub-estimation; see Section 3.2.

Cheng and Shang (2013) have recently uncovered a *joint asymptotics phenomenon* in partial smoothing spline models: parametric estimate and (point-wise) nonparametric smoothing spline estimate become asymptotically independent after the parametric bias is removed. This corresponds to a special case of Part (ii) of Theorem 3.4 for polynomially decaying kernels with $s = 1$ and $h \rightarrow \infty$. Therefore, case (ii) in Theorem 3.4 generalizes this new phenomenon to the partially linear kernel ridge regression models. When $h \nrightarrow 0$, e.g., $h \asymp r^{-1}$ for finite rank kernel, the semi-nonparametric estimation in consideration essentially reduces to a parametric one. Hence, it is not surprising that the asymptotic dependence remains.

Remark 3.1—Theorem 3.4 implies that $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0^{(j)}\right) \rightsquigarrow N\left(\mathbf{0},\sigma^{\mathbf{2}}\mathbf{\Omega}^{\mathbf{-1}}\right)$ when $\lambda = o(n^{-1/2})$. When the error $\in$ follows a Gaussian distribution, it is well known that $\hat{\boldsymbol{\beta}}^{(j)}$ achieves the semiparametric efficiency bound (Kosorok, 2007). Hence, the semiparametric efficient estimate can be the 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 is one important feature of massive data: strength-borrowing.

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 non-asymptotic senses when s does not grow too fast and λ is chosen according to the entire sample size. In this section, we employ $\bar{f}$ to boost the estimation efficiency of $\hat{\boldsymbol{\beta}}^{(j)}$ from semiparametric level to parametric level. This leads to our final estimate for heterogeneity, i.e., $\check{\boldsymbol{\beta}}^{(j)}$ defined in (3.15). More importantly, $\check{\boldsymbol{\beta}}^{(j)}$ possesses the limit distribution as if the commonality in each sub-population were known, and hence satisfies the “oracle rule”. This interesting efficiency boosting phenomenon will be empirically verified in Section 6.

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

$$\check{\beta}^{(j)}=\underset{\boldsymbol{\beta}\in\mathbb{R}^{\mathrm{P}}}{argmin}\frac{1}{n}\sum_{i\in L_j}\left(Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}-\bar{\mathbf{f}}\left(\mathbf{Z_i}\right)\right)^2. \qquad (3.15)$$

Theorem 3.5 below shows that $\check{\beta}^{(j)}$ 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, i.e., (3.16), and thus the homogeneous data setting is trivially excluded. This lower bound requirement slows down the convergence rate of $\check{\beta}^{(j)}$, i.e., $\sqrt{n}$, such that $\bar{f}$ can be treated as if it were known.

Theorem 3.5—Suppose Assumption 3.1 and 3.2 hold. If s satisfies

$$s^{-1}=o\left(h^2\,log^{-4}\,N\right), \qquad (3.16)$$

$$s=o\left(Nh^2(\omega\left(\mathscr{F},1\right)+log\,N)^{-2}\right), \qquad (3.17)$$

$$s(Nh)^{-1}\,log^4\,N+\lambda=o\left(h^2(\omega\left(\mathscr{F},1\right)+log\,N)^{-2}\,log^{-2}\,N\right), \qquad (3.18)$$

and we choose $\lambda = o((Nh)^{-1})$, then we have

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

where $\Sigma=\left[\boldsymbol{X}\boldsymbol{X}^T\right]$.

Recall that $\boldsymbol{X}$ and Z are not assumed to be independent. Hence, the parametric efficiency bound Σ^{-1} is not larger than the semiparametric efficiency bound Ω^{-1}.

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 sub-populations may share the same underlying distribution. In this section, we consider testing for the heterogeneity among sub-populations. 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 sub-populations.

Consider a general class of pairwise heterogeneity testing:

$$H_0{:}Q\left(\boldsymbol{\beta}_0^{(j)}-\boldsymbol{\beta}_0^{(k)}\right)=\mathbf{0} \quad \text{for} \quad \mathbf{j}\neq\mathbf{k}, \qquad (3.19)$$

where $Q=\left(Q_1^T,\ldots,Q_q^T\right)^T$ is a $q \times q$ matrix with $q \geq p$. The general formulation (3.19) 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 $\hat{\boldsymbol{\beta}}$ or its improved version $\check{\boldsymbol{\beta}}$. Let $C_\alpha \subset {}^q$ be a confidence region satisfying $(\boldsymbol{b}\in C_\alpha)=1-\alpha$ for any $\boldsymbol{b} \sim N(0, I_q)$. Specifically, we have the following α-level Wald tests:

$$\begin{aligned}\Psi_1= & \ I\left\{Q\left(\hat{\boldsymbol{\beta}}^{(j)}-\hat{\boldsymbol{\beta}}^{(k)}\right)\notin\ \sqrt{2/n}\sigma\left(Q\Omega^{-1}Q^T\right)^{1/2}C_\alpha\right\},\\ \Psi_2= & \ I\left\{Q\left(\check{\boldsymbol{\beta}}^{(j)}-\check{\boldsymbol{\beta}}^{(k)}\right)\notin\ \sqrt{2/n}\sigma\left(Q\Sigma^{-1}Q^T\right)^{1/2}C_\alpha\right\}.\end{aligned}$$

The consistency of the above tests are guaranteed by Theorem 3.6 below. In addition, we note that the power of the latter test is larger than the former; see the analysis below Theorem 3.6. The price we need to pay for this larger power is to require a lower bound on *s*.

Theorem 3.6—Suppose that the conditions in Theorem 3.4 are satisfied. Under the null hypothesis specified in (3.19), we have

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

Moreover, under the conditions in Theorem 3.5, we have

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

where $\Sigma=\left[\boldsymbol{X}\boldsymbol{X}^T\right]$.

The larger power of $\boldsymbol{\Psi}_2$ 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 $H_0{:}\beta_{01}^{(j)}-\beta_{01}^{(k)}=0$, i.e., $Q=$ (1, 0, 0 ..., 0). In this case, we have $\Psi_1=I\left\{|\hat{\beta}_1^{(j)}-\hat{\beta}_1^{(k)}|>\sqrt{2}\sigma\left[\Omega^{-1}\right]_{11}^{1/2}z_{\alpha/2}/\sqrt{n}\right\}$, and

$\Psi_2 = I\left\{|\check{\beta}_1^{(j)} - \check{\beta}_1^{(k)}| > \sqrt{2}\sigma\left[\Sigma^{-1}\right]_{11}^{1/2} z_{\alpha/2}/\sqrt{n}\right\}$. The (asymptotic) power function under the alternative that $\beta_{01}^{(j)} - \beta_{01}^{(k)} = \beta^*$ for some non-zero $\boldsymbol{\beta}^*$ is

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

where $W \sim N(0, 1)$ and $\boldsymbol{\sigma}^*$ is $\sqrt{2}\sigma\left[\Omega^{-1}\right]_{11}^{1/2}$ for $\boldsymbol{\Psi}_1$ and $\sqrt{2}\sigma\left[\Sigma^{-1}\right]_{11}^{1/2}$ for $\boldsymbol{\Psi}_2$. Hence, a smaller $\boldsymbol{\sigma}^*$ gives rise to a larger power, and $\boldsymbol{\Psi}_2$ is more powerful than $\boldsymbol{\Psi}_1$. Please see Section 6 for empirical support for this power comparison.

We next consider a simultaneous testing that is applied to a large number of sub-populations:

$$H_0 : \boldsymbol{\beta}^{(j)} = \widetilde{\boldsymbol{\beta}}^{(j)} \quad \text{for all} \quad j \in \mathscr{G}, \qquad (3.20)$$

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

$$H_1 : \boldsymbol{\beta}^{(j)} \neq \widetilde{\boldsymbol{\beta}}^{(j)} \quad \text{for some} \quad j \in \mathscr{G}. \qquad (3.21)$$

The above $\hat{\boldsymbol{\beta}}^{(j)}$'s are pre-specified for each $j \in \mathscr{G}$. If all $\widetilde{\boldsymbol{\beta}}^{(j)}$'s are the same, then it becomes a type of heterogeneity test for the group of sub-populations indexed by $\mathscr{G}$. Here we allow $|\mathscr{G}|$ to be as large as s, and thus it can increase with n. Let $\hat{\boldsymbol{\Sigma}}^{(j)}$ be the sample covariance matrix of $\boldsymbol{X}$ for the j-th sub-population, i.e., $n^{-1}\sum_{i \in L_j} \boldsymbol{X}_i \boldsymbol{X}_i^T$. Define the test statistic

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

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

$$W_{\mathscr{G}} := \max_{j \in \mathscr{G}, 1 \le k \le p} \frac{1}{\sqrt{n}} \sum_{i \in L_j} \left(\hat{\boldsymbol{\Sigma}}^{(j)}\right)_k^{-1} \boldsymbol{X}_i e_i,$$

where e_i's are i.i.d. $N(0, \sigma 2)$ independent of the data and $\left(\hat{\boldsymbol{\Sigma}}^{(j)}\right)_k^{-1}$ is the k-th row of $\left(\hat{\boldsymbol{\Sigma}}^{(j)}\right)^{-1}$. Let $c_{\mathscr{G}}(\alpha) = inf\left\{t \in : (W_{\mathscr{G}} \le t|) \ge 1 - \alpha\right\}$. We employ the recent Gaussian approximation and multiplier bootstrap theory (Chernozhukov et al., 2013) to obtain the following theorem.

Theorem 3.7—Suppose Assumptions 3.1 and 3.2 hold. In addition, suppose (3.17) and (3.18) in Theorem 3.5 hold. For any $\mathscr{G} \subset \{1, 2, \ldots, s\}$ with $|\mathscr{G}|=d$, if (i) $s \gtrsim h^{-2}\, log\,(pd)\; log^4\, N$, (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 $\lambda = o((Nh)^{-1})$, we have

$$\underset{\alpha\in(0,1)}{sup}\;|\,(T_{\mathscr{G}}>c_{\mathscr{G}}\,(\alpha))-\alpha|=o\,(1)\,.$$

Remark 3.2—We can perform heterogeneity testing even without specifying $\widetilde{\beta}^{(j)}$'s. This can be done by simply reformulating the null hypothesis as follows (for simplicity we set $\mathscr{G}=[s]$): $H_0 : \alpha^{(j)} = 0$ for $j \in [s - 1]$, where $\alpha^{(j)} = \beta^{(j)} - \beta^{(j+1)}$ for $j = 1,\ldots,s - 1$. The test statistic is $T'_{\mathscr{G}}=max_{1\le j\le s-1}\; max_{1\le k\le p}\alpha_k^{(j)}$. The bootstrap quantity is defined as

$$W'_{\mathscr{G}} := \underset{1\le j\le s-1,1\le k\le p}{max} \frac{1}{\sqrt{n}}\sum_{i\in L_j}\left(\hat{\mathbf{\Sigma}}^{(j)}\right)_k^{-1}\boldsymbol{X}_i e_i - \frac{1}{\sqrt{n}}\sum_{i\in L_{j+1}}\left(\hat{\mathbf{\Sigma}}^{(j+1)}\right)_k^{-1}\boldsymbol{X}_i e_i.$$

The proof is similar to that of Theorem 3.7 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 sub-populations 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. Rather, we present the bounds that have less complicated forms but are still sufficient to 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, $h \asymp r^{-1}$. Combining this fact with Theorem 3.4, we get the following corollary for finite rank kernels:

Corollary 4.1—Suppose Assumption 3.1 – 3.3 hold and $s \to \infty$. For any $z_0 \in \mathscr{X}$, if $\lambda = o(N^{-1/2})$, $\log(\lambda^{-1}) = o(N^2 \log^{-12} N)$ and $s=o\left(\frac{N}{\sqrt{log\ \lambda^{-1}}\ log^6\ N}\right)$, then

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

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

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 $h \asymp r^{-1}$ into (3.16). This lower bound is clearly smaller than the upper bound. Hence, the efficiency boosting is feasible.

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

Corollary 4.2

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

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

for some constant *C*.

4.2. Example II: Exponential Decay Kernel

We next consider the RKHS for which the kernel has exponentially decaying eigenvalues, i.e., $\mu_\ell=exp\left(-\alpha\ell^p\right)$ for some $\alpha > 0$. In this case, we have $h \asymp \left(log\,\lambda^{-1}\right)^{-1/p}$ by explicit calculations.

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

$$\left(\begin{array}{c} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0^{(j)}\right) \\ \sqrt{Nh}\left(\bar{f}\left(z_0\right)-f_0\left(z_0\right)\right) \end{array}\right) \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}}\left(\begin{array}{cc} \Omega^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_{z_0}^2 \end{array}\right)\right),$$

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

Corollary 4.3 implies the shrinking rate of the confidence interval for $f_0(z_0)$ as $(Nh)^{-1/2}$. This motivates us to choose λ (equivalently h) as large as possible. Plugging such a λ into the upper bound of s yields $s{=}o\left(N\,log^{-(7p+4)/p}N\right)$. For example, when $p = 1 (p = 2)$, the upper bound is $s{=}o\left(N\,log^{-11}\,N\right)\left(s{=}o\left(N\,log^{-9}\,N\right)\right)$. Note that this upper bound for s only differs from that for the finite rank kernel up to some logrithmic term. This is mainly because RKHS with exponentially decaying eigenvalues has an effective dimension (log $N)^{1/p}$ (for the above λ). Again, by (3.16) we get the lower bound of $s \gtrsim \left(log\,\lambda^{-1}\right)^{2/p} log^2$ 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

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

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

$$\sigma^2_{z_0} \le \underset{N\to\infty}{lim}\,\sigma^2 c_\phi^2 h \sum_{\ell=0}^{\infty}(1{+}\lambda\eta\,exp\,(-2\,(log\,\eta)\,\ell))^{-2}{=}c_\phi^2{\cdot}2\sigma^2\,log\,(1/\eta) \le 1.7178\sigma^2,$$

where equality follows from Lemma C.1 in Appendix C with the case $t = 2$. Hence, a (conservative) $100(1–\alpha)\%$ confidence interval for $f_0(z_0)$ is given by $\bar{f}(z_0) \pm 1.3106\sigma z_{\alpha/2}/\sqrt{Nh}$.

Corollary 4.4

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

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

We know that the above rate is minimax optimal according to Zhang et al. (2013). Note that the upper bound for s required here is similar as 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, i.e., $\mu_\ell{=}c\ell^{-2\nu}$ for some $\nu > 1/2$. Hence, we can explicitly calculate that $h = \lambda^{1/(2\nu)}$. The resulting penalized estimate is called as “partial smoothing spline” in the statistics literature; see Gu (2013); Wang (2011).

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

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

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

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

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

When N is large, the above bound reduces to $s = o\left(N^{\frac{8\nu^2-8\nu+1}{2\nu(4\nu+1)}} log^{-6} N\right)$. 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 $h \asymp \lambda^{\frac{1}{2\nu}}$ into (3.16). When $\lambda \asymp N^{-\frac{2\nu}{4\nu+2}}$, we get $s \gtrsim N^{\frac{1}{4\nu+1}} log^2 N$. 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:

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

and eigenvalues

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

Hence, Assumption 3.2 trivially holds. Under the above eigensystem, 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:

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

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

$$\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\right) \\ \sqrt{Nh}\left(\bar{f}(z_0) - f_0(z_0)\right) \end{pmatrix} \rightsquigarrow N\left(\mathbf{0}, \sigma^{\mathbf{2}} \begin{pmatrix} \Omega^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_{z_0}^2 \end{pmatrix}\right). \qquad (4.3)$$

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

Corollary 4.6—Suppose that Assumption 3.1 - 3.3 hold. If we choose $\lambda = N^{-\frac{2\nu}{2\nu+1}}$, and $s = o\left(N^{\frac{4\nu^2-4\nu+1}{4\nu^2+2\nu}} log^{-4} N\right)$, the combined estimator achieves optimal rate of convergence, i.e.,

$$\left[\| \bar{f} - f_0\|_{L_2(\mathbb{P}_Z)}^2\right] \le CN^{-\frac{2\nu}{2\nu+1}}. \qquad (4.4)$$

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 sub-populations 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 sub-populations 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.

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 $\hat{\boldsymbol{\beta}}^{(j)}$ and $\hat{f}^{(j)}$ when they are based on the *j*-th subsample. Thus, the averaged estimator is defined as

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

Comparing to the oracle estimator, the aggregation procedure reduces the computational complexity in terms of the entire sample size *N* to the sub-sample 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 and 5.2 below state conditions under which the divide-and-conquer estimators maintain the same statistical properties as oracle estimate, i.e., so-called oracle property.

Our first contribution is a non-asymptotic upper bound for MSE($\bar{f}$).

Proposition 5.1

Suppose that the conditions in Theorem 3.1 hold. We have that the divide and conquer estimator $\bar{f}$ satisfies

$$MSE\left(\bar{f}\right) \le C_1\sigma^2(Nh)^{-1} + 2\|f_0\|_{\mathscr{H}}^2\lambda + C_2\lambda^2 + s^{-1}a\,(n, s, h, \lambda, \omega)\,, \tag{5.1}$$

where $a(n, s, h, \lambda, \omega)$, C_1 and C_2 are constants defined in Theorem 3.1.

Our second contribution is on the joint asymptotic distribution under the same conditions for (s, λ) required in the heterogeneous data setting.

Proposition 5.2

Suppose that the conditions in Theorem 3.4 hold. If we choose $\lambda = o(N^{-1/2})$, then

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

where $\Sigma_{12}^* = \Sigma_{21}^{*T} = \sigma^2\Omega^{-1}\gamma_{z_0}$ and $\Sigma_{22}^* = \sigma^2\left(\sigma_{z_0}^2 + \gamma_{z_0}^T\Omega^{-1}\gamma_{z_0}\right)$. Moreover, if $h \to 0$, then $\gamma_{z_0} = \mathbf{0}$. In this case, $\Sigma_{12}^* = \Sigma_{21}^{*T} = \mathbf{0}$ and $\Sigma_{22}^* = \sigma^2\sigma_{z_0}^2$.

The conclusion of Proposition 5.2 holds no matter *s* is fixed or diverges (once the condition for *s* in Theorem 3.4 are satisfied).

In view of Propositions 5.1 and 5.2, we note that the above upper bound and joint asymptotic distribution are exactly the same as those for the oracle estimate, i.e., $s = 1$.

6. Numercial Experiment

In this section, we empirically examine the impact of the number of sub-populations on the statistical inference built on $\left(\hat{\boldsymbol{\beta}}^{(j)}, \bar{f}\right)$. 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 $Z \sim$ Uniform $(-1, 1)$ and $X = (W + Z)/2$, where $W \sim$ Unfiform$(-1,1)$, such that X and Z are dependent. It is easy to show that $\Omega = E[(X - E[X|Z])^2)] = 1/12$ and $\Sigma = E[X^2] = 1/6$. To design that heterogeneous data setting, we let $\beta_0^{(j)}=j$ for $j = 1,2,...,s$ on the j-th 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 $Beta(a_1, a_2)$.

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

$$\left[\hat{Y}^{(j)} \pm 1.96\sigma \sqrt{x_0^T\Omega^{-1}x_0/n+\sigma_{z_0}^2/\left(Nh\right)+1}\right],$$

where $\hat{Y}^{(j)} - x_0^T\hat{\boldsymbol{\beta}}^{(j)} + \bar{f}\left(z_0\right)$ is the predicted response. The unknown error variance σ is estimated by $\left(\hat{\sigma}^{(j)}\right)^2=n^{-1}\sum_{i\in L_j}\left(Y_i - X_i^T\hat{\boldsymbol{\beta}}^{(j)} - \hat{f}^{(j)}\left(Z_i\right)\right)^2/\left(n - Tr\left(A\left(\lambda\right)\right)\right)$, where $A(\lambda)$ denotes the smoothing matrix, followed by an aggregation $\bar{\sigma}^2=1/s\sum_{j=1}^{s}\left(\hat{\sigma}^{(j)}\right)^2$. 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 $s = 2^0, 2^1,...,2^{t-3}$ when $N = 2^{\mathrm{t}}$. 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 decreaes. 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 $\hat{\boldsymbol{\beta}}^{(j)}$ and $\check{\boldsymbol{\beta}}^{(j)}$, we construct the following two types of 95% confidence intervals for $\beta_0^{(j)}$.

$$
\begin{aligned}
CI_1 &= \left[\hat{\boldsymbol{\beta}}^{(j)} \pm 1.96\Omega^{-1/2} n^{-1.2}\ \bar{\sigma}\right],\\
CI_2 &= \left[\check{\boldsymbol{\beta}}^{(j)} \pm 1.96\Sigma^{-1/2} n^{-1.2}\ \bar{\sigma}\right].
\end{aligned}
$$

Obviously, CI$_2$ is shorter than CI$_1$. However, Theorem 3.5 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.

Lastly, we consider the heterogeneity testing. In Figure 5, we compare tests $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ 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 $\boldsymbol{\Delta} = 0.5, 1, 1.5$. It is clearly seen that both tests are consistent, and their powers increase as $\boldsymbol{\Delta}$ or *N* increases. In addition, we observe that $\boldsymbol{\Psi}_2$ has uniformly larger powers than $\boldsymbol{\Psi}_1$.

7. Proof of Main Results

In this section, we present main proofs of Theorem 3.1, 3.3 and 3.4 in the main text.

7.1. Proof of Theorem 3.1

Proof—We start from analyzing the minimization problem (3.2) on each sub-population. Recall $m = (\boldsymbol{\beta}, f)$ and $U = (\boldsymbol{X}, Z)$. The objective function can be rewritten as

$$
\begin{aligned}
\frac{1}{n}\sum_{i\in L_j}\left(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \mathbf{f}\left(\mathbf{Z_i}\right)\right)^2 + \lambda\|f\|_{\mathscr{H}}^2 &= \frac{1}{n}\sum_{i\in L_j}\left(Y_i - m\left(U_i\right)\right)^2 + \langle P_\lambda m, m\rangle_{\mathscr{A}}\\
&= \frac{1}{n}\sum_{i\in L_j}\left(Y_i - \langle RU_i, m\rangle_{\mathscr{A}}\right)^2 + \langle P_\lambda m, m\rangle_{\mathscr{A}}
\end{aligned}
$$

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

$$
\frac{1}{n}\sum_{i\in L_j} R_{U_i}\left(\hat{m}^{(j)}\left(U_i\right) - Y_i\right) + P_\lambda \hat{m}^{(j)} = 0,
$$

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

$$
-\frac{1}{n}\sum_{i\in L_j} R_{U_i}\varepsilon_i + \frac{1}{n}\sum_{i\in L_j} R_{U_i}\left(\hat{m}^{(j)}\left(U_i\right) - m_0^{(j)}\left(U_i\right)\right) + P_\lambda \hat{m}^{(j)} = 0,
$$

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

$$\mathbb{E}_U\left[R_U \Delta m^{(j)}(U)\right] + P_\lambda \Delta m^{(j)} = \frac{1}{n}\sum_{i \in L_j} R_{U_i}\varepsilon_i - P_\lambda m_0^{(j)} - \frac{1}{n}\sum_{i \in L_j}\left(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\left[R_U \Delta m^{(j)}(U)\right]\right). \tag{7.1}$$

The L.H.S. of (7.1) can be rewritten as

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

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

$$\hat{m}^{(j)} - m_0^{(j)} = \frac{1}{n}\sum_{i \in L_j} R_{U_i}\varepsilon_i - P_\lambda m_0^{(j)} - \frac{1}{n}\sum_{i \in L_j}\left(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\left[R_U \Delta m^{(j)}(U)\right]\right). \tag{7.2}$$

We will show that the first term in the R.H.S. of (7.2) weakly converges to a normal distribution, the second term contributes to the estimation bias, and that the last term is an asymptotically ignorable remainder term. We denote the last term as $Rem^{(j)} := \frac{1}{n}\sum_{i \in L_j}\left(R_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\left[R_U \Delta m^{(j)}(U)\right]\right)$. 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:

$$\begin{aligned} Rem_\beta^{(j)} := &\ \frac{1}{n}\sum_{i \in L_j}\left(L_{U_i}\Delta m^{(j)}(U_i) - \mathbb{E}_U\left[L_U \Delta m^{(j)}(U)\right]\right) \\ Rem_f^{(j)} := &\ \frac{1}{n}\sum_{i \in L_j}\left(NU_i \Delta m^{(j)}(U_i) - \mathbb{E}_U\left[N_U \Delta m^{(j)}(U)\right]\right). \end{aligned}$$

Similarly, (7.2) can be rewritten into the following two equations:

$$\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)} = \frac{1}{n}\sum_{i \in L_j} L_{U_i}\varepsilon_i - L_\lambda f_0 - Rem_\beta^{(j)}, \tag{7.3}$$

and

$$\hat{f}^{(j)} - f_0 = \frac{1}{n}\sum_{i \in L_j} N_{U_i}\varepsilon_i - N_\lambda f_0 - Rem_f^{(j)}, \tag{7.4}$$

for all j = 1,...,s. Taking average of (7.4) for all j over s, and by definition of $\bar{f}$, we have

$$\bar{f} - f_0 = \frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i - N_\lambda f_0 - \frac{1}{s}\sum_{j=1}^{s} Rem_f^{(j)}, \tag{7.5}$$

where we used $1/N\sum_{i=1}^{N} N_{U_i}\varepsilon_i = 1/s\sum_{j=1}^{s} 1/n\sum_{i \in L_j} N_{U_i}\varepsilon_i$. By (7.5), it follows that

$$\left[\| \bar{f} - f_0\|_{\mathscr{C}}^{2}\right] \le 3\left[\|\frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i\|_{\mathscr{C}}^{2}\right] + 3\|N_\lambda f_0\|_{\mathscr{C}}^{2} + 3\left[\|\frac{1}{s}\sum_{j=1}^{s} Rem_f^{(j)}\|_{\mathscr{C}}^{2}\right]. \tag{7.6}$$

By Lemma A.5 and the fact that each $N_{U_i}\varepsilon_i$ is i.i.d., it follows that

$$\left[\|\frac{1}{N}\sum_{i=1}^{N} N_{U_i}\varepsilon_i\|_{\mathscr{C}}^{2}\right] = \frac{1}{N}\left[\|N_U \varepsilon\|_{\mathscr{C}}^{2}\right] \le C_1\sigma^2\frac{1}{Nh}, \tag{7.7}$$

and

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

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 sub-populations that

$$\left[\|\frac{1}{s}\sum_{j=1}^{s} Rem_f^{(j)}\|_{\mathscr{C}}^{2}\right] = \frac{1}{s^2}\sum_{j=1}^{s}\left[\|Rem_f^{(j)}\|_{\mathscr{C}}^{2}\right]. \tag{7.9}$$

Therefore it suffices to bound $\left[\|Rem_f^{(j)}\|_{c}^{2}\right]$. We have the following lemma that controls this term:

Lemma 7.1—Suppose Assumptions 3.1, 3.2 and Condition (3.6) hold. We have for all j = 1,...,s

$$\left[\|Rem^{(j)}\|_{\mathscr{A}}^{2}\right] \leq a\left(n, s, h, \lambda, \omega\right),$$

for sufficiently large *n*. Moreover, the inequality also holds for $\left[\|Rem_{\beta}^{(j)}\|_{2}^{2}\right]$ and $\left[\|Rem_{f}^{(j)}\|_{\mathscr{C}}^{2}\right]$

Combining (7.6) - (7.9) and Lemma 7.1, and by the fact that $\|\bar{f}-f_0\|_{L_2(\mathbb{P}_Z)}^{2} \leq \|\bar{f}-f_0\|_{\mathscr{C}}^{2}$, we complete the proof of Theorem 3.1.

7.2. Proof of Theorem 3.3

Proof—Recall that $m_0^{(j)*}=\left(\boldsymbol{\beta}_0^{(j)*}, f_0^*\right)=(id-P_\lambda)\, m_0^{(j)}$ where $m_0^{(j)}=\left(\beta_0^{(j)}, f_0\right)$. This implies that $\beta_0^{(j)*}=\beta_0^{(j)}-L_\lambda f_0$ and (7.5), for arbitrary ***x*** and *z*$_0$,

$$\begin{aligned}\left(\boldsymbol{x}^T, 1\right)\begin{pmatrix}\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0^{(j)*}\right)\\ \sqrt{Nh}\left(\bar{f}\left(z_0\right)-f_0^*\left(z_0\right)\right)\end{pmatrix} &= \sqrt{n}\boldsymbol{x}^T\left(\hat{\boldsymbol{\beta}}^{(j)}-\boldsymbol{\beta}_0^{(j)*}\right)+\sqrt{Nh}\left(\bar{f}_{N,\lambda}\left(z_0\right)-f_0^*\left(z_0\right)\right)\\ &= \underbrace{\frac{1}{\sqrt{n}}\sum_{i\in L_j}\boldsymbol{x}^T L_{U_i}\varepsilon_i+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}h^{1/2}N_{U_i}\left(z_0\right)\varepsilon_i}_{(I)}+\underbrace{\sqrt{n}\boldsymbol{x}^T Rem_{\beta}^{(j)}+\sqrt{Nh}s^{-1}}_{(II)}\sum_{j=1}^{s}Rem_f^{(j)}\left(z_0\right)\end{aligned}$$

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 ***x*** is arbitrary, we apply Wold device to conclude the proof of joint asymptotic normality.

Asymptotic normality of (I): We present that result for showing asymptotic normality of (I) in the following lemma and defer its proof to supplemental material.

Lemma 7.2.—Suppose that as $N \to \infty, h\|\widetilde{K}_{z_0}\|_{L_2(\mathbb{P}_Z)}^{2} \to \sigma_{z_0}^2$, $h^{1/2}\left(W_\lambda \boldsymbol{A}\right)\left(z_0\right) \to a_{z_0} \in {}^{p}$ and $h^{1/2}\boldsymbol{A}\left(z_0\right) \to -\gamma_{z_0} \in {}^{p}$. We have

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

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

(ii) if *s* is fixed, then

$$(I) \rightsquigarrow N\left(0, \sigma^2\left(\boldsymbol{x}^T\Omega^{-1}\boldsymbol{x}+\Sigma_{22}+2s^{-1/2}\boldsymbol{x}^T\Sigma_{12}\right)\right). \quad (7.11)$$

Control of the remainder term (II): We now turn to bound the remainder term (II). We need the following lemma:

Lemma 7.3—Suppose Assumption 3.1, 3.2 and Condition (3.6) hold. We have the following two sets of results that control the remainder terms:

(i) For all *j* = 1,...,*s*

$$\|Rem^{(j)}\|_{\mathscr{A}} = o_P\left(b_{n,s}\right).$$

where $b_{n,s} = Ch^{-1}n^{-1/2}r_{n,s}\left(\omega\left(\mathscr{F},1\right) + log\ N\right)$ and $r_{n,s} = (log\ N)^2(nh)^{-1/2} + \lambda^{1/2}$. Also, $\|Rem_\beta^{(j)}\|_2 = o_P\left(b_{n,s}\right)$ and $\|Rem_f^{(j)}\|_{\mathscr{C}} = o_P\left(b_{n,s}\right)$.

(ii) Moreover, we have

$$\begin{aligned} \left\|\frac{1}{s}\sum_{j=1}^{s} Rem^{(j)}\right\|_{\mathscr{A}} &= o_P\left(sw^{-1/2}b_{n,s}\ log\ N\right) \\ \|\frac{1}{s}\sum_{j=1}^{s} Rem_\beta^{(j)}\|_2 &= o_P\left(s^{-1/2}b_{n,s}\ log\ N\right) \\ \|\frac{1}{s}\sum_{j=1}^{s} Rem_f^{(j)}\|_{\mathscr{C}} &= o_P\left(s^{-1/2}b_{n,s}\ log\ N\right). \end{aligned}$$

By Lemma 7.3, we have

$$\sqrt{n}|\boldsymbol{x}^T Rem_\beta^{(j)}| \le \sqrt{n}\|\boldsymbol{x}\|_2\|Rem_\beta^{(j)}\|_2 = o_P\left(n^{1/2}b_{n,s}\right) = o_p\left(\sqrt{N}s^{-1/2}b_{n.s}\right), \quad (7.12)$$

where we used the boundedness of ***x***. Also,

$$\begin{aligned} \sqrt{Nh}\left|s^{-1}\sum_{j=1}^{s} Rem_f^{(j)}\left(z_0\right)\right| &\le \sqrt{Nh}\|\widetilde{K}_{z_0}\|_{\mathscr{C}}\left\|s^{-1}\sum_{j=1}^{s} Rem_f^{(j)}\right\|_{\mathscr{C}} \\ &\lesssim \sqrt{N}\left\|s^{-1}\sum_{j=1}^{s} Rem_f^{(j)}\right\|_{\mathscr{C}} \\ &= o_P\left(\sqrt{N}s^{-1/2}b_{n,s}\ log\ N\right), \end{aligned} \quad (7.13)$$

where the second inequality follows from Lemma A.4. Therefore by (7.12) and (7.13), we have

$$(II) = o_P\left(\sqrt{N}s^{-1/2}b_{n,s}\ log\ N\right). \quad (7.14)$$

Now by definition of $b_{n,s}$ and condition (3.7), we have $(II) = o_P(1)$. Combining (7.10) and (7.14), it follows that if $s \to \infty$, then

$$\left(\boldsymbol{x}^T, 1\right) \begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\right) \\ \sqrt{Nh}\left(\bar{f}\left(z_0\right) - f_0^*\left(z_0\right)\right) \end{pmatrix} \rightsquigarrow N\left(0, \sigma^2\left(\boldsymbol{x}^T \Omega^{-1} \boldsymbol{x} + \Sigma_{22}\right)\right).$$

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

$$\left(\boldsymbol{x}^T, 1\right) \begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)*}\right) \\ \sqrt{Nh}\left(\bar{f}\left(z_0\right) - f_0^*\left(z_0\right)\right) \end{pmatrix} \rightsquigarrow N\left(0, \sigma^2\left(\boldsymbol{x}^T \Omega^{-1} \boldsymbol{x} + \Sigma_{22} + 2 s^{-1/2} \boldsymbol{x}^T \Sigma_{12}\right)\right).$$

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

7.3. Proof of Lemma 7.3: Controlling the Remainder Term

Proof—(i) We first derive the bound of $\|Rem^{(j)}\|_{\mathscr{A}}$. Recall

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

Let $Z_n\left(m\right) - c_r^{-1} h^{1/2} n^{-1/2} \sum_{i \in L_j}\left\{m\left(U_i\right) R_{U_i} - \left[m\left(U\right) R_U\right]\right\}$, 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 $Rem^{(j)} = c_r^{-1}\sqrt{nh} Z_n\left(\Delta m^{(j)}\right)$. We apply Lemma F.1 to obtain an exponential inequality for $sup_{m \in \mathscr{F}} \|Z_n\left(m\right)\|_{\mathscr{A}}$. The first step is to show that $Z_n(m)$ is a sub-Gaussian process by Lemma G.1. Let $g\left(U_i, m\right) = c_r^{-1}\sqrt{nh}\left(m\left(U_i\right) R_{U_i} - \left[m\left(U\right) R_U\right]\right)$. Now for any m_1 and m_2,

$$\begin{aligned} &\|g\left(U_i, m_1\right) - g\left(U_i, m_2\right)\|_{\mathscr{A}} = c_r^{-1}\sqrt{nh}\left\{\|\left(m_1\left(U_i\right) - m_2\left(U_i\right)\right) R_{U_i}\|_{\mathscr{A}}\right. \\ &\quad \left. + \|\mathbb{E}\left[\left(m_1\left(U\right) - m_2\left(U\right)\right) R_U\right]\|_{\mathscr{A}}\right\} \\ &\le 2\sqrt{n}\|m_1 - m_2\|_{sup}, \end{aligned}$$

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

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

Then by Lemma F.1, we have

$$\left(\underset{m\in\mathscr{F}}{sup}\|Z_n\left(m\right)\|_{\mathscr{A}}\ge C\omega\left(\mathscr{F},diam\left(\mathscr{F}\right)\right)+x\right)\le C\,exp\left(\frac{-x^2}{C\,diam(\mathscr{F})^2}\right), \tag{7.16}$$

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

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

$$\|\widetilde{m}\|_{sup}\le c_r h^{-1/2}(2d_{n,s})^{-1}\|\Delta m^{(j)}\|_{\mathscr{A}}\le 1/2,$$

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

$$\|\widetilde{f}\|_{\mathscr{H}}\le\lambda^{-1/2}\|\widetilde{m}\|_{\mathscr{A}}\le\lambda^{-1/2}/\left(2d_{n,s}\right)\|\Delta m^{(j)}\|_{\mathscr{A}}\le c_r^{-1}h^{1/2}\lambda^{-1/2}$$

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

$$\left(\left\{\|Z_n\left(\widetilde{m}\right)\|_{\mathscr{A}}\ge C\omega\left(\mathscr{F},1\right)+x\right\}\cap\mathscr{E}\right)\le C\,exp\left(-x^2/C\right), \tag{7.17}$$

By the definition of $\widetilde{m}$, and the relationship $Rem^{(j)}=c_r^{-1}\sqrt{nh}Z_n\left(\Delta m^{(j)}\right)$, we calculate that $Z_n\left(\widetilde{m}\right)=\left(1/2\right)h^{1/2}n^{1/2}d_{n,s}^{-1}Rem^{(j)}=\left(1/2\right)c_r^{-1}hn^{1/2}r_{n,s}^{-1}Rem^{(j)}$. Plugging the above form of $Z_n(\widetilde{m})$ into (7.17) and letting $x=\log N$ in (7.17), we have

$$\left(\left\{\|Rem^{(j)}\|_{\mathscr{A}}\ge b_{n,s}\right\}\cap\mathscr{E}\right)\le C\,exp\left(-log^2\,N/C\right), \tag{7.18}$$

where we used the definition that $b_{n,s}=Ch^{-1}n^{-1/2}r_{n,s}\left(\omega\left(\mathscr{F},1\right)+log\,N\right)$. Therefore we have

$$\begin{aligned}\left(\|Rem^{(j)}\|_{\mathscr{A}} \geq b_{n,s}\right) & \leq \left(\left\{\|Rem^{(j)}\|_{\mathscr{A}} \geq b_{n,s}\right\} \cap \mathscr{E}\right) + (\mathscr{E}^{c}) \\ & \leq C\, exp\left(-log^{2} N/C\right) + (\mathscr{E}^{c}) .\end{aligned} \qquad (7.19)$$

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

Lemma 7.4: If Assumption 3.1, 3.2 and Condition (3.6) are satisfied, then there exist a constant c such that

$$(\mathscr{E}^{c}) = \left(\|\Delta m^{(j)}\|_{\mathscr{A}} \geq r_{n,s}\right) \lesssim n\, exp\left(-c\, log^{2} N\right).$$

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

By Lemma 7.4 and (7.19) we have

$$\left(\|Rem^{(j)}\|_{\mathscr{A}} \geq b_{n,s}\right) \lesssim n\, exp\left(-c\, log^{2} N\right). \qquad (7.20)$$

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

$$\left(\{\cap_{j} \mathscr{A}_{j}\} \cap \left\{\|s^{-1}\sum_{j=1}^{s} Rem^{(j)}\|_{\mathscr{A}} > s^{-1/2} b_{n,s}\, log\, N\right\}\right) \leq 2\, exp\left(-log^{2} N/2\right). \qquad (7.21)$$

Moreover, by (7.20),

$$\left(\mathscr{A}_{j}^{c}\right) \lesssim n\, exp\left(-c\, log^{2} N\right). \qquad (7.22)$$

Hence it follows that

$$\left(\|s^{-1}\sum_{j=1}^{s} Rem^{(j)}\|_{\mathscr{A}} > s^{1/2} b_{n,s}\, log\, N\right) \leq \left(\left\{\cap_{j=1}^{s} \mathscr{A}_{j}\right\} \cap \left\{\|s^{-1}\sum_{j=1}^{s} Rem^{(j)}\|_{\mathscr{A}} > s^{-1} b_{n,s}\, log\, N\right\}\right) + \left(\cup_{j} \mathscr{A}_{j}^{c}\right)$$

as $N \to \infty$, where the last inequality follows from (7.21), (7.22) and union bound. This finishes the proof.

We can apply similar arguments as above to bound $\|Rem_f^{(j)}\|_{\mathscr{C}}$ and $\|1/s\sum_{j=1}^{s}Rem_f^{(j)}\|_{\mathscr{C}}$, by changing $\omega\left(\mathscr{F},1\right)$ to $\omega\left(\mathscr{F}_2,1\right)$, which is dominated by $\omega\left(\mathscr{F},1\right)$. The bounds of $\nabla\|Rem_\beta^{(j)}\|_2$ and $\|1/s\sum_{j=1}^{s}Rem_\beta^{(j)}\|_2$ then follow from triangular inequality.

7.4. Proof of Theorem 3.4

Proof—In view of Theorem 3.4, we first prove

$$\begin{pmatrix} \sqrt{n}\left(\boldsymbol{\beta}_0^{(j)*}-\boldsymbol{\beta}_0^{(j)}\right) \\ \sqrt{Nh}\left(f_0^*\left(z_0\right)-f_0\left(z_0\right)-W_\lambda f_0\left(z_0\right)\right) \end{pmatrix} \to \mathbf{0} \qquad (7.23)$$

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

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

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

$$\boldsymbol{A}\left(z\right)=\left\langle \boldsymbol{A},\widetilde{K}_z\right\rangle_{\mathscr{C}}=\left\langle \boldsymbol{B},\widetilde{K}_z\right\rangle_{L_2\left(\mathbb{P}_Z\right)}=\sum_{\ell=1}^{\infty}\frac{\left\langle \boldsymbol{B},\phi_\ell\right\rangle_{L_2\left(\mathbb{P}_Z\right)}}{1+\lambda/\mu_\ell}\phi_\ell\left(z\right), \qquad (7.25)$$

Applying Cauchy-Schwarz, we obtain

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

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

$$\boldsymbol{A}(z_0)^T L_\lambda f_0 \le \|\boldsymbol{A}\left(z_0\right)\|_2\|L_\lambda f_0\|_2 \lesssim \lambda.$$

Therefore, if we choose $\lambda=o\left(\left(Nh\right)^{-1/2}\wedge n^{-1/2}\right)$, 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 $\alpha_{z_0}=lim_{N\to\infty}h^{1/2}W_\lambda\boldsymbol{A}\left(z_0\right)$. By Lemma A.2 and (7.25), we have

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

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

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

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research Sponsored by NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841

Research Sponsored by NSF CAREER Award DMS-1151692, DMS-1418042, Simons Foundation 305266. 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.

References

Aitkin M, Rubin DB. Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society. Series B (Methodological). 1985:67–75.

Bach F. Sharp analysis of low-rank kernel matrix approximations. arXiv preprint arXiv. 2012:1208.2015.

Berlinet, A., Thomas-Agnan, C. Reproducing kernel Hilbert spaces in probability and statistics. Vol. 3. Springer; 2004.

Birman MŠ, Solomjak M. Piecewise-polynomial approximations of functions of the classes img align= absmiddle alt= w_p^ α tex_sm_2343_img1/img. Sbornik: Mathematics. 1967; 2:295–317.

Carl B, Triebel H. Inequalities between eigenvalues, entropy numbers, and related quantities of compact operators in banach spaces. Mathematische Annalen. 1980; 251:129–133.

Chen, X., Xie, M. Tech. rep., Technical Report 2012-01. Dept. Statistics, Rutgers Univ; 2012. A split-and-conquer approach for analysis of extraordinarily large data..

Cheng G, Shang Z. Joint asymptotics for semi-nonparametric models under penalization. arXiv. 2013:1311.2628.

Chernozhukov V, Chetverikov D, Kato K, et al. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. The Annals of Statistics. 2013; 41:2786–2819.

Fan J, Zhang W. Statistical estimation in varying coefficient models. Annals of Statistics. 1999:1491–1518.

Figueiredo MA, Jain AK. Unsupervised learning of finite mixture models. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2002; 24:381–396.

Gu, C. Smoothing spline ANOVA models. Vol. 297. Springer; 2013.

Guo W. Inference in smoothing spline analysis of variance. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002; 64:887–898.

Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological). 1993:757–796.

Huang J, Zhang T. The benefit of group sparsity. The Annals of Statistics. 2010; 38:1978–2004.

Kleiner A, Talwalkar A, Sarkar P, Jordan M. The big data bootstrap. arXiv preprint arXiv. 2012:1206.6415.

Kosorok, MR. Introduction to empirical processes and semiparametric inference. Springer; 2007.

Krasikov I. New bounds on the hermite polynomials. arXiv preprint math/0401310. 2004

Lafferty JD, Lebanon G. Diffusion kernels on statistical manifolds. 2005

Mammen E, van de Geer S. Penalized quasi-likelihood estimation in partial linear models. The Annals of Statistics. 1997:1014–1035.

McDonald R, Hall K, Mann G. 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. 2010

McLachlan, G., Peel, D. Finite mixture models. John Wiley & Sons; 2004.

Meinshausen N, Bühlmann P. Maximin effects in inhomogeneous large-scale data. arXiv preprint arXiv. 2014:1406.0596.

Mendelson, S. In Computational Learning Theory. Springer; 2002. Geometric parameters of kernel machines..

Nardi Y, Rinaldo A. On the asymptotic properties of the group lasso estimator for linear models. Electronic Journal of Statistics. 2008; 2:605–633.

Obozinski, G., Wainwright, MJ., Jordan, MI. Union support recovery in high-dimensional multivariate regression.. Communication, Control, and Computing, 2008 46th Annual Allerton Conference on; IEEE; 2008.

Pinelis I. Optimum bounds for the distributions of martingales in banach spaces. The Annals of Probability. 1994:1679–1706.

Raskutti G, Wainwright MJ, Yu B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. The Journal of Machine Learning Research. 2014; 15:335–366.

Shang Z, Cheng G. Local and global asymptotic inference in smoothing spline models. The Annals of Statistics. 2013; 41:2608–2638.

Shawe-Taylor, J., Cristianini, N. Kernel methods for pattern analysis. Cambridge university press; 2004.

Sollich, P., Williams, CK. Deterministic and Statistical Methods in Machine Learning. Springer; 2005. Understanding gaussian process regression using the equivalent kernel.; p. 211-228.

Städler N, Bühlmann P, van de Geer S. #2113$_1$-penalization for mixture regression models. Test. 2010; 19:209–256.

Steinwart, I., Hush, DR., Scovel, C., et al. Optimal rates for regularized least squares regression. COLT; 2009.

Stewart GW, Sun J.-g. Matrix perturbation theory. 1990

Stone CJ. Additive regression and other nonparametric models. The annals of Statistics. 1985:689–705.

Tropp JA. User-friendly tail bounds for sums of random matrices. Foundations of Computational Mathematics. 2012; 12:389–434.

Tsybakov, AB., Zaiats, V. Introduction to nonparametric estimation. Vol. 11. Springer; 2009.

Van Der Vaart, AW., Wellner, JA. Weak Convergence. Springer; 1996.

van Handel R. Probability in high dimension: Lecture notes. 2014

Wahba, G. Spline models for observational data. Vol. 59. Siam; 1990.

Wang X, Dunson DB. Parallel mcmc via weierstrass sampler. arXiv preprint arXiv. 2013:1312.4605.

Wang, Y. Smoothing splines: methods and applications. CRC Press; 2011.

Wasserman L. Steins method and the bootstrap in low and high dimensions: A tutorial. 2014

Williamson RC, Smola AJ, Scholkopf B. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. Information Theory, IEEE Transactions on. 2001; 47:2516–2532.

Yang Y, Barron A. Information-theoretic determination of minimax rates of convergence. Annals of Statistics. 1999:1564–1599.

Zhang T. Learning bounds for kernel regression using effective data dimensionality. Neural Computation. 2005; 17:2077–2098. [PubMed: 15992491]

Zhang Y, Duchi J, Wainwright M. Divide and conquer kernel ridge regression. Conference on Learning Theory. 2013

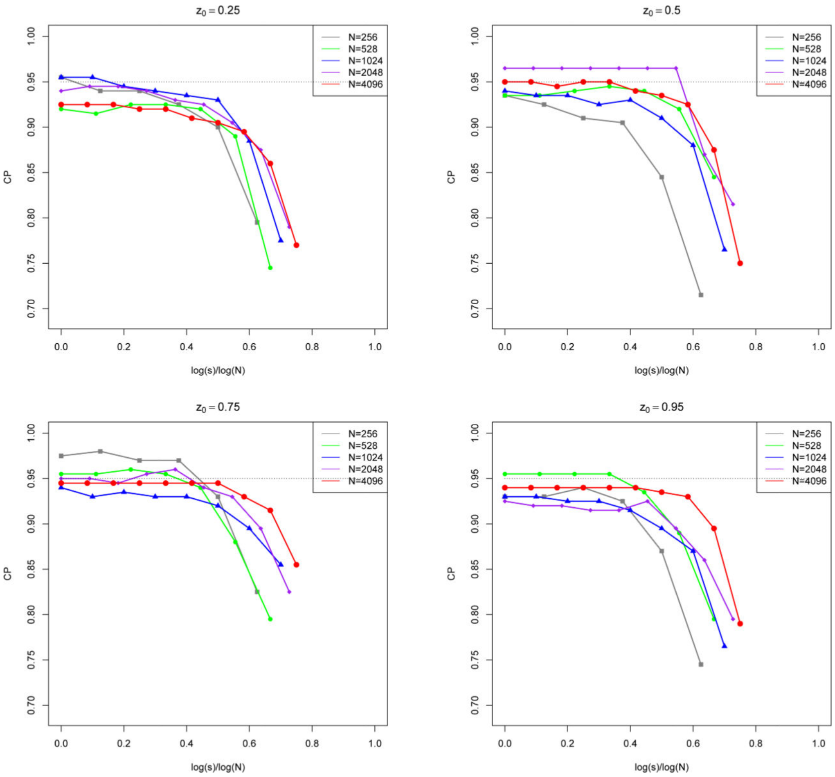

Fig 1.

Coverage probability of 95% predictive interval with different choices of s and N

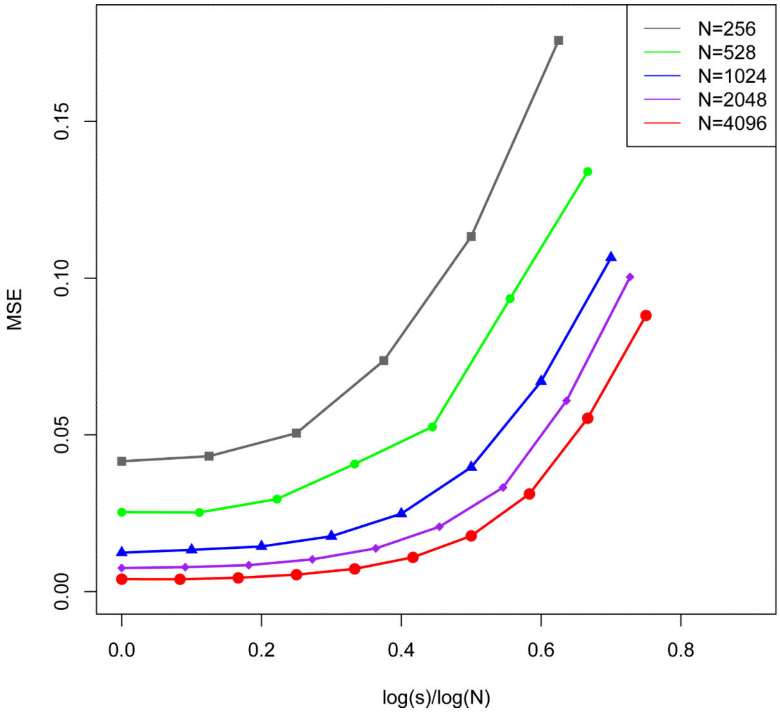

Fig 2.

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

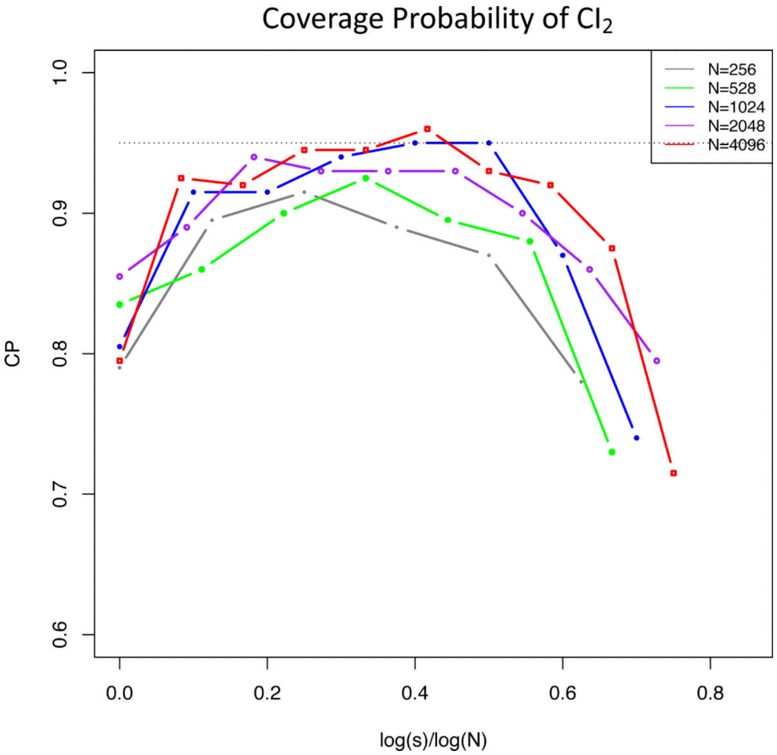

Fig 3.

Coverage probability of 95% *confidence interval based on* $\check{\beta}$

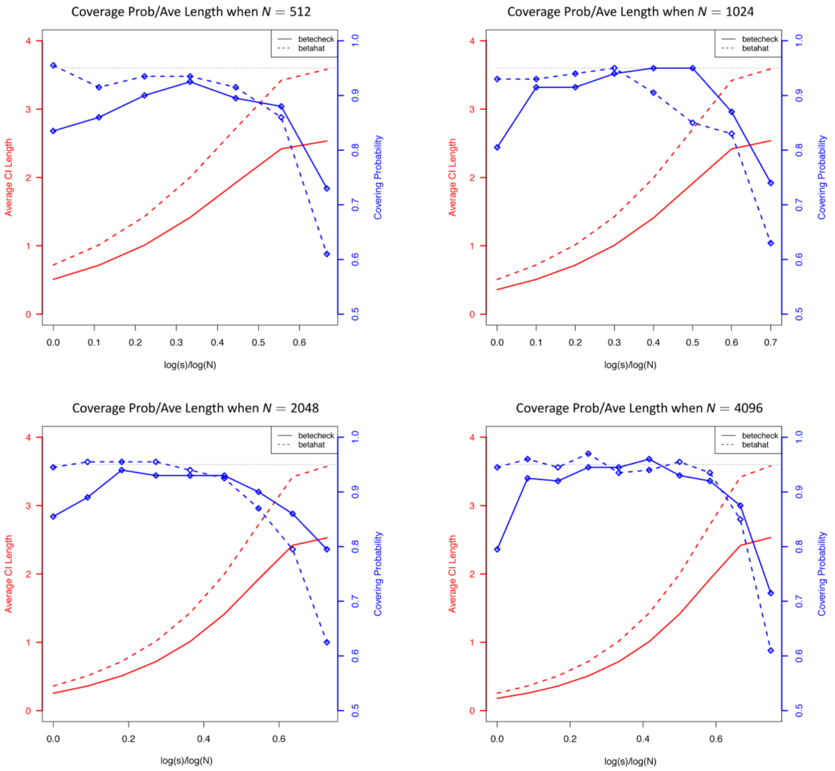

Fig 4.

Covergae probabilities and average lengths of 95% *confidence intervals constructed based on* $\hat{\beta}$ *and* $\check{\beta}$. *In the above figures, dashed lines represent* CI_1, *which is constructed based on* $\check{\beta}$, *and solid lines represent* CI_2, *which is constructed based on* $\hat{\beta}$.

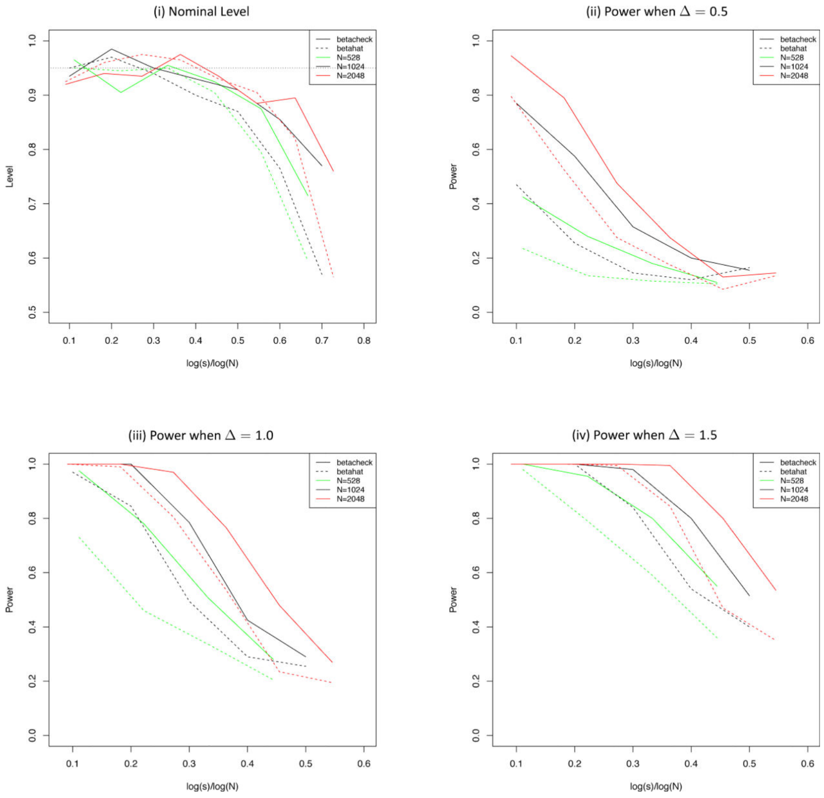

Fig 5.

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