The Annals of Statistics
2019, Vol. 47, No. 5, 2922–2949
https://doi.org/10.1214/18-AOS1769

PROJECTED SPLINE ESTIMATION OF THE NONPARAMETRIC FUNCTION IN HIGH-DIMENSIONAL PARTIALLY LINEAR MODELS FOR MASSIVE DATA

BY HENG LIAN[1], KAIFENG ZHAO AND SHAOGAO LV[2]

City University of Hong Kong, Philips Research China and Nanjing Audit University

In this paper, we consider the local asymptotics of the nonparametric function in a partially linear model, within the framework of the divide-and-conquer estimation. Unlike the fixed-dimensional setting in which the parametric part does not affect the nonparametric part, the high-dimensional setting makes the issue more complicated. In particular, when a sparsity-inducing penalty such as lasso is used to make the estimation of the linear part feasible, the bias introduced will propagate to the nonparametric part. We propose a novel approach for estimation of the nonparametric function and establish the local asymptotics of the estimator. The result is useful for massive data with possibly different linear coefficients in each subpopulation but common nonparametric function. Some numerical illustrations are also presented.

1. Introduction. In this paper, we consider divide-and-conquer methodology for high-dimensional partially linear models, focusing on the estimation and asymptotic distribution of the nonparametric function. In recent years, there has been an increasing research interest on dealing with data sets so large that they need to be partitioned into smaller subsets to be analyzed one by one, or even to be distributed to multiple local machines to be analyzed, due to the whole data set is too large to be loaded into a single machine. The final estimator is typically obtained by averaging the local estimates. For example, Chen and Xie (2014) considered averaging the local estimates for generalized linear models after dividing the data, Zhang, Duchi and Wainwright (2015) studied the nonparametric problem of averaging the local kernel ridge estimates and Kleiner et al. (2014) proposed averaging the bootstrap estimates after subsampling the original data, calling their method “bag of little bootstraps.” The common messages of these works, roughly speaking, is that the divide-and-conquer approach yields a pooled estimator of

Received April 2018; revised July 2018.

[1] Supported in part by Hong Kong RGC general research fund 11301718 and by Project 11871411 from NSFC and the Shenzhen Research Institute, City University of Hong Kong.

[2] Supported in part by NSFC 11871277.

MSC2010 subject classifications. Primary 62G08; secondary 62G20.

Key words and phrases. Asymptotic normality, B-splines, local asymptotics, profiled estimation.

the same statistical performance as the estimator using the entire sample. Shi, Lu and Song (2017) and Banerjee, Durot and Sen (2017) established that the pooled estimator even outperforms the estimator based on the entire sample in some nonstandard problems where the convergence rate is slower than $\sqrt{n}$.

For semiparametric models, Zhao, Cheng and Liu (2016) considered a pooled estimator of the nonparametric function in a partially linear model and found that as long as the number of partitions does not grow too fast and the smoothing parameter is chosen according to the entire sample size N, the asymptotic properties of the estimator can be the same as the estimator based on the entire data. Furthermore, it was shown under appropriate assumptions the existence of the fixed-dimensional linear part in a partially linear model does not affect the estimation of the nonparametric function, and the aggregated estimator possesses the same asymptotic distribution as the "oracle estimate" computed when all data, as well as the linear coefficients, are available. A natural open question is that whether and how the asymptotic distribution of the nonparametric part is affected by the unknown linear part in the high-dimensional setting with possibly $p > N$. Even if the entire data is analyzed without a divide-and-conquer strategy, the answer to this question seems unclear. It turns out aggregating the standard spline estimator of the nonparametric function when a lasso penalty is used for the linear part does not work in the high-dimensional setting and the high-dimensional linear part needs to be dealt with carefully.

The high-dimensional partially linear model can be motivated by the genetic study where a response variable is related to a large set of genes, while there is an additional environmental variable whose relationship with the response is required to be nonlinear possibly based on biological knowledge. Although the sample size used in such studies is typically small due to cost consideration, after combining different studies together the data can potentially be massive.

In this paper, we consider partially linear model estimators with a high-dimensional linear part, with a lasso penalty for the linear part. We note even though partially linear models with a high-dimensional linear part has been investigated, the asymptotic distribution of the nonparametric function has not been considered in the literature. We use polynomial splines to approximate the nonparametric functions (Huang, Zhang and Zhou (2007), Liang and Li (2009), Xie and Huang (2009)) which is computationally more convenient than using kernel-based estimation (Fan and Yao (2003), Härdle and Liang (2007)). For spline estimators with a lasso penalty, both the nonparametric part and the parametric part can be estimated simultaneously in a single step. However, the resulting shrinkage bias of the linear part will propagate to the nonparametric part. In particular, this makes it hard to establish the asymptotic distribution of the nonparametric function, unless the dimension of the parametric part is sufficiently small, or at least the number of nonzero components in the linear part is sufficiently small, so that its convergence rate is dominated by the convergence rate of the nonparametric part. In this paper, we propose a novel estimation method for the nonparametric function, which we call a projected spline estimator, since it is related to the projection/profile technique often used in semiparametric models. The projection alleviates bias and makes it possible to establish its asymptotic normality under reasonable conditions.

Our results directly apply to the setting where the data are naturally partitioned into subpopulations and the true linear coefficients in different subpopulations are allowed to be different (see model (2.2)). This is an extension to the high-dimensional setting of the heterogeneous data context considered in Zhao, Cheng and Liu (2016). The latter focused on fixed-dimensional problems without a lasso penalty in which the linear part has a $\sqrt{n}$ convergence rate and is asymptotically unbiased, and thus does not affect the convergence rate of the nonparametric part. On the other hand, we argue that the bias in the high-dimensional linear part can have a nontrivial effect on the estimator of the nonparametric part.

The main contribution of this work is to establish the asymptotic distribution of the aggregated estimator for the nonparametric function in a massive data setting where the entire dataset is partitioned into subpopulations. The partition is necessary due to the data set is too big to be analyzed simultaneously. The nonparametric functions are assumed to be the same across subpopulations, while the linear part can be the same or different in different subpopulations. To deal with the bias from the penalized estimation of the linear part, our proposed projection/profile approach is similar to the one often used in the literature for parametric and semiparametric models (Li (2000), Wang et al. (2011)). However, as far as we know, there is no work that explicitly used the projection for the estimation of the nonparametric function, due to that it is not useful in the fixed-dimensional setting. A further technical challenge here is that we are trying to profile out an ultrahigh dimensional linear part.

In the special case that the data are homogeneous (different partitions have the same linear coefficients), aggregation of the linear part is also of interest, which has been considered in Lv and Lian (2017) in a reproducing kernel Hilbert space framework, and thus we do not consider distributed estimation of the linear part in this paper. Since Lv and Lian (2017) focused on the linear part, the problem of bias propagation to the nonparametric part was not studied there. Compared to the profiling procedure for the parametric part, the profile technique for the nonparametric part here involves novel ideas and techniques.

The rest of the article is organized as follows. In the next section, we present the model setup and explain the possible large bias when both the nonparametric and the parametric part are estimated simultaneously. We then propose a profiled method to address the bias problem, which makes it possible to aggregate the estimators in a massive data setting. In Section 3, we present the asymptotic properties of the estimator, followed by some simulation studies in Section 4. We conclude with a discussion in Section 5.

Notation. For any vector $a = (a_j)$, $\|a\|_0 = \sum_j I\{a_j \neq 0\}$, $\|a\|_1 = \sum_j |a_j|$, $\|a\| = (\sum_j a_j^2)^{1/2}$, $\|a\|_\infty = \max_j |a_j|$ denotes its ℓ_0, ℓ_1, ℓ_2 and ℓ_∞ norm, respectively. For a matrix $A = (a_{ij})$, we define $\|A\| = (\sum_{i,j} a_{ij}^2)^{1/2}$, $\|A\|_{\mathrm{op}}$ its operator norm and $\|A\|_{\max} = \max_{i,j} |a_{ij}|$. Throughout the paper, C denotes a generic positive constant whose value can be different at different places.

2. High-dimensional partially linear model.

We consider the following partially linear model (PLM):

$$y_i = f(w_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \dots N, \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^{\mathrm{T}}$ are the predictors in the linear part with p diverging with, or even larger than, the sample size N and $(y_i, w_i, \mathbf{x}_i)$ are i.i.d. copies of $(y, w, \mathbf{x})$. We assume the sample size N is so large that it is impossible to analyze the entire data set at once. So we divide observations into m parts/subpopulations $\mathcal{S}_j$, $j = 1, \dots, m$ with $\bigcup_j \mathcal{S}_j = \{1, \dots, N\}$, and each is analyzed separately before being aggregated.

Slightly more generally, we can allow the subpopulations to be heterogeneous in the sense that the linear coefficients could be different. That is, within subpopulation j,

$$y_i = f(w_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}^{(j)} + \epsilon_i, \qquad i \in \mathcal{S}_j. \tag{2.2}$$

There is no extra complication in dealing with (2.2) compared to (2.1) using our methodology proposed below, and thus we will stick with the more general setting (2.2). For simplicity of notation, we always assume all subpopulations have the same size $n = N/m$ which is an integer.

For the nonparametric function, we consider polynomial splines for estimation. Assume the support of the distribution of w is $[0, 1]$ for simplicity. Let $\tau_0 = 0 < \tau_1 < \cdots < \tau_{K'} < 1 = \tau_{K'+1}$ be a partition of $[0, 1]$ into subintervals $[\tau_k, \tau_{k+1}), k = 0, \dots, K'$ with K' internal knots. We only restrict our attention to equally spaced knots. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q - 1$ and globally $q - 2$ times continuously differentiable on $[0, 1]$. The collection of splines with a fixed sequence of knots has a B-spline basis $B(t) = (B_1(t), \dots, B_K(t))^{\mathrm{T}}$ with $K = K' + q$. We assume the B-spline basis is normalized to have $\sum_{k=1}^K B_k(t) = \sqrt{K}$. Such normalization is not essential and is just imposed to simplify some expressions in theoretical derivations later.

We assume $f \in C^\alpha([0, 1])$ with $\alpha \geq 1$, where

$$C^\alpha([0, 1]) = \{f : [0, 1] \to R, f^{(t)}(x) \text{ is Lipschitz continuous of order } r\},$$

where t is the largest integer strictly less than α and $r = \alpha - t$. Then there exists a K-vector $\boldsymbol{\theta}_0$ such that

$$\sup_t \left| f(t) - B^{\mathrm{T}}(t)\boldsymbol{\theta}_0 \right| \leq CK^{-\alpha}, \tag{2.3}$$

which is possible by splines' approximation property (de Boor (2001), Huang (2003), Schumaker (2007)). We will use a particular coefficient vector defined by

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} E\big[\big(f(w) - B^{\mathrm{T}}(w)\boldsymbol{\theta}\big)^2\big]. \tag{2.4}$$

By Theorem A.1 of Huang (2003), $\boldsymbol{\theta}_{0j}$ defined in this way satisfies (2.3).

We first study the estimators in one subpopulation, and thus suppress the superscript (j) which we will use to denote the quantities based on subpopulation j toward the end of the current section. For the linear part, when p is large, one typically assumes the true parameter $\boldsymbol{\beta}_0$ is sparse in order to feasibly obtain some good estimate of $\boldsymbol{\beta}_0$. We follow this tradition by assuming $\|\boldsymbol{\beta}_0\|_0 = s$ and use a lasso penalty on $\boldsymbol{\beta}$ (Tibshirani (1996)). We can estimate $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$ by

$$(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) = \arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.5}$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$, $\mathbf{Z}_{n\times K} = (B(w_1), \dots, B(w_n))^{\mathrm{T}}$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^{\mathrm{T}}$.

In the estimation of partially linear models, often the nonparametric part is profiled out first to obtain an estimator of $\boldsymbol{\beta}_0$. More specifically, under our current framework, it is easy to see that for any given $\boldsymbol{\beta}$, the minimizer of $\boldsymbol{\theta}$ in the displayed above is given by

$$\boldsymbol{\theta} = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{2.6}$$

or equivalently,

$$\boldsymbol{\theta} - \boldsymbol{\theta}_0 = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\boldsymbol{\epsilon} + \mathbf{R} - \mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)), \tag{2.7}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^{\mathrm{T}}$, $\mathbf{R} = (r_1, \dots, r_n)$ and $r_i = f(w_i) - B^{\mathrm{T}}(w_i)\boldsymbol{\theta}_0$. By plugging (2.6) into (2.5), we have

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\big\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\big\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.8}$$

where for any matrix $\mathbf{A}$, $\mathbf{P}_{\mathbf{A}} := \mathbf{A}(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}$ is a projection matrix. The classical profiled estimator in the fixed-dimensional setting without using penalty is simply

$$\big((\mathbf{X} - \mathbf{P}_{\mathbf{Z}}\mathbf{X})^{\mathrm{T}}(\mathbf{X} - \mathbf{P}_{\mathbf{Z}}\mathbf{X})\big)^{-1}(\mathbf{X} - \mathbf{P}_{\mathbf{Z}}\mathbf{X})^{\mathrm{T}}\mathbf{Y}, \tag{2.9}$$

the minimizer of the first term in (2.8). After obtaining $\widehat{\boldsymbol{\beta}}$, we can plug it back into (2.6) to get

$$\boldsymbol{\theta}^* = (\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

which we call the *plug-in estimator*.

Except for the projection $\mathbf{P}_{\mathbf{Z}}$, (2.8) is basically the standard lasso problem, and thus one expects that similar strategy such as that used in Bickel, Ritov and Tsybakov (2009); Belloni and Chernozhukov (2013) will yield convergence rate

$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = O_p(s\sqrt{\log(p \vee N)/n})$ (see Theorem 3.1). After plugging $\widehat{\boldsymbol{\beta}}$ into (2.6) to obtain $\widehat{\boldsymbol{\theta}}$, it is seen that the bias inherent in $\widehat{\boldsymbol{\beta}}$ will propagate to $\widehat{\boldsymbol{\theta}}$, and thus we expect that the bias in $\widehat{\boldsymbol{\beta}}$ makes it hard to obtain the asymptotic distribution, unless $s\sqrt{\log(p \vee N)/n}$ is small compared to the typical convergence rate of the nonparametric part $\sqrt{K/N}$. Even if s is small, $s\sqrt{\log(p \vee N)/n}$ can still be much larger than $\sqrt{K/N}$ if n is small compared to N, which is a phenomenon nonexistent in the classical setting where the entire data set is analyzed without partitioning. Note that in the big data setting where there are multiple partitions, the bias will survive aggregation and remains being the same order as the bias before aggregation.

To motivate our new estimator to be proposed, by examining (2.9) we note that the rows of $\mathbf{X} - \mathbf{P_Z X}$ is an approximation to $\mathbf{x}_i - E[\mathbf{x}|w = w_i]$ which is uncorrelated with $f(w_i)$. Thus one initial idea is to define the estimator for the nonparametric part as

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} &= \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Y}\\ &= \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\boldsymbol{\epsilon} + \mathbf{R} + \mathbf{Z}\boldsymbol{\theta}_0 + \mathbf{X}\boldsymbol{\beta}_0),\end{aligned}$$

where $\widehat{\boldsymbol{\gamma}} \in \mathbb{R}^{p\times K}$ is an estimator of $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K) := (E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}] = (E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}E[\mathbf{x}B^{\mathrm{T}}(w)]$. Note that by the definition of $\boldsymbol{\gamma}$, $B(w_i) - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_i$ is uncorrelated with $\mathbf{x}_i$. We call $\widetilde{\boldsymbol{\theta}}$ the *naive profiled estimator*.

However, although $(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}$ has mean zero due to the projection, the nonstochastic terms are still too large for our purpose. In fact, by easy algebra, we have

$$\begin{aligned}\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 &= \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Y} - \boldsymbol{\theta}_0\\ &= \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\bigl(\boldsymbol{\epsilon} + \mathbf{R} + \mathbf{X}(\widehat{\boldsymbol{\gamma}}\boldsymbol{\theta}_0 + \boldsymbol{\beta}_0)\bigr), \qquad (2.10)\end{aligned}$$

and $(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}\boldsymbol{\theta}_0 + \boldsymbol{\beta}_0)$ turns out to be too large for our purpose. We thus make two key modifications. The first and more obvious change is to replace $\mathbf{Y}$ by $\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ so that $\mathbf{X}\boldsymbol{\beta}_0$ in the expression above will become $\mathbf{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}})$. The second and less obvious change is to replace $((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}}))^{-1}$ by $((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z})^{-1}$ whose purpose is to remove $\mathbf{X}\widehat{\boldsymbol{\gamma}}\boldsymbol{\theta}_0$. In fact, with these changes, now we define

$$\widehat{\boldsymbol{\theta}} = \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z}\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

and by substituting $\mathbf{Y} = \boldsymbol{\epsilon} + \mathbf{R} + \mathbf{Z}\boldsymbol{\theta}_0 + \mathbf{X}\boldsymbol{\beta}_0$ and subtracting $\boldsymbol{\theta}_0$ from both sides, we easily get (note $\boldsymbol{\theta}_0$ has disappeared in the right-hand side below)

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = \bigl((\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z}\bigr)^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})\bigl(\boldsymbol{\epsilon} + \mathbf{R} - \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\bigr),$$

and we will establish the asymptotic normality of this estimator. Note also that compared with (2.7), $\mathbf{Z}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is replaced with $(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ with the latter hopefully being of a smaller order due to that $(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}$ has mean zero. $\widehat{\boldsymbol{\theta}}$ is our *proposed (profiled) estimator*. Although in form the difference between $\widetilde{\boldsymbol{\theta}}$

and $\widehat{\boldsymbol{\theta}}$ is minor, without careful analysis it is hard to see beforehand why $\widehat{\boldsymbol{\theta}}$ works while $\widetilde{\boldsymbol{\theta}}$ does not.

Since $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ is a large $p \times p$ matrix, we need additional assumptions to estimate $\boldsymbol{\gamma}$. Among possibly multiple approaches, for example, by assuming sparsity of $(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}$, we adopt the more direct approach of assuming the sparsity of $\boldsymbol{\gamma}$. More specifically, we assume $\|\boldsymbol{\gamma}_k\|_0 \le s_k$ and for simplicity of notation assume $s_1 = \cdots = s_K = s$. Note $\boldsymbol{\gamma}_k$ can be interpreted as the coefficients when regressing $B_k(w)$ on $\mathbf{x}$, and thus sparsity of $\boldsymbol{\gamma}_k$ can be naturally assumed as for the standard lasso approach in a regression framework. Furthermore, we can define the estimator of $\boldsymbol{\gamma}_k$ by

$$\widehat{\boldsymbol{\gamma}}_k = \arg\min \frac{1}{2}\|\mathbf{Z}_k - \mathbf{X}\boldsymbol{\gamma}_k\|^2 + \lambda_k\|\boldsymbol{\gamma}_k\|_1, \tag{2.11}$$

where $\mathbf{Z}_k$ is the kth column of $\mathbf{Z}$ and $\lambda_k > 0$ is a tuning parameter. Again for simplicity, since $\boldsymbol{\gamma}_k$ has the same sparsity as $\boldsymbol{\beta}_0$, we can use $\lambda_1 = \cdots = \lambda_K = \lambda$. In practice, we will use different λ_k for $k = 1, \dots, K$.

Now consider the partitioned setting (2.2). Here, we can first estimate $\boldsymbol{\beta}_0^{(j)}$ by

$$\widehat{\boldsymbol{\beta}}^{(j)} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}^{(j)}})(\mathbf{Y}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\beta})\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.12}$$

where $\mathbf{Y}^{(j)}$, $\mathbf{Z}^{(j)}_{n\times K}$ and $\mathbf{X}^{(j)}_{n\times p}$ are defined as before, using only observations in $\mathcal{S}_j$. Then we compute

$$\widehat{\boldsymbol{\theta}}^{(j)} = ((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)})^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}(\mathbf{Y}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\beta}}^{(j)}),$$

where $\widehat{\boldsymbol{\gamma}}^{(j)} = (\widehat{\boldsymbol{\gamma}}_1^{(j)}, \dots, \widehat{\boldsymbol{\gamma}}_K^{(j)})$ is obtained from (2.11) using only observations in $\mathcal{S}_j$.

Once we obtained $\widehat{\boldsymbol{\theta}}^{(j)}$, $j = 1, \dots, m$, these m estimates can be pooled to finally yield

$$\check{\boldsymbol{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\widehat{\boldsymbol{\theta}}^{(j)}.$$

In particular, $f(x)$ is estimated by $\check{f}(x) = B^{\mathrm{T}}(x)\check{\boldsymbol{\theta}}$.

3. Asymptotic properties.

The following assumptions are adopted to show the asymptotic properties under model (2.2), and treat (2.1) as a special case.

(A1) The observations in different subpopulations are independent and are independent and identically distributed within each subpopulation satisfying model (2.2).

(A2) w has a density whose support is $[0, 1]$ and is bounded away from zero and infinity. The error has variance σ^2 and is independent of predictors with a sub-Gaussian distribution.

(A3) $f \in C^{\alpha}([0,1])$ with $\alpha \geq 1$. Here,

$$C^{\alpha}\big([0,1]\big) = \big\{f : [0,1] \to R,\, f^{(t)}(x) \text{ is Lipschitz continuous of order } r\big\},$$

where t is the largest integer strictly less than α and $r = \alpha - t$.

(A4) $\mathbf{x} = (x_1, \dots, x_p)^{\mathrm{T}}$ has sub-Gaussian components and eigenvalues of $E[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ are bounded away from zero and infinity. Let $\mathbf{h}_{ik} = \mathbf{x}_i(B_k(w_i) - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_k)$ and we assume $\mathbf{h}_{ik}$ has subexponential components (note $\mathbf{h}_{ik}$ has mean zero by the definition of $\boldsymbol{\gamma}$) with parameters $(C_1, C_2\sqrt{K})$ for some constants $C_1, C_2 > 0$.

(A5) $E[x_j|w]$ as a function of w is in $C^{\alpha'}([0,1])$ with some $\alpha' \geq 1$. We use B-splines with order $q \geq \max\{\alpha, \alpha'\}$. Eigenvalues of $E[(\mathbf{x} - E[\mathbf{x}|w])(\mathbf{x} - E[\mathbf{x}|w])^{\mathrm{T}}]$ are bounded away from zero and infinity.

(A6) The true parameter $\boldsymbol{\beta}_0^{(j)}$ is sparse with $\|\boldsymbol{\beta}_0^{(j)}\|_0 \leq s$. We also assume $\|\boldsymbol{\gamma}_k\|_0 \leq s$. Furthermore, $\|\boldsymbol{\gamma}_k\|_1$ is bounded.

(A7) (restricted eigenvalue condition) For some constants $c > 1$ and $\kappa > 0$,

$$\inf_{\|\delta_{T^c}\|_1 \leq c\|\delta_T\|_1} \frac{\|(\mathbf{X} - E[\mathbf{X}|\mathbf{w}])\delta\|}{\sqrt{N}\|\delta\|} \geq \kappa,$$

where $T = \{j : |\beta_{0j}| \neq 0\}$ and δ_T is the subvector of δ containing only components in T. Here, $E[\mathbf{X}|\mathbf{w}]$ is the $N \times p$ matrix with entries $E[x_j|w = w_i]$.

Some of the assumptions are standard while others call for more explanations. Compact support for w is standard in order to construct splines, but we do not assume x_j is bounded. Sub-Gaussianity assumption for $\mathbf{x}$ and ϵ is convenient for applying some concentration inequality. On the other hand, for $x_j(B_k(w) - \mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}_k)$, it is more reasonable to use a subexponential assumption. Chapter 2 of Wainwright (2018) contains a comprehensive discussion of supexponential random variables. Remember that a mean zero subexponential random variable x with parameters (a, b) by definition satisfies that $E[e^{tx}] \leq e^{a^2t^2/2}$ for $|t| < 1/b$. If we assume that components of $\mathbf{x}$ are sub-Gaussian, that the conditional density $p(w|x_j)$ is uniformly bounded, then we have $E[B_k^r(w)|x_j] \leq C\sqrt{K}^{r-2}$, and thus $E[x_j^r B_k^r(w)] = E[x_j^r E[B_k^r(w)|x_j]] \leq Cr!(C\sqrt{K})^{r-2}$ by Theorem 2.1(III) of Wainwright (2018). Then by Proposition 2.3 of Wainwright (2018), $x_j B_k(w)$ is subexponential with parameters $(C_1, C_2\sqrt{K})$. Furthermore, suppose we assume $\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}_k$ is sub-Gaussian. This would be true, for example, if $\mathbf{x}$ is sub-Gaussian (in the sense that $\mathbf{x}^{\mathrm{T}}\mathbf{a}$ is sub-Gaussian for any unit norm vector $\mathbf{a}$) and $\|\boldsymbol{\gamma}_k\|_2$ is bounded, or if components of $\mathbf{x}$ are bounded and $\|\boldsymbol{\gamma}_k\|_1$ is bounded (in this case $\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}_k$ is easily seen to be bounded, and thus sub-Gaussian). Then $x_j\mathbf{x}^{\mathrm{T}}\boldsymbol{\gamma}_k$ is subexponential (based on the fact that product of two sub-Gaussian random variables is subexponential). These then imply the second statement of (A4). In (A6), we assume both $\boldsymbol{\beta}_0^{(j)}$ and $\boldsymbol{\gamma}_k$ are sparse. Although this is common and a familiar assumption for $\boldsymbol{\beta}_0^{(j)}$, sparsity of $\boldsymbol{\gamma}_k$ is less common. But it can be motivated from the similar principle that motivated the sparsity of $\boldsymbol{\beta}_0^{(j)}$. Without sparsity assumption, it is hard, if not impossible, to feasibly estimate $\boldsymbol{\gamma}_k$. This also roughly means only a few of the components of $\mathbf{x}$ is related to $B_k(w)$. On the other hand, the strict sparsity can be relaxed. For example, one can assume the existence of an s-sparse $\widetilde{\boldsymbol{\gamma}}_k$ with $\|\boldsymbol{\gamma}_k - \widetilde{\boldsymbol{\gamma}}_k\|_1 \le a_n$ where a_n represents sparse approximation error. Then it is easy to show that $\sum_i (\mathbf{x}_i^{\mathrm{T}}(\boldsymbol{\gamma}_k - \widetilde{\boldsymbol{\gamma}}_k))^2/n \le Ca_n^2$ with high probability. Following similar arguments in Theorem 1 of Belloni and Chernozhukov (2013), we have $\|\widehat{\boldsymbol{\gamma}}_k - \widetilde{\boldsymbol{\gamma}}_k\|_1 \le s\sqrt{\log(p \vee N)/n} + \sqrt{s}a_n^2$, and thus $\|\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\|_1 \le s\sqrt{\log(p \vee N)/n} + \sqrt{s}a_n^2 + a_n$. A larger bound for $\|\widehat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\|_1$ leads to larger bound in the statement of Lemma 5, but does not invalidate our main results with corresponding modification of assumptions in the statement of Theorem 3.2 below.

Similarly, the assumption that $\|\boldsymbol{\gamma}_k\|_1$ is bounded can be relaxed to allow it to be (slowly) diverging. This makes the bound in Lemma 5 as well as the bound in (3.14) and (3.15) slightly larger but it is easy to see that the assumptions in the statement of Theorem 3.2 below can be modified accordingly so that the result still holds. Finally, (A5) is standard in semiparametric models, for example, in Xie and Huang (2009) and we provide a more detailed discussion of the restricted eigenvalue condition (A7) in the Appendix.

THEOREM 3.1. *Under Assumptions* (A1)–(A7) *and that* $(s^2 + Ks)\log(p \vee N) = o(n)$, $s = o(K^{2\alpha'})$ *and that* $\lambda = C\sqrt{n\log(p \vee N)}$ *for some* C *large enough, with probability at least* $1 - (p \vee N)^{-C}$, *for all* m *machines*,

$$\|\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0^{(j)}\|_1 \le s\sqrt{\log(p \vee N)/n}, \tag{3.1}$$

$$\|\widehat{\boldsymbol{\gamma}}_k^{(j)} - \boldsymbol{\gamma}_k\|_1 \le s\sqrt{\log(p \vee N)/n}. \tag{3.2}$$

Note the appearance of $\log N$ *above instead of* $\log n$ *is due to that we require rates uniform over all* m *machines*.

PROOF. In the proof, we suppress the superscripts (j) and the following arguments apply to each subpopulation separately.

Based on the profiled penalized least squares problem (2.8), we have

$$\frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\|^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 \le \frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0)\|^2 + \lambda\|\boldsymbol{\beta}_0\|_1.$$

Using $\mathbf{Y} = \boldsymbol{\epsilon} + \mathbf{R} + \mathbf{Z}\boldsymbol{\theta}_0 + \mathbf{X}\boldsymbol{\beta}_0$ with $\mathbf{R} = (r_1, \dots, r_n)$ and $r_i = f(w_i) - B^{\mathrm{T}}(w_i)\boldsymbol{\theta}_0$, simple algebra on the displayed above yields

$$\begin{aligned}&\frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 - (\boldsymbol{\epsilon} + \mathbf{R})^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \lambda\|\widehat{\boldsymbol{\beta}}_T\|_1 + \lambda\|\widehat{\boldsymbol{\beta}}_{T^c}\|_1\\ &\quad \le \lambda\|\boldsymbol{\beta}_{0T}\|_1,\end{aligned}$$

where $T = \{j : \beta_{0j} \neq 0\}$ is the support of the true $\boldsymbol{\beta}_0$ while $T^c = \{1, \dots, p\}\backslash T$, and $\boldsymbol{\beta}_{0T}$ denotes the subvector of $\boldsymbol{\beta}_0$ containing only components indexed by T, for example.

Denote $\Delta := \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$, using $\|\widehat{\boldsymbol{\beta}}_{T^c}\|_1 = \|\Delta_{T^c}\|_1$ and $\|\boldsymbol{\beta}_{0T}\|_1 - \|\widehat{\boldsymbol{\beta}}_T\|_1 \leq \|\Delta_T\|_1$, the display above becomes

$$\frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta\|^2 - (\boldsymbol{\epsilon} + \mathbf{R})^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta \leq \lambda\|\Delta_T\|_1 - \lambda\|\Delta_{T^c}\|_1.$$

By Lemmas 1 and 2, $|(\boldsymbol{\epsilon} + \mathbf{R})^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta| \leq \|(\boldsymbol{\epsilon} + \mathbf{R})^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty\|\Delta\|_1 \leq (\lambda/c)\|\Delta\|_1 = (\lambda/c)(\|\Delta_T\|_1 + \|\Delta_{T^c}\|_1)$ for some $c > 1$, which leads to

$$\frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|^2 + \lambda\left(1 - \frac{1}{c}\right)\|\Delta_{T^c}\|_1 \leq \lambda\left(1 + \frac{1}{c}\right)\|\Delta_T\|_1. \tag{3.3}$$

In particular, the above implies

$$\|\Delta_{T^c}\|_1 \leq \frac{c+1}{c-1}\|\Delta_T\|_1. \tag{3.4}$$

Then Lemma 4 shows that

$$\frac{1}{2}\|(\mathbf{I} - \mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta\|^2 \geq Cn\|\Delta\|^2. \tag{3.5}$$

Combining (3.3) and (3.5), we get

$$n\|\Delta\|_2^2 \leq C\lambda\|\Delta_T\|_1 \leq C\lambda\sqrt{s}\|\Delta\|_2,$$

or,

$$\|\Delta\| \leq C\frac{\lambda\sqrt{s}}{n}.$$

Furthermore, by (3.4), we have

$$\|\Delta\|_1 = \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1 \leq C\|\Delta_T\|_1 \leq C\sqrt{s}\|\Delta\|_2 \leq C\frac{\lambda s}{n},$$

which proved (3.1).

Proof of (3.2) is similar. We have by similar arguments that, with now $\Delta = \widehat{\boldsymbol{\gamma}}_k^{(j)} - \boldsymbol{\gamma}_k$ and $T = \{j : |\boldsymbol{\gamma}_{kj}| \neq 0\}$,

$$\frac{1}{2}\|\mathbf{X}\Delta\|^2 - \mathbf{h}_k^{\mathrm{T}}\Delta \leq \lambda\|\Delta_T\|_1 - \lambda\|\Delta_{T^c}\|_1, \tag{3.6}$$

where $\mathbf{h}_k = \sum_i \mathbf{h}_{ik}$. Using Bernstein's inequality instead of Hoeffding's inequality used in Lemma 2, we can have with probability at least $1 - (p \vee N)^{-C}$, $\|\mathbf{h}_k\|_\infty \leq \lambda/c$ and the same arguments for $\widehat{\boldsymbol{\beta}}$ lead to the conclusion (3.2). □

The following main result shows the asymptotic property of the estimator.

THEOREM 3.2. *Under assumptions assumed in Theorem 3.1, and that* $s\sqrt{K}\log(p \vee N)/n = o(1/\sqrt{N})$, $s\sqrt{\log(p \vee N)/n}K^{-\alpha+1/2} = o(1/\sqrt{N})$, $K^2\log(p \vee N)/n = o(1/\sqrt{N})$, $K^{-\alpha+1}\log(p \vee N)/\sqrt{n} = o(1/\sqrt{N})$ *and* $N = o(K^{2\alpha+2\alpha'-1})$, *we have*

$$B^{\mathrm{T}}(x)(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = B^{\mathrm{T}}(x)\big(E[(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})]\big)^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{\epsilon} + o_p(\sqrt{K/N}).$$

In particular, this implies

$$\frac{\check{f}(x) - f(x) + O(K^{-\alpha})}{(\sigma^2 B^{\mathrm{T}}(x)(NE[(B(w) - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w) - \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}])^{-1}B(x))^{1/2}} \xrightarrow{d} N(0, 1). \tag{3.7}$$

REMARK 1. The assumptions on the relative order of s, K, p, N can be easily satisfied. For example, assume $\alpha = \alpha'$ and set $K \asymp N^{1/(2\alpha+1)}$ which is the optimal number of knots based on the entire sample, the constraints are satisfied when $s^2K^{\alpha+5/2}\log^2(p \vee N) = o(n)$. Note that the assumptions in the statement of Theorem 3.2 actually impose a constraint on m. For example, with $K \asymp N^{1/(2\alpha+1)}$, $s^2K^{\alpha+5/2}\log^2(p \vee N) = o(n)$ can be equivalently written as $m = o(\sqrt{N}/(s^2K^2\log^2(p \vee N)))$. On the other hand, there is not theoretical low bound for m and m can be fixed or even $m = 1$. But our interest is in the setting that the data is too big to be analyzed as a whole, so one wants to use large m to alleviate computational burden.

REMARK 2. The key mechanism by which the novel projected estimator reduces the propagation of bias from the linear part is seen in (3.9) in the proof. In the standard estimator based on (2.7), instead of (3.9), we would have to bound the term $\|((\mathbf{Z}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)})^{-1}(\mathbf{Z}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)\|$. Even when p is fixed, $\|(\mathbf{Z}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}\|$ would have order $O(n)$, and thus $\|((\mathbf{Z}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)})^{-1}(\mathbf{Z}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)} \times (\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)\| = O_p(\|\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0\|)$ would fail to be smaller than $1/\sqrt{N}$.

REMARK 3. For the naive profiled estimator, the term (2.10) causing trouble that cannot be easily dealt with is

$$((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)}))^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)}\boldsymbol{\theta}_0. \tag{3.8}$$

In fact, it can be show that, similar to the proof of Lemma 5, $\|((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)}))^{-1}\|_{\mathrm{op}} = O_p(1/n)$. Furthermore, $\|(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)} \times \widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}\|_{\max} \le \lambda = C\sqrt{n\log(p \vee N)}$ by the first-order condition for estimating $\boldsymbol{\gamma}$. $\|\widehat{\boldsymbol{\gamma}}^{(j)}\boldsymbol{\theta}_0\|_1$ can be bounded by $\|\widehat{\boldsymbol{\gamma}}^{(j)}\|_1\|\boldsymbol{\theta}_0\|_\infty = O(\sqrt{K})$. These bounds however lead to a bound for (3.8) that is larger than $n^{-1/2}$ which is too large for our purpose. Thus although we do not have a rigorous proof, it seems difficult to establish asymptotic normality of $\widetilde{\boldsymbol{\theta}}$. On the other hand, when p is fixed, we can define $\widehat{\boldsymbol{\gamma}}^{(j)}$ as the least squares solution that minimizes the first term of (2.11) without penalization. This makes $(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}$ exactly equal to zero, and thus this problem does not appear in the fixed-dimensional case. It is also easy to see, by following our proof of Theorem 3.2, that for the fixed-dimensional case the asymptotic distribution of the naïve profiled estimator is the same as (3.7).

REMARK 4. Under mild assumptions, it can be shown that $E[(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}]$ has eigenvalues bounded and bounded away from zero, then the denominator of (3.7) is of order $\sqrt{K/N}$ since $\|B(x)\|^2 \asymp K$, which is the standard convergence rate for the spline-based nonparametric function estimation. In fact, to see the order of the eigenvalues, we first assume that $E[\mathbf{v}\mathbf{v}^{\mathrm{T}}]$ has eigenvalues bounded and bounded away from zero, where we define $\mathbf{v}=(B(w),\mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$. Note that this is a reasonable assumption even in the high-dimensional case. Once this is assumed, using the identity

$$(\mathbf{I},-\boldsymbol{\gamma}^{\mathrm{T}})\mathbf{v}\mathbf{v}^{\mathrm{T}}(\mathbf{I},-\boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}=(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}},$$

we see that $E[(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}]$ has eigenvalues bounded and bounded away from zero as long as the operator norm of $\boldsymbol{\gamma}$ (its largest singular value) is bounded. Given $\boldsymbol{\gamma}=(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}E[\mathbf{x}B^{\mathrm{T}}(w)]$, with eigenvalues of $(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}$ and operator norm of $E[\mathbf{x}B^{\mathrm{T}}(w)]$ bounded (this is easily seen by $(\mathbf{a}^{\mathrm{T}}E[\mathbf{x}B^{\mathrm{T}}(w)]\mathbf{b})^2\le(\mathbf{a}^{\mathrm{T}}E[\mathbf{x}\mathbf{x}^{\mathrm{T}}]\mathbf{a})(\mathbf{b}^{\mathrm{T}}E[B(w)B^{\mathrm{T}}(w)]\mathbf{b})$), we have indeed $\|\boldsymbol{\gamma}\|_{\mathrm{op}}$ is bounded.

REMARK 5. When the dimension p is fixed, we can actually show that $E[(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}]=E[B(w)B^{\mathrm{T}}(w)](1+o(1))$ so the asymptotic distribution would reduce to the more familiar form for nonparametric regression, and the effect of the linear part would disappear. This is because, using $\boldsymbol{\gamma}=(E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}E[\mathbf{x}B^{\mathrm{T}}(w)]$, we have

$$\begin{aligned}&E[(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})(B(w)-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}]\\&\quad=E[B(w)B^{\mathrm{T}}(w)]-E[B(w)\mathbf{x}^{\mathrm{T}}](E[\mathbf{x}\mathbf{x}^{\mathrm{T}}])^{-1}E[\mathbf{x}B^{\mathrm{T}}(w)]\end{aligned}$$

and $E[B(w)x_j]=O(1/\sqrt{K})$, so the first term above of order $O(1)$ dominates.

REMARK 6. When $q=\alpha$, as shown in Zhou, Shen and Wolfe (1998) and also mentioned in Section 5.3 of Huang (2003) (note that our definition of $\boldsymbol{\theta}_0$ has nothing to do with the linear part, and thus the bias is the same as for nonparametric models), the bias can be written more explicitly as $-f^{(q)}(x)h_k^q/q!\mathcal{B}_q((x-\tau_k)/h_k)+o(K^{-d})$ when $x\in(\tau_k,\tau_{k+1}]$ where $\mathcal{B}_q(\cdot)$ is the qth Bernoulli polynomial and $h_k=\tau_{k+1}-\tau_k$. If $q\ge\alpha+1$, then as Huang (2003) showed, the bias become smaller with order $o(K^{-\alpha})$.

PROOF OF THEOREM 3.2. Using superscript (j) to denote the estimates based on the jth subpopulation, we consider the aggregated estimator that satisfies

$$\check{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (1/m)\sum_{j=1}^{m}\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1} \times (\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\big(\boldsymbol{\epsilon}^{(j)} + \mathbf{R}^{(j)} - \mathbf{X}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)\big).$$

We have

$$\begin{aligned}
&\left\|(1/m)\sum_{j=1}^{m}\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)\right\| \\
&\quad\le \max_j \big\|\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}(\widehat{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0)\big\| \\
&\quad\le Cs\sqrt{K}\log(p\vee N)/n = o(1/\sqrt{N}),
\end{aligned} \tag{3.9}$$

where we used that $\|((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)})^{-1}\|_{\mathrm{op}}$ is of order $O_p(1/n)$ (Lemma 5) and $\|(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{X}^{(j)}\|_\infty \le \lambda$ by the KKT condition from the definition of $\widehat{\boldsymbol{\gamma}}^{(j)}$, as well as Theorem 3.1. Thus we focus on

$$(1/m)\sum_{j=1}^{m}\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}(\boldsymbol{\epsilon}^{(j)} + \mathbf{R}^{(j)}). \tag{3.10}$$

If the second appearance of $\widehat{\boldsymbol{\gamma}}^{(j)}$ above is replaced by the true $\boldsymbol{\gamma}$, we get

$$(1/m)\sum_{j=1}^{m}\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}(\boldsymbol{\epsilon}^{(j)} + \mathbf{R}^{(j)}). \tag{3.11}$$

The difference between (3.10) and (3.11) is

$$\begin{aligned}
&\left\|(1/m)\sum_{j=1}^{m}\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{X}^{(j)})^{\mathrm{T}}(\boldsymbol{\epsilon}^{(j)} + \mathbf{R}^{(j)})\right\| \\
&\quad\le \max_j \big\|\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}(\widehat{\boldsymbol{\gamma}}^{(j)} - \boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{X}^{(j)})^{\mathrm{T}}(\boldsymbol{\epsilon}^{(j)} + \mathbf{R}^{(j)})\big\| \\
&\quad\le \max_j \big\|\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}\big\|_{\mathrm{op}} \\
&\qquad\times \sqrt{K}\max_k \|\widehat{\boldsymbol{\gamma}}_k^{(j)} - \boldsymbol{\gamma}_k\|_1\big(\|(\mathbf{X}^{(j)})^{\mathrm{T}}\boldsymbol{\epsilon}^{(j)}\|_\infty + \|(\mathbf{X}^{(j)})^{\mathrm{T}}\mathbf{R}^{(j)}\|_\infty\big) \\
&\quad\le \frac{s\sqrt{K}\log(p\vee N) + \sqrt{n\log(p\vee N)}sK^{-\alpha+1/2}}{n} = o(1/\sqrt{N}),
\end{aligned}$$

using Theorem 3.1 and Lemma 5. Let

$$\mathbf{A}^{(j)} = \Big[\big((\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1} - \big(E[\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z} - \mathbf{X}\boldsymbol{\gamma})]\big)^{-1}\Big](\mathbf{Z}^{(j)} - \mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}.$$

We have by Lemma 5, and the identity $\mathbf{A}^{-1}-\mathbf{B}^{-1}=\mathbf{A}^{-1}(\mathbf{B}-\mathbf{A})\mathbf{B}^{-1}$,

$$\max_j\|\mathbf{A}^{(j)}\|\le C(K/n^2)\big(s^2\log(p\vee N)+\sqrt{n\big(K+\log(p\vee N)\big)\log K}\big)\sqrt{nK}.$$

Thus

$$\begin{aligned}
&(1/m)\sum_{j=1}^m\big(\big((\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\widehat{\boldsymbol{\gamma}}^{(j)})^{\mathrm{T}}\mathbf{Z}^{(j)}\big)^{-1}\\
&\quad-\big(E\big[(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})\big]\big)^{-1}\big)\\
&\quad\times(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}(\boldsymbol{\epsilon}^{(j)}+\mathbf{R}^{(j)})\\
&\quad=(1/m)\sum_{j=1}^m\mathbf{A}^{(j)}(\boldsymbol{\epsilon}^{(j)}+\mathbf{R}^{(j)})\\
&\quad=O_p\bigg((1/m)\sqrt{\sum_j\|\mathbf{A}^{(j)}\|^2}+(1/m)\sqrt{n}K^{-\alpha}\sum_j\|\mathbf{A}^{(j)}\|\bigg)\\
&\quad=O_p\big((1/\sqrt{m}+\sqrt{n}K^{-\alpha})(K/n^2)\big(s^2\log(p\vee N)\\
&\qquad+\sqrt{n\big(K+\log(p\vee N)\big)}\log K\big)\sqrt{nK}\big)\\
&\quad=o_p(1/\sqrt{N}).
\end{aligned}$$

Thus

$$\begin{aligned}
\check{\boldsymbol{\theta}}-\boldsymbol{\theta}_0&=\frac{1}{m}\sum_{j=1}^m\big(E\big[(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})\big]\big)^{-1}(\mathbf{Z}^{(j)}-\mathbf{X}^{(j)}\boldsymbol{\gamma})^{\mathrm{T}}\\
&\quad\times(\boldsymbol{\epsilon}^{(j)}+\mathbf{R}^{(j)})+o_p(1/\sqrt{N})\\
&=\big(E\big[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})\big]\big)^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\boldsymbol{\epsilon}+\mathbf{R})+o_p(1/\sqrt{N})
\end{aligned}$$

and

$$\begin{aligned}
B^{\mathrm{T}}(x)(\check{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)&=B^{\mathrm{T}}(x)\big(E\big[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})\big]\big)^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\boldsymbol{\epsilon}+\mathbf{R})\\
&\quad+o_p(\sqrt{K/N}).
\end{aligned}$$

Although asymptotic normality can be shown from the above, we have the bias term, based on a simple bound, $B^{\mathrm{T}}(x)(E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})])^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{R}=O_p(K^{-\alpha+1})$, while the stochastic term $B^{\mathrm{T}}(x)(E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})])^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\boldsymbol{\epsilon}=O_p(\sqrt{K/N})$. Thus undersmoothing with $N/K^{d-1/2}\to 0$ needs to be used. To avoid this, we carry out a more careful analysis of the bias term.

Using Lemma 6,

$$
\begin{aligned}
&B^{\mathrm{T}}(x)\big(E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})]\big)^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})\mathbf{R}\\
&\quad= B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{R}\\
&\qquad - B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\\
&\qquad - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big)^{-1}\\
&\qquad \times \big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\\
&\quad= B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{R}\\
&\qquad - B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}\\
&\qquad \times \big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\\
&\qquad - B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\\
&\qquad - E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\\
&\qquad \times E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}\big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R},
\end{aligned}
\tag{3.12}
$$

where the last equality used the Woodbury matrix identity. By Lemma 7, and that $E[B(w_i)r_i] = 0$ (due to the definition of $\boldsymbol{\theta}_0$ in (2.4)), the first term in (3.12) is

$$B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{R} = O_p\big(\sqrt{K/N}K^{-\alpha}\big) = o_p\big(\sqrt{K/N}\big).$$

Now consider the second term in (3.12). By Lemma 7, $\|B^{\mathrm{T}}(x)(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\boldsymbol{\gamma}^{\mathrm{T}}\|_1 \le \sqrt{K}/N$. Furthermore,

$$
\begin{aligned}
&\big\|\big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\big\|_\infty\\
&\quad\le \big\|\big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}|\mathbf{w}]\big)\mathbf{R}\big\|_\infty\\
&\qquad + \big\|\big(E[\mathbf{X}^{\mathrm{T}}|\mathbf{w}] - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\big\|_\infty\\
&\quad\le \sqrt{N\log(p\vee N)}K^{-\alpha} + NK^{-\alpha-\alpha'},
\end{aligned}
\tag{3.13}
$$

using that $(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}|\mathbf{w}])\mathbf{R}$ has mean zero. Thus the second term in (3.12) is

$$
\begin{aligned}
&\big|B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}\big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\big|\\
&\quad\le \big\|B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\boldsymbol{\gamma}^{\mathrm{T}}\big\|_1\big\|\big(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}\mathbf{Z}^{\mathrm{T}}\big)\mathbf{R}\big\|_\infty\\
&\quad\le \big(\sqrt{K}/N\big)\big(\sqrt{N\log(p\vee N)}K^{-\alpha} + NK^{-\alpha-\alpha'}\big) = o_p\big(\sqrt{K/N}\big).
\end{aligned}
\tag{3.14}
$$

Now consider the third term in (3.12). By Lemma 7,

$$\big\|B^{\mathrm{T}}(x)\big(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]\big)^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\big(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]\big)^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]\big\| = O\big(\sqrt{K}\big),$$

since

$$\|E[\mathbf{Z}_k^{\mathrm{T}}\mathbf{X}]\boldsymbol{\gamma}\| \le \|E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\|_\infty \sqrt{K} \max_j \|\boldsymbol{\gamma}_j\|_1 \le C. \tag{3.15}$$

Furthermore, by (3.13),

$$\begin{aligned} &\|E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{R}\| \\ &\quad \le \sqrt{K} \max_k \|\boldsymbol{\gamma}_k\|_1 \|(\mathbf{X}^{\mathrm{T}} - E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}})\mathbf{R}\|_\infty \\ &\quad = O_p(\sqrt{K}(\sqrt{N\log(p \vee N)}K^{-\alpha} + NK^{-\alpha-\alpha'})). \end{aligned}$$

These imply the third term in (3.12) is again $o_p(\sqrt{K/N})$.

Finally, the asymptotic normality straightforwardly follows from the central limit theorem. □

4. Simulations. We illustrate the performances of the distributed estimators for high-dimensional partially linear models via simulations. We generate data from model (2.2) with $f(x) = 4\sin(2\pi x)/(2 - \sin(2\pi x))$. For the covariates, we first generate a $(p+1)$-vector $\mathbf{x}_i^*$ from a multivariate normal distribution with mean zero, marginal variances 1 and all pairwise correlations equal to 0.5. Then the first component of $\mathbf{x}_i^*$ is used as w_i, after applying the cumulative distribution function of the standard normal distribution to map it to [0, 1], and the remaining components of $\mathbf{x}_i^*$ become the components of $\mathbf{x}_i$. We set

$$\boldsymbol{\beta} = (\pm 1, \pm 2, \pm 1, \pm 0.5, \pm 2, 0, 0, \ldots, 0)^{\mathrm{T}}, \tag{4.1}$$

or

$$\boldsymbol{\beta} = (\underbrace{\pm 0.5, \pm 1, \pm 0.5, \pm 0.25, \pm 1, \pm 0.5, \pm 1, \ldots, \pm 1}_{s=20}, 0, \ldots, 0)^{\mathrm{T}}, \tag{4.2}$$

where the sign in the coefficients are generated randomly for each partition. While this can create heterogeneity of data such that $\boldsymbol{\beta}$ assumes different values in different partitions, it is clear that the performance of the estimators does not depend on the sign of the components since the distribution of $\mathbf{x}_i$ is symmetric. Also note that the two $\boldsymbol{\beta}$ vectors have the same Euclidean norm. The errors are generated from $N(0, 3^2)$.

The tuning parameter λ in the penalties are selected by 5-fold cross-validation for each subpopulation. We use cubic splines in our simulation (spline order equals 4) and we set the number of internal knots to be simply $N^{1/9}$, which is the theoretical optimal order. Since our sample size is generally large, we find the results are insensitive to the choice of the number of knots as long as it is not too small. For the estimation of f, we compute the centralized estimator (CE), the naive estimator which solves (2.5) and directly aggregates the $\boldsymbol{\theta}$ estimates (NE) and the proposed estimator motivated by projection (PE). We compare the estimators by mean squared errors given by $\int(\hat{f}(x) - f(x))^2 \, dx$.

First, we set $N = 5000$, $m = 1, 5, 10, 20, 25, 50$ ($m = 1$ is the centralized estimator) and $p = 5000$. The first row of Figure 1 shows errors of the estimators that change with m, based on 200 data sets generated. We see the performances generally deteriorate with the increase of m. For the estimation of the function f, we see that the proposed estimator is better than both plug-in estimator and the naive estimator. Furthermore, for the first setting of β value (4.1), we separately compute the bias (we take the absolute value of it for visualization) and standard error and show them in Figure 2. We see that bias is generally larger than standard error and both increases with m.

In the second set of simulations, we still use $p = 5000$ and consider larger sample sizes $N = 5000, 7000, 9000, 11{,}000, 13{,}000, 15{,}000$, and fix the number of samples in each subpopulation to be $n = 1000$ (and thus the number of machines m increases with N from 5 to 15). From the reported results in the second row of Figure 1, it is seen that the proposed estimator has errors decreasing with total sample size, with errors close to the central estimator while the errors of the plug-in estimator and the naive estimator are larger.

In the third setting, we fix $N = 5000$ and consider larger dimensions $p = 5000, 7000, 9000, 11{,}000, 13{,}000$ and the MSEs are shown in the third row of Figure 1. The errors are increasing with the increase of dimension as expected and again it shows the proposed estimator works well.

Finally, we investigate the coverage probability of the pointwise confidence interval for the proposed estimator with $N = 5000$, $p = 5000$ and $m = 1$, 5, 10, 20, 25, 50 (same as the setup used in the first simulation setting above). This is based on the derived asymptotic normality for the proposed estimator, and thus we cannot construct the confidence interval for the plug-in estimator and the naive estimator. The coverage probabilities are found for the nonparametric function at the points $w = 0.1, 0.2, \ldots, 0.9$. We observe from Figure 3 that the coverage is satisfactory when m is small while it fails when m becomes large, for which the coverage probability is particularly low close to the boundary.

The simulations are carried out on the computational cluster Katana in the University of New South Wales. For the first set of simulations, for example, computation time decreases from 7 hours for the central estimator to about half an hour with $m = 10$, to finish all 200 repetitions. For sample size $N = 15{,}000$ of the second set of simulations, the central estimator takes about 20 hours while the distributed estimator takes 2 hours to finish.

5. Conclusion. In this paper, we considered the estimation of the nonparametric function in high-dimensional partially linear models. The newly proposed estimator is designed to block partially the propagation of estimation error from the high-dimensional linear part to the nonparametric part. This is important in a big and/or heterogeneous data setting in which the linear part is estimated at a

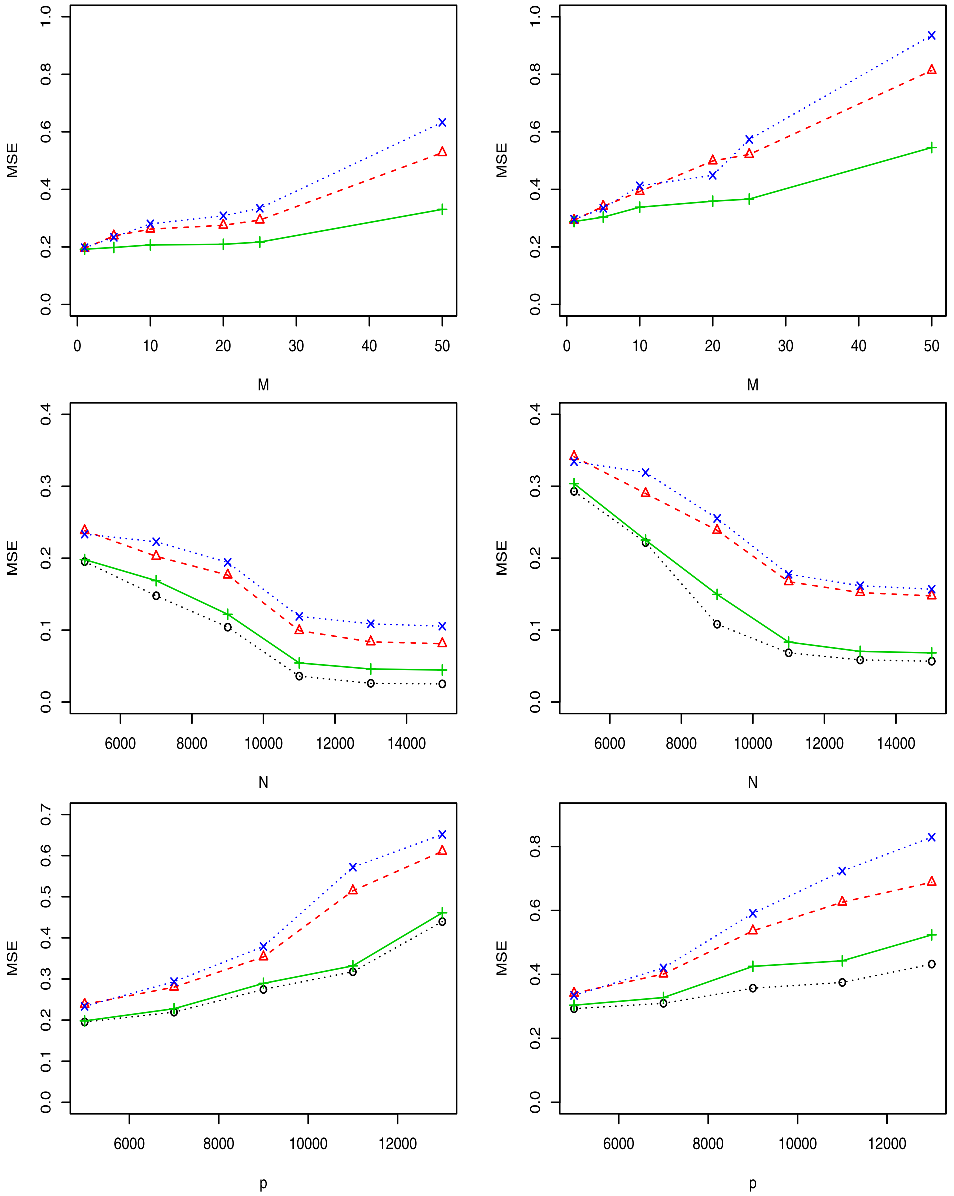

FIG. 1. *First row*: *the MSE of estimates with* $m \in \{1, 5, 10, 15, 20, 25, 50\}$ ($m = 1$ *represents the centralized estimator*) *with* $N = p = 5000$. *Second row*: *the MSE of estimates with* $p = 5000$ *and* $N \in \{5000, 7000, 9000, 11000, 13000, 15000\}$. *Third row*: *the MSE of estimates with* $N = 5000$ *and* $p \in \{5000, 7000, 9000, 11000, 13000\}$. $\circ$(*black*): *centralized estimator* (*CE*); $\triangle$(*red*): *plug-in estimator* $\boldsymbol{\theta}^*$ (*PIE*); $\times$(*blue*): *naive profile estimator* $\widetilde{\boldsymbol{\theta}}$; $+$(*green*): *the proposed profile estimator* $\widehat{\boldsymbol{\theta}}$ (*PPE*). *The left column is for* $\boldsymbol{\beta}_0$ *as in* (4.1) *while the right column is for* $\boldsymbol{\beta}_0$ *as in* (4.2).

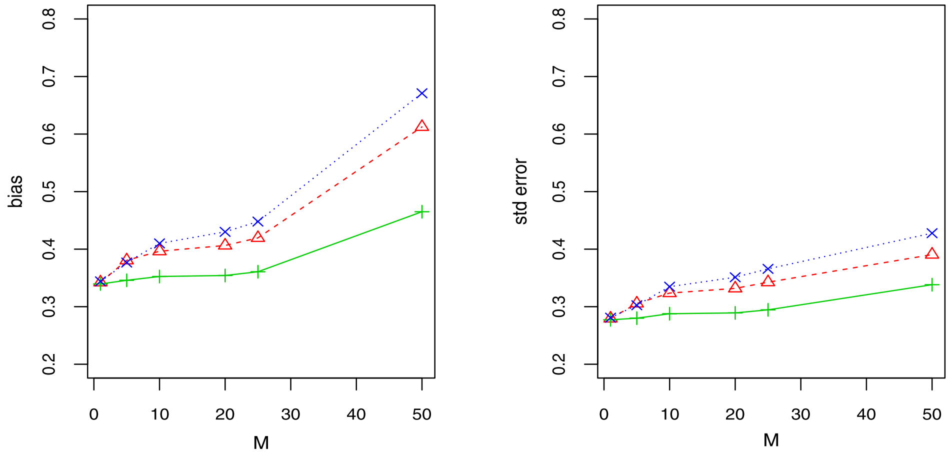

FIG. 2. *The absolute bias (left panel) and standard error (right panel) of estimates with $m \in \{1, 5, 10, 15, 20, 25, 50\}$ for the first set of simulations for $\boldsymbol{\beta}_0$ value in (4.1).*

rate much slower than the estimator based on the entire sample, as is the estimate based on a small subset of the data. We demonstrated the asymptotic normality of the nonparametric part despite that the estimator for the linear part can only be estimated at a slower rate.

Extension of the proposal to more complicated settings can be entertained. These may include additive partially linear models with multiple nonparametric functions, generalized partially linear models and models with a nonsmooth loss

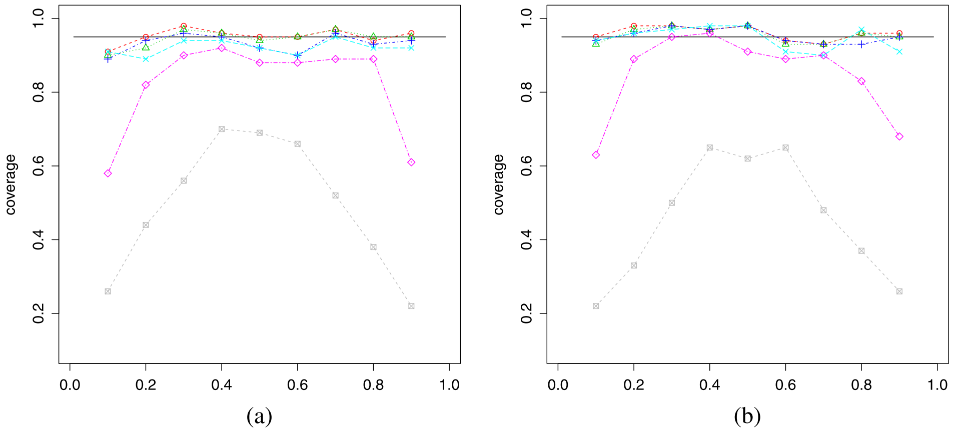

FIG. 3. *Coverage probabilities of CI for $f(w)$ at $w = 0.1, 0.2, \ldots, 0.9$ for the two choices of $\boldsymbol{\beta}_0$. The colors red, green, blue, cyan, magenta and gray represent the result for $m = 1, 5, 10, 20, 25, 50$, respectively. The left panel is for $\boldsymbol{\beta}_0$ as in (4.1) while the right panel is for $\boldsymbol{\beta}_0$ as in (4.2).*

function such as in quantile regression. These are interesting models for future investigation.

APPENDIX: LEMMAS

In the following, the superscripts (j) are suppressed and the results in Lemmas 1–5 apply to each subpopulation.

LEMMA 1. *$\|\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty \le \lambda/c$ for some $c>1$, with probability at least $1-(p\vee N)^{-C}$.*

PROOF. We write

$$\begin{aligned}
&\|\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty \\
&\quad\le \big\|\mathbf{R}^{\mathrm{T}}\big(\mathbf{X}-E[\mathbf{X}|\mathbf{w}]\big)\big\|_\infty + \big\|\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})E[\mathbf{X}|\mathbf{w}]\big\|_\infty \\
&\qquad + \big\|\mathbf{R}^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}\big(\mathbf{X}-E[\mathbf{X}|\mathbf{w}]\big)\big\|_\infty \\
&\quad=: I_1+I_2+I_3,
\end{aligned}$$

where $E[\mathbf{X}|\mathbf{w}]=(E[\mathbf{x}|w=w_1],\dots,E[\mathbf{x}|w=w_n])^{\mathrm{T}}$ as defined previously. By our definition of $\boldsymbol{\theta}_0$, $|r_i|\le CK^{-\alpha}$. Since $r_i(x_{ij}-E[x_j|w=w_i])$ has mean zero, Bernstein's inequality yields

$$I_1 \le C\sqrt{n}K^{-\alpha}\sqrt{\log(p\vee N)} \le C\sqrt{n\log(p\vee N)},$$

with probability at least $1-(p\vee N)^{-C}$. Similarly, all bounds below hold with probability at least $1-(p\vee N)^{-C}$ even without explicitly stating so.

Next, we write

$$\begin{aligned}
I_2 &= \big\|\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})E[\mathbf{X}|\mathbf{w}]\big\|_\infty \\
&= \big\|\mathbf{R}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\big(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0\big)\big\|_\infty \\
&= \big\|\mathbf{R}^{\mathrm{T}}\big(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0\big)\big\|_\infty \\
&\quad + \|\mathbf{R}^{\mathrm{T}}\mathbf{Z}\|\,\big\|(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\big\|_{\mathrm{op}}\max_j\big\|\mathbf{Z}^{\mathrm{T}}\big(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_{0j}\big)\big\|,
\end{aligned}$$

where $\boldsymbol{\alpha}_0=(\boldsymbol{\alpha}_{01},\dots,\boldsymbol{\alpha}_{0p})$ and $\boldsymbol{\alpha}_{0j}$ is such that

$$\big\|E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_{0j}\big\| \le C\sqrt{n}K^{-\alpha'}, \tag{A.1}$$

which is possible by assumption (A5).

This implies

$$\begin{aligned}
&\big\|\mathbf{R}^{\mathrm{T}}\big(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0\big)\big\|_\infty \\
&\qquad \le CnK^{-\alpha-\alpha'} = o(\lambda),
\end{aligned} \tag{A.2}$$

and

$$\|\mathbf{Z}^{\mathrm{T}}(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_{0j})\| = O_p(nK^{-\alpha'}). \tag{A.3}$$

Also, by the definition of $\boldsymbol{\theta}_0$, $R_i B_{1k}(w_{1i})$ has mean zero, and thus by Bernstein's inequality,

$$\|\mathbf{R}^{\mathrm{T}}\mathbf{Z}\| \le C\sqrt{nK\log(p \vee N)}K^{-\alpha}. \tag{A.4}$$

Combining (A.2), (A.3) and (A.4), we have

$$I_2 = o_p(\lambda).$$

Finally,

$$I_3 \le \|\mathbf{R}^{\mathrm{T}}\mathbf{Z}\| \|(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\|_{\mathrm{op}} \max_j \|\mathbf{Z}^{\mathrm{T}}(\mathbf{X}_j - E[\mathbf{X}_j|\mathbf{w}])\|.$$

Similar to the bound for I_1, since $\mathbf{Z}^{\mathrm{T}}(\mathbf{X}_j - E[\mathbf{X}_j|\mathbf{w}])$ has mean zero,

$$\max_j \|\mathbf{Z}^{\mathrm{T}}(\mathbf{X}_j - E[\mathbf{X}_j|\mathbf{w}])\|_\infty \le C\sqrt{nK\log(p \vee N)}. \tag{A.5}$$

Thus

$$I_3 = o_p(\lambda). \qquad \square$$

LEMMA 2. *With probability at least* $1-(p \vee N)^{-C}$, $\|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty \le C(\sqrt{n\log(p \vee N)})$.

PROOF. By Hoeffding's inequality, we get

$$P(\|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty > a|\{\mathbf{x}_i, w_i\}) \le Cp\exp\left\{-C\frac{a^2}{\max_j \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}_j\|^2}\right\}.$$

Using $\max_j \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}_j\|^2 \le \max_j \|\mathbf{X}_j\|^2 = O_p(n)$, we get $\|\boldsymbol{\epsilon}^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\|_\infty \le C(\sqrt{n\log(p \vee N)})$. $\square$

LEMMA 3. *The eigenvalues of* $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$ *are bounded away from zero and infinity, with probability at least* $1-(p \vee N)^{-C}$.

PROOF. Since $|B_k(w)B_{k'}(w)| \le K$, $E[B_k^2(w)B_{k'}^2(w)] \le CK$, this is trivially based on Bernstein's inequality

$$P\left(\left|\sum_i B_k(w_i)B_{k'}(w_i) - nE[B_k(w)B_{k'}(w)]\right| > a\right) < C\exp\left\{-C\frac{a^2}{aK+nK}\right\}.$$

$\square$

LEMMA 4.

$$\inf_{\|\Delta_{T^c}\|_1 \le c\|\Delta_T\|_1} \frac{\|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta\|}{\sqrt{n}\|\boldsymbol{\Delta}\|} \ge \kappa/2,$$

with probability at least $1-(p\vee N)^{-C}$.

PROOF.

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})\mathbf{X}\Delta\| \ge{}& \|(\mathbf{X}-E[\mathbf{X}|\mathbf{w}])\Delta\| - \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})E[\mathbf{X}|\mathbf{w}]\Delta\| \\ &- \|\mathbf{P}_{\mathbf{Z}}(\mathbf{X}-E[\mathbf{X}|\mathbf{w}])\Delta\|.\end{aligned}$$

We have

$$\begin{aligned}&\|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})E[\mathbf{X}|\mathbf{w}]\Delta\|^2 \\ &\quad= \|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0)\Delta\|^2 \\ &\quad\le \|\boldsymbol{\Delta}\|_1^2 \|(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0)^{\mathrm{T}}(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})(E[\mathbf{X}|\mathbf{w}]-\mathbf{Z}\boldsymbol{\alpha}_0)\|_\infty \\ &\quad\le CnK^{-2\alpha'}\|\boldsymbol{\Delta}\|_1^2,\end{aligned}$$

where $\boldsymbol{\alpha}_0$ is defined in (A.1), and thus

$$\begin{aligned}\|(\mathbf{I}-\mathbf{P}_{\mathbf{Z}})E[\mathbf{X}|\mathbf{w}]\Delta\| &\le C\sqrt{n}K^{-\alpha'}\|\boldsymbol{\Delta}\|_1 \\ &\le C\sqrt{n}K^{-\alpha'}\|\boldsymbol{\Delta}_T\|_1 \\ &\le C\sqrt{n}K^{-\alpha'}\sqrt{s}\|\boldsymbol{\Delta}_T\| \\ &\le C\sqrt{n}K^{-\alpha'}\sqrt{s}\|\boldsymbol{\Delta}\| \\ &= o(\sqrt{n})\|\boldsymbol{\Delta}\|.\end{aligned}$$

Furthermore,

$$\begin{aligned}&\|\mathbf{P}_{\mathbf{Z}}(\mathbf{X}-E[\mathbf{X}|\mathbf{w}])\Delta\|^2 \\ &\quad\le \|\Delta\|_1^2\|(\mathbf{X}-E[\mathbf{X}|\mathbf{w}])^{\mathrm{T}}\mathbf{P}_{\mathbf{Z}}(\mathbf{X}-E[\mathbf{X}|\mathbf{w}])\|_\infty \\ &\quad\le \|\Delta\|_1^2\|(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\|_{\mathrm{op}}\max_j\|\mathbf{Z}(\mathbf{X}_j-E[\mathbf{X}_j|\mathbf{w}])\|^2 \\ &\quad\le C\|\Delta\|_1^2 K\log(p\vee N),\end{aligned}$$

and thus

$$\begin{aligned}&\|\mathbf{P}_{\mathbf{Z}}(\mathbf{X}-E_{\mathrm{add}}[\mathbf{X}])\Delta\| \\ &\quad\le C\sqrt{K\log(p\vee N)}\|\Delta\|_1 \\ &\quad\le C\sqrt{sK\log(p\vee N)}\|\Delta\| \\ &\quad= o(\sqrt{n})\|\Delta\|.\end{aligned}$$

The proof is complete by assumption (A7). □

LEMMA 5. *With probability at least* $1-(p\vee N)^{-C}$,

$$\begin{aligned}&\|(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{Z}-E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})]\|_{\max}\\&\quad\le Cs^2\log(p\vee N)+C\sqrt{n\bigl(K+\log(p\vee N)\bigr)}\log K.\end{aligned}$$

PROOF. We have

$$\begin{aligned}&(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})-(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})\\&\quad=-(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})+(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})\\&\qquad+(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}).\end{aligned}$$

For the first term above,

$$\begin{aligned}\max_{k,k'}\bigl|(\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\widehat{\boldsymbol{\gamma}}_{k'}-\boldsymbol{\gamma}_{k'})\bigr|&\le\|\mathbf{X}^{\mathrm{T}}\mathbf{X}\|_{\max}\max_k\|\widehat{\boldsymbol{\gamma}}_k-\boldsymbol{\gamma}_k\|_1^2\\&\le Cs^2\log(p\vee N).\end{aligned}$$

By the KKT condition for (2.11), $\|\mathbf{X}^{\mathrm{T}}(\mathbf{Z}_k-\mathbf{X}\widehat{\boldsymbol{\gamma}}_k)\|_\infty\le\lambda$, and thus

$$\|(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})\|_{\max}\le C\lambda s\sqrt{\log(p\vee N)/n}.$$

Thus

$$\|(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})-(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})\|_{\max}\le Cs^2\log(p\vee N).\tag{A.6}$$

Let $\Omega=\{\mathbf{x}:\|\mathbf{x}\|_\infty\le C\sqrt{\log(p\vee N)}\}$. Then $\{\mathbf{x}_i\in\Omega,\forall i=1,\dots,N\}$ holds with probability at least $1-(p\vee N)^{-C}$. We have $|B_k(w_i)-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_k|I\{\mathbf{x}_i\in\Omega\}\le C\sqrt{K+\log(p\vee N)}$ and $E|B_k(w_i)-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_k|^2I\{\mathbf{x}_i\in\Omega\}\le C$. Thus by Bernstein's inequality,

$$\begin{aligned}&P\Bigl(\sum_i|B_k(w_i)-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_k|^2I\{\mathbf{x}_i\in\Omega\}-nE\bigl[|B_k(w_i)-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\gamma}_k|^2I\{\mathbf{x}_i\in\Omega\}\bigr]>a\Bigr)\\&\quad\le C\exp\biggl\{-C\frac{a^2}{a(K+\log(p\vee N))+n(K+\log(p\vee N))}\biggr\}.\end{aligned}$$

Thus with probability at least $1-(p\vee N)^{-C}$,

$$\begin{aligned}&\|(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})-E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})]\|_{\max}\\&\quad\le C\sqrt{n\bigl(K+\log(p\vee N)\bigr)}\log K.\end{aligned}$$

Finally, using the KKT condition again, we have

$$\|\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})\|_{\max}\le\max_k\|\widehat{\boldsymbol{\gamma}}_k\|_1\|\mathbf{X}^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\widehat{\boldsymbol{\gamma}})\|_{\max}\le C\lambda,$$

and the lemma is proved by combining the two displayed equations above and (A.6). □

LEMMA 6.

$$\begin{aligned}&(E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})])^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\\&\quad=(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}}-(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\\&\qquad\times(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]-E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}])^{-1}\\&\qquad\times(\mathbf{X}^{\mathrm{T}}-E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}}).\end{aligned}$$

PROOF. Since $\boldsymbol{\gamma}=(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]$,

$$\begin{aligned}&(E[(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})])^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\\&\quad=(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}]-E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}])^{-1}(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\\&\quad=((E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}+(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}]\\&\qquad\times(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]-E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}])^{-1}\\&\qquad\times E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1})(\mathbf{Z}-\mathbf{X}\boldsymbol{\gamma})^{\mathrm{T}}\\&\quad=(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}}\\&\qquad+(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]-E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}])^{-1}\\&\qquad\times E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}\mathbf{Z}^{\mathrm{T}}\\&\qquad-(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}\mathbf{X}^{\mathrm{T}}\\&\qquad-(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}]-E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}])^{-1}\\&\qquad\times E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}](E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}E[\mathbf{Z}^{\mathrm{T}}\mathbf{X}](E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}\mathbf{X}^{\mathrm{T}},\end{aligned}$$

where the second equality used the Woodbury identity. Let $\mathbf{A}=(E[\mathbf{X}^{\mathrm{T}}\mathbf{X}])^{-1}$, $\mathbf{B}=E[\mathbf{X}^{\mathrm{T}}\mathbf{Z}]$ and $\mathbf{C}=(E[\mathbf{Z}^{\mathrm{T}}\mathbf{Z}])^{-1}$. Based on the expression above, we only need to show that

$$\mathbf{A}+(\mathbf{A}^{-1}-\mathbf{BCB}^{\mathrm{T}})^{-1}\mathbf{BCB}^{\mathrm{T}}\mathbf{A}=(\mathbf{A}^{-1}-\mathbf{BCB}^{\mathrm{T}})^{-1}.$$

The above identity is easily validated by left-multiplying both sides by $(\mathbf{A}^{-1}-\mathbf{BCB}^{\mathrm{T}})$. □

LEMMA 7. *For any x and any (random) matrix $\mathbf{A}$,*

$$\|B^{\mathrm{T}}(x)(NE[B(w)B^{\mathrm{T}}(w)])^{-1}\mathbf{A}\|\le C(\sqrt{K}/N)\max_j\|\mathbf{A}_{j\cdot}\|.$$

The above also holds if $\|\cdot\|$ is replaced by any other vector norm on both sides.

REMARK 7. This key lemma improves upon a more naive bound based on $\|B^{\mathrm{T}}(x)(NE[B(w)B^{\mathrm{T}}(w)])^{-1}\mathbf{A}\| \leq \|B(x)\|\|(NE[B(w)B^{\mathrm{T}}(w)])^{-1}\|_{\mathrm{op}}\|\mathbf{A}\| \leq C(\sqrt{K_1}/N)\|\mathbf{A}\|$. It uses the fact that $B(x)$ only has at most q nonzero components and the peculiar structure of $(NE[B(w)B^{\mathrm{T}}(w)])^{-1}$ as in the proof.

PROOF OF LEMMA 7. Let $\mathbf{e}_j$ be the unit vector with a single one in the jth position, since $B(x)$ only has at most q nonzero components and its nonzero components are bounded by $\sqrt{K}$. Thus we can write $B(x) = \sum_j v_j \mathbf{e}_j$ where at most q of the values v_j is nonzero and these values are no larger than $\sqrt{K}$. Thus we only need to prove that for any j,

$$\|\mathbf{e}_j^{\mathrm{T}}(NE[B(w)B^{\mathrm{T}}(w)])^{-1}\mathbf{A}\| \leq (C/n)\max_j \|\mathbf{A}_{j\cdot}\|. \tag{A.7}$$

By Theorem 2.2 of Demko (1977), the (j, j') entry of $(NE[B(w)B^{\mathrm{T}}(w)])^{-1}$ is bounded by $(C/N)\gamma^{|j-j'|}$ for some $\gamma < 1$. Thus

$$\begin{aligned}
&\|\mathbf{e}_j^{\mathrm{T}}(NE[B(w)B^{\mathrm{T}}(w)])^{-1}\mathbf{A}\| \\
&\quad \leq (C/N)\sum_{j'}\gamma^{|j-j'|}\|\mathbf{A}_{j'\cdot}\| \\
&\quad \leq (C/N)\max_j \|\mathbf{A}_{j\cdot}\|.
\end{aligned}$$

□

DISCUSSIONS ON ASSUMPTION (A7)

We will first discuss how the restricted eigenvalue condition (A7) can be implied by a related sparse Riesz condition (SRC) and then consider how sparse Riesz condition can be satisfied in the semiparametric case. We do not aim to conduct a comprehensive study on these eigenvalue assumptions or provide very general sufficient conditions for (A7). The main goal is just to show that (A7) is a reasonable assumption and we make some further simplifying assumptions to facilitate this discussion.

We will relate (A7) to sparse Riesz condition as in Bickel, Ritov and Tsybakov (2009). Let $A \subseteq \{1, \ldots, p\}$ and denote by $\mathbf{u}_{iA}$ the subvector of $\mathbf{u}_i := \mathbf{x}_i - E[\mathbf{x}|w = w_i]$ containing only components associated with predictors in A. Define $c^*(v) = \sup_{|A|\leq v, \|\delta\|=1}\sum_i \delta^{\mathrm{T}}\mathbf{u}_{iA}\mathbf{u}_{iA}^{\mathrm{T}}\delta/N$ and $c_*(v) = \inf_{|A|\leq v, \|\delta\|=1}\sum_i \delta^{\mathrm{T}}\mathbf{u}_{iA}\mathbf{u}_{iA}^{\mathrm{T}}\delta/N$. Conditions on the magnitudes of $c^*(v)$ and $c_*(v)$ are usually referred to as sparse Riesz conditions.

The following discussions are mainly adapted from Bickel, Ritov and Tsybakov (2009), in particular, their proof of Lemma 4.1 and we only focus on the modifications required. The paper of Bickel, Ritov and Tsybakov (2009) contains other sufficient conditions for restricted eigenvalue assumption but here we only focus on part (ii) of their Lemma 4.1.

In the following discussion, we suppose the covariates are rearranged such that $T = \{1, \dots, s\}$ are the indices for all the nonzero components in $\boldsymbol{\beta}_0$ and $T^c = \{s + 1, \dots, p\}$. Also, for the p-dimensional vector δ as used in (A7), we write $\delta = (\delta_1, \dots, \delta_p)$.

Given δ satisfying the constraint $\sum_{j\in T^c} |\delta_j| \le c \sum_{j\in T} |\delta_j|$, we partition T^c into subsets of size ℓ with last subset of size $\le \ell$ (we will set $\ell = s \log N - s$ later). Thus we write $T^c = \bigcup_{h=1}^{H} T_h$ where T_h contains the indices j corresponding to ℓ largest $|\delta_j|$ outside of $\bigcup_{k=1}^{h-1} T_k$. Let $T_{01} = T \cup T_1$. Using the same arguments as in the proof of Lemma 4.1 of Bickel, Ritov and Tsybakov (2009), we get $\frac{\sum_i \delta^{\mathrm{T}}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\delta}{N} \ge (\sqrt{c_*(s+\ell)} - c\sqrt{c^*(\ell)}\sqrt{s/\ell}))\|\delta_{T_{01}}\|^2$. Furthermore, since the kth largest value among $|\delta_j|, s + 1 \le j \le p$ satisfies $|\delta_j| \le \sum_{s+1\le j\le p} |\delta_j|/k$, we have $\|\delta_{T_{01}^c}\|^2 \le (\sum_{s+1\le j\le p} |\delta_j|)^2 \sum_{k\ge \ell+1}(1/k^2) \le (\sum_{s+1\le j\le p} |\delta_j|)^2/\ell$, and thus

$$
\begin{aligned}
\|\delta\| &\le \|\delta_{T_{01}}\| + \|\delta_{T_{01}^c}\| \le \|\delta_{T_{01}}\| + \frac{\sum_{s+1\le j\le p} |\delta_j|}{\sqrt{\ell}} \\
&\le \|\delta_{T_{01}}\| + c\frac{\sum_{1\le j\le s} |\delta_j|}{\sqrt{\ell}} \le \|\delta_{T_{01}}\| + c\sqrt{s/\ell}\|\delta_T\| \\
&\le (1 + c\sqrt{s/\ell})\|\delta_{T_{01}}\|.
\end{aligned}
$$

Then we have

$$
\frac{\sum_i \delta^{\mathrm{T}}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}\delta}{(N)\|\delta\|^2} \ge C(1 + c\sqrt{s/\ell})^{-2}\bigl(\sqrt{c_*(s+\ell)} - c\sqrt{c^*(\ell)}\sqrt{s/\ell})\bigr).
$$

As a result, the above is bounded away from zero if $s + \ell = s \log N$ and $c_*(s \log N)$ is bounded away from zero, and $c^*(\ell)$ is bounded.

We now consider how $c^*(v)$ and $c_*(v)$ can be bounded and bounded away from zero for v not too large. The calculation below is a standard application of Bernstein's inequality for subexponential random variables. Assuming x_j is sub-Gaussian and also naturally that $E[x_j|w]$ is a bounded function of w, then $u_j := x_j - E[x_j|w]$ is also sub-Gaussian. It is well known that the product of two sub-Gaussian random variables is a subexponential random variable. Then Bernstein's inequality yields

$$
P\left(\left|\sum_{i=1}^{N} u_j u_{j'}/N - E[u_j u_{j'}]\right| > t\right) \le C \exp\{-Cnt^2/(t + 1)\}
$$

and thus by union bound

$$
P\left(\sup_{j,j'}\left|\sum_{i=1}^{N} u_j u_{j'}/N - E[u_j u_{j'}]\right| > t\right) \le Cp^2 \exp\{-Cnt^2/(t + 1)\},
$$

which in turn implies

$$\sup_{|A|\le v}\left\|\sum_i \mathbf{u}_{iA}\mathbf{u}_{iA}^{\mathrm{T}}/N - E\left[\mathbf{u}_{iA}\mathbf{u}_{iA}^{\mathrm{T}}\right]\right\|_{\mathrm{op}} = O_p\left(v\sqrt{\log p/N}\right).$$

When $v\sqrt{\log p/N} = o(1)$, the above together with Assumption (A5) implies $c^*(v)$ and $c_*(v)$ are bounded and bounded away from zero. This means (A7) can be satisfied under suitable conditions and concludes our discussion.

Acknowledgments. The authors sincerely thank the Editor, Professor Edward George, the Associate Editor and three anonymous reviewers for their insightful comments that lead to significant improvement of the work.

REFERENCES

BANERJEE, M., DUROT, C. and SEN, B. (2017). Divide and conquer in non-standard problems and the super-efficiency phenomenon. *Ann. Statist.* To appear.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. MR3308656

DE BOOR, C. (2001). *A Practical Guide to Splines*, Revised ed. *Applied Mathematical Sciences* **27**. Springer, New York. MR1900298

DEMKO, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM J. Numer. Anal.* **14** 616–619. MR0455281

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. *Springer Series in Statistics*. Springer, New York. MR1964455

HAERDLE, W., LIANG, H. and GAO, J. (2007). *Partially linear models*. Springer, New York.

HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827

HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Stat.* **34** 451–477. MR2368793

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677

LI, Q. (2000). Efficient estimation of additive partially linear models. *Internat. Econom. Rev.* **41** 1073–1092. MR1790072

LIANG, H. and LI, R. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104** 234–248. MR2504375

LV, S. and LIAN, H. (2017). A debiased distributed estimation for sparse partially linear models in diverging dimensions. Available at arXiv:1708.05487.

SCHUMAKER, L. L. (2007). *Spline Functions*: *Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR2348176

SHI, C., LU, W. and SONG, R. (2017). A massive data framework for M-estimators with cubic-rate. *J. Amer. Statist. Assoc.* To appear.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

WAINWRIGHT, M. (2018). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Draft book.

WANG, L., LIU, X., LIANG, H. and CARROLL, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.* **39** 1827–1851. MR2893854

XIE, H. and HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37** 673–696. MR2502647

ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340. MR3450540

ZHAO, T., CHENG, G. and LIU, H. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist.* **44** 1400–1437. MR3519928

ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277

H. LIAN
DEPARTMENT OF MATHEMATICS
CITY UNIVERSITY OF HONG KONG
83 TAT CHEE AVE
KOWLOON
HONG KONG
E-MAIL: henglian@cityu.edu.hk

K. ZHAO
BIG DATA & AI
PHILIPS RESEARCH CHINA
718 LINGSHI ROAD
SHANGHAI 200040
CHINA
E-MAIL: kaifengzhao66@hotmail.com

S. LV
DEPARTMENT OF STATISTICS AND MATHEMATICS
NANJING AUDIT UNIVERSITY
NANJING 211815
CHINA
E-MAIL: lvsg716@nau.edu.cn