High Dimensional Inference in Partially Linear Models

Ying Zhu
Purdue University (West Lafayette)
Department of Statistics
Department of Computer Science

Zhuqing Yu
AbbVie Inc.
North Chicago, IL

Guang Cheng
Purdue University (West Lafayette)
Department of Statistics

Abstract

We propose two semiparametric versions of the debiased Lasso procedure for the model $Y_i = X_i\beta_0 + g_0(Z_i) + \varepsilon_i$, where the parameter vector of interest β_0 is high dimensional but sparse (exactly or approximately) and g_0 is an unknown nuisance function. Both versions are shown to have the same asymptotic normal distribution and do not require the minimal signal condition for statistical inference of any component in β_0. We further develop a simultaneous hypothesis testing procedure based on multiplier bootstrap. Our testing method takes into account of the dependence structure within the debiased estimates, and allows the number of tested components to be exponentially high.

1 Introduction

Semiparametric regression is a longstanding statistical tool that leverages the flexibility of nonparametric models while avoiding the "curse of dimensionality" (see, e.g., Bickel, et al., 1998). A leading example of semiparametric regression models is the partially linear regression

$$Y_i = X_i\beta_0 + g_0(Z_i) + \varepsilon_i,\ i = 1, ..., n. \tag{1}$$

In (1), $\beta_0 \in \mathbb{R}^p$ is an unknown vector and g_0 is an unknown function; $X := (X_i)_{i=1}^n \in \mathbb{R}^{n\times p}$ and $Z := (Z_i)_{i=1}^n \in \mathbb{R}^{n\times d}$ are observed covariates (X_i and Z_i denote the ith row of X and Z, respectively), $Y := (Y_i)_{i=1}^n \in \mathbb{R}^n$ are the response variables, $\varepsilon = (\varepsilon_i)_{i=1}^n \in \mathbb{R}^n$ is a noise vector with $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = 1$, and independent of (X, Z). Throughout the paper, we assume that the data $\{X_i, Z_i, Y_i\}_{i=1}^n$ are *i.i.d.*. The goal of this paper is to establish statistical inference results, e.g., confidence intervals and hypothesis testing, for the high dimensional component β_0 in presence of the nuisance function g_0. In particular, we assume that $p \geq n$ and β_0 exhibits sufficient sparsity (meaning that the ordered coefficients in β_0 decay sufficiently fast). Our method also works when Z_i is high dimensional ($d \geq n$) provided that the function classes $\mathbb{E}(X_{ij}|Z_i)$s and $\mathbb{E}(Y_i|Z_i)$ belong to exhibit certain sparsity features, e.g., a sparse additive decomposition structure as defined in Raskutti, et al., (2012).

For statistical inference of β_0 in (1), existing results mainly focus on the regime where p increases with n but smaller than n, for example, Li and Liang (2008), Xie and Huang (2009), and Cheng, et al. (2015). Sherwood and Wang (2016) allow $p \geq n$ but require the minimal signal condition. Such results therefore suffer the problems arising from the nonuniformity of limit theory. More recently, Javanmard and Montanari (2014), van de Geer, et al. (2014), and Zhang and Zhang (2014) have proposed the debiased Lasso for high dimensional linear models. These estimators are non-sparse, have a limiting normal distribution, and do not require the minimal signal condition. For the linear model $Y = X\beta_0 + \varepsilon$, given an initial Lasso estimate $\hat{\beta}$ of β_0, the debiased Lasso adds a correction term to $\hat{\beta}_j$ (the jth component of $\hat{\beta}$) to remove the bias introduced by regularization. In particular, the correction term takes the form of

$$\hat{\Gamma}_j \frac{1}{n} X^T \left(Y - X\hat{\beta}\right). \tag{2}$$

In (2), $n^{-1}X^T\left(Y - X\hat{\beta}\right)$ is the sample analogue of the population score function $\mathbb{E}\left(X_i^T\left(Y_i - X_i\beta_0\right)\right)$; $\hat{\Gamma}_j$ denotes the jth row of $\hat{\Gamma}$ where $\hat{\Gamma}$ is an approximate inverse of $n^{-1}X^TX$, whose population counterpart is $\mathbb{E}\left(X_i^TX_i\right)$. In our model (1), additional bias arises due to the presence of g_0; consequently, the standard debiased Lasso cannot rid of the effect from g_0 and thus will not have a limiting distribution centered around

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

zero. Instead, we propose two modified versions of the debiased Lasso estimators for β_0. Both versions are shown to be asymptotically unbiased for β_0, have the same limiting (normal) distribution, and do not require the minimal signal condition.

Our modified debiased Lasso estimators use a "nonparametric projection" strategy to remove the impact of g_0 in (1). Such a strategy has been used in the semiparametric inference literature where p is assumed to be small relative to n (e.g., Robinson, 1988; Donald and Newey, 1994; Liang and Li, 2009). To be more specific, by taking the conditional expectations of the left side and the right side of (1) with respect to Z_i, we obtain

$$\mathbb{E}\left(Y_i|Z_i\right) = \mathbb{E}\left(X_i|Z_i\right)\beta_0 + g_0(Z_i) \tag{3}$$

where we exploit the fact that $\mathbb{E}\left(\varepsilon_i|Z_i\right) = 0$. Subtracting $\mathbb{E}\left(Y_i|Z_i\right)$ from Y_i and $\mathbb{E}\left(X_i|Z_i\right) + g_0(Z_i)$ from $X_i + g_0(Z_i)$ in (1) yields

$$\tilde{Y}_i = \tilde{X}_i\beta_0 + \varepsilon_i \tag{4}$$

where $\tilde{Y}_i := Y_i - \mathbb{E}\left(Y_i|Z_i\right)$, $\tilde{X}_{ij} := X_{ij} - \mathbb{E}\left(X_{ij}|Z_i\right)$ and $\tilde{X}_i := \left(\tilde{X}_{ij}\right)_{j=1}^{p}$ (which is a $p-$dimensional row vector).

Relating (4) to the linear model $Y_i = X_i\beta_0 + \varepsilon_i$, given nonparametric surrogates $\hat{Y}_i := Y_i - \hat{\mathbb{E}}\left(Y_i|Z_i\right)$ of $\tilde{Y}_i$ and $\hat{X}_{ij} := X_{ij} - \hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$ of $\tilde{X}_{ij}$ $(j = 1, ..., p)$, we simply replace Y_i with $\hat{Y}_i$, X_i with the row vector $\hat{X}_i := \left(\hat{X}_{ij}\right)_{j=1}^{p}$, and $\hat{\Gamma}_j$ with the jth row $(\hat{\Theta}_j)$ of an approximate inverse (denoted as $\hat{\Theta}$) of $n^{-1}\sum_{i=1}^{n}\hat{X}_i^T\hat{X}_i$ in (2). This yields our first semiparametric version of the debiased procedure

$$\hat{b}_j := \hat{\beta}_j + \hat{\Theta}_j\frac{1}{n}\hat{X}^T\left(\hat{Y} - \hat{X}\hat{\beta}\right), \tag{5}$$

where $\hat{\beta}$ is an initial estimate of β_0. Alternatively, by noting that $n^{-1}\hat{X}^T\left(\hat{Y} - \hat{X}\hat{\beta}\right)$ in (5) is simply the sample analogue of the population score function $\mathbb{E}\left(\tilde{X}_i^T\varepsilon_i\right)$, we arrive at our second debiased procedure

$$\tilde{b}_j := \hat{\beta}_j + \hat{\Theta}_j\frac{1}{n}\hat{X}^T\left(Y - X\hat{\beta} - \hat{g}\right), \tag{6}$$

where $\hat{g}$ is an estimate of g_0.

We provide theoretical implications on the impact of the estimation errors associated with the p nonparametric surrogates $\hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$s in our modified debiased Lasso procedures when each of $\hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$s concerns a large family of (regularized) nonparametric least squares estimators. These implications also hold true for the surrogate $\hat{\mathbb{E}}\left(Y_i|Z_i\right)$ (which matters to (5)) and the surrogate $\hat{g}(Z_i)$ (which matters to (6)). After careful theoretical analysis, we find that if the error of the nonparametric regression *per se* (with respect to the prediction norm) is $O_p\left(r_n\right)$, it only contributes $O_p\left(r_n^2\right)$ in the asymptotic expansions of $\hat{b}_j - \beta_{0j}$ and $\tilde{b}_j - \beta_{0j}$ for any $j = 1, ..., p$, where r_n is related to the optimal rate for the nonparametric regression. This result implies that even with p much larger than n (and/or with the dimension d of Z_i much larger than n), the limiting distribution of our modified debiased estimators for any individual component in β_0 may behave as if the unknown conditional expectations $\mathbb{E}\left(X_{ij}|Z_i\right)$s and $\mathbb{E}\left(Y_i|Z_i\right)$ as well as the unknown function $g_0(Z_i)$ were known.

This theoretical finding motivates us to consider a multiplier-bootstrap-based *simultaneous* hypothesis testing procedure for any sub-vector of β_0. This extends the method developed by Zhang and Cheng (2017) from linear regressions to more flexible partially linear regressions. Our simultaneous testing procedure takes into account of the dependence structure within our debiased estimators, and allows the number of tested components to be exponentially high.

We illustrate the theoretical finding with three specific examples in terms of $\dim\left(Z_i\right)$ and the function class that $\mathbb{E}\left(X_{ij}|Z_i\right)$s and $\mathbb{E}\left(Y_i|Z_i\right)$ belong to. With regard to the specific forms of $\hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$s and $\hat{\mathbb{E}}\left(Y_i|Z_i\right)$, several modern techniques for the projection step are considered and the rates achieved by these practical procedures are compared with the theoretical results. The techniques discussed in the paper include the smoothing splines estimator in Sobolev balls, the Lasso (Tibshirani, 1996) and Slope (Su and Candés, 2016) in sparse linear regression models, and the l_1-regularized kernel ridge regression (Raskutti, et al., 2012) in sparse additive models.

The testing procedures proposed in this paper are important for evaluations of policy interventions in social science as well as clinical trials in precision medicine. Suppose U_i is a binary variable that indicates whether or not individual i receives "treatment" or not, and $Z_i = (Z_{i1}, ..., Z_{id})$ is the high dimensional vector of control variables. Let us consider the following model:

$$Y_i = U_i\alpha_0 + \sum_{l=1}^{d}U_iZ_{il}\gamma_{0l} + g_0\left(Z_i\right) + \varepsilon_i,\ i = 1, ..., n.$$

In particular, an interesting hypothesis would be H_0 : $\gamma_{01} = \gamma_{02} = \cdots = \gamma_{0d} = 0$; that is, the interactions of the treatment variable and the high dimensional controls have no effects on the outcome of interest, Y_i.

The rest of the paper is organized as follows. Section 2 presents the detailed construction of the modified debiased estimators for β_0 and the simultaneous testing procedure. Section 3 establishes the main theoretical results. The proposed method is illustrated with simulated studies and a real data example in Section 4. All technical details are deferred to the supplementary material.

Notation. The l_q-norm of a $p-$dimensional vector Δ is denoted by $\left\|\Delta\right\|_q$ for $1 \leq q \leq \infty$. For a matrix $H \in \mathbb{R}^{p_1 \times p_2}$, write $\left\|H\right\|_\infty := \max_{i,j}|H_{ij}|$ to be the elementwise $l_\infty-$norm of H. Let H_j denote the jth row of H. For a matrix $H \in \mathbb{R}^{m\times m}$, the minimum eigenvalue of H is denoted by $\Lambda^2_{\min}(H)$ and the maximum eigenvalue of H is denoted by $\Lambda^2_{\max}(H)$. The $\mathcal{L}^2(\mathbb{P}_n)-$norm of the vector $\Delta := \{\Delta(X_i)\}_{i=1}^n$, denoted by $\left\|\Delta\right\|_n$, is given by $\left[\frac{1}{n}\sum_{i=1}^n(\Delta(X_i))^2\right]^{\frac{1}{2}}$. For functions $f(n)$ and $g(n)$, write $f(n) \succsim g(n)$ to mean that $f(n) \geq cg(n)$ for a universal constant $c \in (0, \infty)$ and similarly, $f(n) \precsim g(n)$ to mean that $f(n) \leq c^{'}g(n)$ for a universal constant $c^{'} \in (0, \infty)$, and $f(n) \asymp g(n)$ when $f(n) \succsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. Also denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$. As a general rule for this paper, all the $c \in (0, \infty)$ constants denote positive universal constants. The specific values of these constants may change from place to place.

2 Main methodology

In this section we discuss the construction of $\hat{b}_j$ and $\tilde{b}_j$ in detail. Note that both (5) and (6) require estimators for the conditional expectations, an initial estimator $\hat{\beta}$ for β_0 (and an estimator $\hat{g}$ for g_0 in $\tilde{b}_j$), and also an approximate inverse $\hat{\Theta}$ for $n^{-1}\sum_{i=1}^n \hat{X}_i^T\hat{X}_i$. We first discuss how to obtain these aforementioned quantities. Given $\hat{b}_j$s and $\tilde{b}_j$s, we then present the simultaneous inference procedure.

Estimators for the conditional expectations

For either (5) or (6), we need to estimate the conditional expectations $\mathbb{E}(X_{ij}|Z_i)$s $(j = 1, ..., p)$. This step is easily paralleable as it involves solving p independent subproblems and each subproblem can be in general solved with an efficient algorithm. In contrast with (6), (5) does not require an estimate of g_0 but an estimate of $\mathbb{E}(Y_i|Z_i)$. Estimating conditional expectations is widely studied in the literature on nonparametric methods. For the purpose of this paper, global properties of the nonparametric estimators $\hat{\mathbb{E}}(X_{ij}|Z_i)$s and $\hat{\mathbb{E}}(Y_i|Z_i)$ are the key to our analysis of the debiased procedures and therefore, we focus on the following least squares estimators

$$\hat{f}_j \in \arg\min_{f_j \in \mathcal{F}_j}\left\{\frac{1}{2n}\sum_{i=1}^n(w_{ij} - f_j(z_i))^2\right\}, \qquad (7)$$

where $w_{i0} = y_i$ and $w_{ij} = x_{ij}$ for $j = 1, ..., p$. Denote $\hat{f}_0(Z_i)$ as $\hat{\mathbb{E}}(Y_i|Z_i)$ and $\hat{f}_j(Z_i)$ as $\hat{\mathbb{E}}(X_{ij}|Z_i)$.

A nonparametric regression problem like (7) is a standard setup in many modern statistics books (e.g., van de Geer, 2000; Wainwright, 2015). Examples of (7) include linear regression, sparse linear regression, series projection, convex regression, Lipschitz Isotonic regression, and kernel ridge regression (KRR). In the case of KRR, we restrict $\mathcal{F}_j$ in (7) to be a compact subset of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, equipped with a norm $\left\|\cdot\right\|_{\mathcal{H}}$; (7)[1] can then be reformulated in its Lagrangian form

$$\hat{f}_j \in \arg\min_{f_j \in \mathcal{H}}\left\{\frac{1}{2n}\sum_{i=1}^n(w_{ij} - f_j(z_i))^2 + \mu_j\left\|f_j\right\|^2_{\mathcal{H}}\right\} \qquad (8)$$

where $\mu_j > 0$ is a regularization parameter. In particular, smoothing spline estimators can be viewed as special cases of KRR.

Initial estimators for β_0 and g_0

In a semiparametric regression model like (1), Zhu (2017) covers a wide spectrum of function classes that the nonparametric component $g_0(\cdot)$ may belong to and provides a general nonasymptotic theory for estimating β_0 and g_0. The estimators $\hat{\beta}$ and $\hat{g}$ in Zhu (2017) can be used as initial estimators in (5)-(6). Given the way $\hat{\beta}$ is obtained in Zhu (2017), the estimated conditional expectations, $\hat{f}_j$s, come in handy as byproducts (therefore, separate estimations for the conditional expectations are not needed in the construction of $\hat{b}_j$ or $\tilde{b}_j$).

For special cases where Z_i has a low dimension and g_0 belongs to the mth order Sobolev ball $\mathcal{S}^m$, other estimators for β_0 and g_0 are also available (see, Müller and van de Geer, 2015; Yu, et al., 2017).

Due to the intractable limiting distribution of Lasso type estimators, these aforementioned papers do not provide any distributional results for their proposed estimators. In Section 3, we take the debiased versions, (5) and (6), of these aforementioned initial estimators and establish the asymptotic normality of individual components in the debiased estimators.

Estimator for the inverse of the population Hessian

Given the estimates $\hat{Y}_i$ of $\tilde{Y}_i$ and $\hat{X}_i$ of $\tilde{X}_i$ via (7), we obtain an approximate inverse $\hat{\Theta}$ of $n^{-1}\sum_{i=1}^n\hat{X}_i^T\hat{X}_i$ using the nodewise regression method proposed by

[1] To be more specific, we let $\mathcal{F}_j$ be a ball of radius R in the norm $\left\|\cdot\right\|_{\mathcal{H}}$ and assume $R \leq 1$ throughout the asymptotic analysis to avoid carrying "R"s around.

van de Geer, et al. (2014). Since our analysis involves establishing $\left\|\hat{\Theta}_j - \Theta_j\right\|_1 = o_p(1)$, as in van de Geer, et al. (2014), we require a sparsity condition on the inverse $\Theta = \Sigma^{-1}$ of the population Hessian $\Sigma := \mathbb{E}\left(\tilde{X}_i^T \tilde{X}_i\right)$. Lack of sparsity in the off-diagonal elements of Θ will cause remainder terms like $\left(\hat{\Theta}_j - \Theta_j\right) \frac{1}{\sqrt{n}} \tilde{X}^T \varepsilon$ in the asymptotic expansions of $\sqrt{n}\left(\hat{b}_j - \beta_{0j}\right)$ or $\sqrt{n}\left(\tilde{b}_j - \beta_{0j}\right)$ to diverge and the resulting limiting distribution may not be well-behaved for any practical purpose. This fact renders the method in Javanmard and Montanari (2014) for constructing $\hat{\Theta}$ inapplicable as their approach is only valid for fixed X, while our analysis accounts for the randomness in X and the estimation errors in $\hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$s.

To apply the nodewise regression method in our context, for each $1 \leq j \leq p$, let us define

$$\hat{\pi}_j \in \arg\min_{\tilde{\pi}_j \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \left\| \hat{X}_j - \hat{X}_{-j} \tilde{\pi}_j \right\|_2^2 + \lambda_j \left\| \tilde{\pi}_j \right\|_1 \right\}, \tag{9}$$

where $\hat{X}_{-j}$ denotes the submatrix of $\hat{X}$ without the jth column. Let

$$\hat{M} := \begin{pmatrix} 1 & -\hat{\pi}_{1,2} & \cdots & -\hat{\pi}_{1,p} \\ -\hat{\pi}_{2,1} & 1 & \cdots & -\hat{\pi}_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\pi}_{p,1} & -\hat{\pi}_{p,2} & \cdots & 1 \end{pmatrix}.$$

Based on $\hat{\pi}_j := \left\{\hat{\pi}_{j,j'};\, j' \neq j\right\}$, for each $1 \leq j \leq p$, we compute

$$\hat{\tau}_j^2 := \frac{1}{n} \left\| \hat{X}_j - \hat{X}_{-j} \hat{\pi}_j \right\|_2^2 + \lambda_j \left\| \hat{\pi}_j \right\|_1$$

and write

$$\hat{T}^2 := \operatorname{diag}\left(\hat{\tau}_1^2, ..., \hat{\tau}_p^2\right).$$

Finally, we define $\hat{\Theta} := \hat{T}^{-2} \hat{M}$.

For later presentations of the theoretical results, we also introduce the population counterparts of the above quantities: let π_j be the population regression coefficients of $\tilde{X}_{ij}$ on $\tilde{X}_{i,-j} = \left\{\tilde{X}_{ij'};\, j' \neq j\right\}$ and

$$\begin{aligned} M &:= \begin{pmatrix} 1 & -\pi_{1,2} & \cdots & -\pi_{1,p} \\ -\pi_{2,1} & 1 & \cdots & -\pi_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_{p,1} & -\pi_{p,2} & \cdots & 1 \end{pmatrix}, \\ T^2 &:= \operatorname{diag}\left(\tau_1^2, ..., \tau_p^2\right), \end{aligned}$$

such that $\tau_j^2 := \mathbb{E}\left[\left(\tilde{X}_{ij} - \tilde{X}_{i,-j}\pi_j\right)^2\right]$ for $j = 1, ..., p$.

Simultaneous inference

From a practical viewpoint, conducting simultaneous inference for a collection of parameters in high-dimensional models may be of greater interest to researchers than inference of a single parameter. To be more specific, suppose we are interested in testing the hypothesis:

$$H_{0,G} \,:\, \beta_{0j} = \tilde{\beta}_j \;\forall j \in G \subseteq \{1, 2, ..., p\}$$

against the alternative $H_{a,G} \,:\, \beta_{0j} \neq \tilde{\beta}_j$ for some $j \in G$. In particular, we allow $|G| \geq n$. Zhang and Cheng (2017) develop a bootstrap-assisted procedure to conduct simultaneous inference in sparse linear models. Here we propose similar test statistics

$$\begin{aligned} T_G &= \max_{j \in G} \sqrt{n} \left| \hat{b}_j - \tilde{\beta}_j \right|, \\ \text{or, } T_G &= \max_{j \in G} \sqrt{n} \left| \tilde{b}_j - \tilde{\beta}_j \right|, \end{aligned}$$

and a multiplier bootstrap version

$$W_G = \max_{j \in G} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \hat{\Theta}_j \hat{X}_i^T \epsilon_i \right|,$$

where $(\epsilon_i)_{i=1}^n$ are *i.i.d.* $\mathcal{N}(0,\, 1)$ random variables. The bootstrap critical value is then given by

$$c_G(\alpha) = \inf\left\{t \in \mathbb{R} \,:\, \mathbb{P}\left(W_G \leq t | Y,\, X,\, Z\right) \geq 1 - \alpha\right\}$$

for any user-defined $\alpha \in (0,\, 1)$. In the case where the variance σ_ε^2 of ε_i in (1) is unknown, we can use

$$W_G = \max_{j \in G} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \hat{\Theta}_j \hat{X}_i^T \hat{\sigma}_\varepsilon \epsilon_i \right|,$$

where $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \left(\hat{Y}_i - X_i \hat{\beta}\right)^2}{n - \left\|\hat{\beta}\right\|_1}$ or $\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \left(Y_i - X_i \hat{\beta} - \hat{g}(Z_i)\right)^2}{n - \left\|\hat{\beta}\right\|_1}$ is an estimator for σ_ε^2.

3 Theoretical results

To make the key point of this paper, we first present the results for the case where β_0 and π_j are assumed to be exactly sparse (Theorem 1). Our additional result (Theorem A.1 in the supplement) relaxes the exact sparsity assumptions and allows relaxes the exact sparsity assumptions and allows β_0 and π_j to be approximately sparse. Note that the (exact or approximate) sparsity of π_j along with the condition $\frac{1}{\tau_j^2} \precsim 1$ implies the sparsity of $\Theta_j = \left(\left[\mathbb{E}\left(\tilde{X}_i^T \tilde{X}_i\right)\right]^{-1}\right)_j$.

We begin with the following definitions. Let

$$\begin{aligned} s_0 &:= \left|\{j : \beta_{0j} \neq 0\}\right|, \\ s_j &:= \left|\left\{j^{'} \neq j,\, 1 \leq j^{'} \leq p : \pi_{j,j'} \neq 0\right\}\right|. \end{aligned}$$

To simplify our notations, we assume that $\left\|\beta_0\right\|_1 \precsim s_0$ and $\left\|\pi_j\right\|_1 \precsim s_j$ in Theorem 1 and Corollary 1. Recall $\mathcal{F}_j$ in (7); for notation simplicity, we assume $\mathcal{F}_j = \mathcal{F}$ from now on. Note that this restriction can be easily relaxed in our analysis. For any radius $\tilde{r}_n > 0$, we define the conditional *local complexity*

$$\mathcal{G}_n(\tilde{r}_n;\, \mathcal{F}) := \mathbb{E}_{\xi}\left[\sup_{f\in\Omega(\tilde{r}_n;\, \mathcal{F})}\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i f(Z_i)\right| \mid \{Z_i\}_{i=1}^{n}\right], \tag{10}$$

where ξ_is are i.i.d. zero-mean sub-Gaussian variables with parameter at most $\sigma^{\dagger}$ and $\mathbb{E}\left(\xi_i|Z_i\right) = 0$ for all $i = 1, ..., n$, and

$$\begin{aligned} \Omega(\tilde{r}_n;\, \mathcal{F}) &:= \left\{f \in \bar{\mathcal{F}} : \left\|f\right\|_n \leq \tilde{r}_n\right\}, \\ \bar{\mathcal{F}} &:= \left\{f = f^{'} - f^{''} : f^{'}, f^{''} \in \mathcal{F}\right\}. \end{aligned}$$

For any star-shaped class $\bar{\mathcal{F}}$ (that is, for any $f \in \bar{\mathcal{F}}$ and $\alpha \in [0,\, 1]$, $\alpha f \in \bar{\mathcal{F}}$), Lemma A8 in Section 4 guarantees that the function $t \mapsto \frac{\mathcal{G}_n(t;\, \mathcal{F})}{t}$ is non-increasing on the interval $(0,\, \infty)$. Therefore, there exists some large enough $\tilde{r}_n > 0$ that satisfies the *critical inequality*

$$\mathcal{G}_n\left(\tilde{r}_n;\, \mathcal{F}\right) \leq \frac{\tilde{r}_n^2}{2}; \tag{11}$$

moreover, (11) has a smallest positive solution r_n (which we will refer to as the *critical radius*). In practice, determining the exact value of this *critical radius* can be difficult; fortunately, reasonable upper bounds on r_n are often available. Here we describe two common methods from existing literature.

By a discretization argument and the Dudley's entropy integral, we may bound (10) by

$$c_0\left(\frac{\sigma^{\dagger}}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_n(t;\, \Omega(\tilde{r}_n;\, \mathcal{F}))}dt + \tilde{r}_n^2\right)$$

for some universal constant $c_0 > 0$, where $N_n(t;\, \Omega(\tilde{r}_n;\, \mathcal{F}))$ is the $t-$covering number of the set $\Omega(\tilde{r}_n;\, \mathcal{F})$. Let $\tilde{r}_n$ be a solution for

$$\frac{\sigma^{\dagger}}{\sqrt{n}}\int_0^{\tilde{r}_n}\sqrt{\log N_n(t;\, \Omega(\tilde{r}_n;\, \mathcal{F}))}dt \precsim \tilde{r}_n^2. \tag{12}$$

The resulting $\tilde{r}_n$ is known to yield an upper bound on the *critical radius* r_n for (11) (see Lemma A9 in the supplement for a formal statement); moreover, such bounds achieve sharp scaling on r_n for a wide variety of function classes (see e.g., Barlett and Mendelson, 2002; Koltchinski, 2006; Wainwright, 2015).

When $\mathcal{F}$ is a ball of radius R in the RKHS norm $\left\|\cdot\right\|_{\mathcal{H}}$, we let

$$\Omega(\tilde{r}_n;\, \mathcal{F}) := \left\{f \in \bar{\mathcal{F}} : \left\|f\right\|_n \leq \tilde{r}_n,\, \left\|f\right\|_{\mathcal{H}} \leq 1\right\}.$$

In this case, we can determine a good upper bound for r_n using the result in Mendelson (2002) who shows that

$$\mathcal{G}_n\left(\tilde{r}_n;\, \mathcal{F}\right) \precsim \sigma^{\dagger}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\min\left\{\tilde{r}_n^2,\, \tilde{\mu}_i\right\}}$$

where $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq ... \geq \tilde{\mu}_n \geq 0$ are the eigenvalues of the underlying kernel matrix for the KRR estimate. Consequently, we can solve for $\tilde{r}_{nj}$ via

$$\sigma^{\dagger}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\min\left\{\tilde{r}_n^2,\, \tilde{\mu}_i\right\}} \precsim \tilde{r}_n^2.$$

This method above is known to yield $\tilde{r}_{nj}$ with sharp scaling for various choices of kernels.

We are now ready to establish our first main result (Theorem 1), which requires a set of assumptions (in addition to those stated at the beginning of Section 1) listed in the supplementary material, due to space limitations.

Theorem 1. *Under Assumptions 1-6 in the supplementary material, if we choose* $\lambda_j \asymp \sqrt{\frac{\log p}{n}}$ *uniformly in* j *in (9), then*

$$\begin{aligned} \frac{\sqrt{n}\left(\hat{b}_j - \beta_{0j}\right)}{\hat{\sigma}_j} &\xrightarrow{D} \mathcal{N}\left(0,\, 1\right), \\ \frac{\sqrt{n}\left(\tilde{b}_j - \beta_{0j}\right)}{\hat{\sigma}_j} &\xrightarrow{D} \mathcal{N}\left(0,\, 1\right), \end{aligned}$$

where $\hat{\sigma}_j^2 = \hat{\Theta}_j\frac{\hat{X}^T\hat{X}}{n}\hat{\Theta}_j^T$, *for each* $j = 1, ..., p$.

Based on Theorem 1, Corollary 1 justifies the use of multiplier bootstrap in testing $H_{0,G}$ even when $|G|$ diverges.

Corollary 1. *Suppose Assumptions 1-4 and 6 hold while (1) in Assumption 2 and Assumption 5 are satisfied with* s_j *replaced by* $\max_{j\in G} s_j$. *Let* $\lambda_j \asymp \sqrt{\frac{\log p}{n}}$ *uniformly in* j *in (9). Assume that* $\frac{(\log pn)^7}{n} \leq C_1 n^{-c_1}$ *for some constants* $c_1, C_1 > 0$, *and there exists a sequence of positive numbers* $\alpha_n \to \infty$ *such that* $\frac{\alpha_n}{p} = o(1)$ *and* $\alpha_n\left(\log p\right)^2\max_{j=1,...,p}\lambda_j\sqrt{s_j} = o(1)$. *Then under the null* $H_{0,G}$, *for any* $G \subseteq \{1, 2, ..., p\}$, *we have*

$$\sup_{\alpha\in(0,1)}\left|\mathbb{P}\left(T_G > c_G(\alpha)\right) - \alpha\right| = o(1).$$

With Corollary 1, the power analysis of T_G then follows from Theorem 2.4 in Zhang and Cheng (2017). The above testing procedure can be easily adapted for constructing simultaneous confidence intervals and support recovery, as we will see in Sections 4.2 and 4.3.

3.1 Theoretical implication of Theorem 1

The technique where we replace X_{ij}s by the estimated partial residuals $\hat{X}_{ij} = X_{ij} - \hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$ as in (5)-(6) is called "partialling out". Note that this technique involves p nonparametric regressions where $p \geq n$. Moreover, the estimation error from each nonparametric regression accumulates in the approximate inverse $\hat{\Theta}$ of $n^{-1}\sum_{i=1}^{n}\hat{X}_i^T\hat{X}_i$. Consequently, we first discuss what makes the "partialling out" strategy work in the statistical inference of β_0 despite that p is high dimensional.

Recall from our previous discussion that $\hat{b}_j - \beta_{0j}$ and $\tilde{b}_j - \beta_{0j}$ can be decomposed into a leading term $\frac{1}{n}\Theta_j\tilde{X}^T\varepsilon$ and several remainder terms. The rates of convergence for the remainder terms that are related to the nonparametric projection step depend on $\max_{j,\,j'}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}\left[\hat{f}_{j'}(Z_i) - f_{j'}(Z_i)\right]\right|$ with $\hat{f}_j$ defined in (7). In particular, we show that

$$\max_{j,\,j'}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{X}_{ij}\left[\hat{f}_{j'}(Z_i) - f_{j'}(Z_i)\right]\right| \leq ct_n^2 \tag{13}$$

for any $t_n \geq r_n$, with probability at least $1 - \exp\left(-c^{'}nt_n^2 + c^{''}\log p\right)$, for some constants c, $c^{'}$, $c^{''} > 0$. For many popular function classes, the *critical radius* r_n defined earlier gives the optimal scaling for bounds on $\left\|\hat{f}_{j'} - f_{j'}\right\|_n$. In particular, for (7), one can show that

$$\max_{j'}\left\|\hat{f}_{j'} - f_{j'}\right\|_n \leq c^{''}t_n \tag{14}$$

for any $t_n \geq r_n$, with probability at least $1 - c_0^{'}\exp\left(-c_1^{'}nt_n^2 + c_2^{'}\log p\right)$.

Note that the orthogonality condition $\mathbb{E}\left(\tilde{X}_{ij}|Z_i\right) = 0$ (for all j) introduced by our partialling out strategy "reduces" the effects of the estimation errors from $\hat{f}_j$: The statistical error contributed by the projection step is r_n^2 instead of the optimal rate r_n that one would expect from the nonparametric regression. Given this observation, for some function $h(s_j,\, s_0)$ of s_j and s_0 only (where the exact form of h is detailed in Assumption 5), as long as

$$\sqrt{n}r_n^2 h(s_j,\, s_0) = o(1),$$

the remainder terms related to (7) in the asymptotic expansions of $\sqrt{n}\left(\hat{b}_j - \beta_{0j}\right)$ and $\sqrt{n}\left(\tilde{b}_j - \beta_{0j}\right)$ are dominated by the leading term $\frac{1}{\sqrt{n}}\Theta_j\tilde{X}^T\varepsilon$, which has an asymptotic normal distribution. Note that the above finding also holds true for the surrogate $\hat{Y}_i = Y_i - \hat{\mathbb{E}}\left(Y_i|Z_i\right)$ (which is used in (5)) and the surrogate $\hat{g}(Z_i)$ (which is used in (6)).

We illustrate the theoretical insight above with three specific examples in terms of $\dim\left(Z_i\right)$ and the function class $\mathcal{F}$ that f_js belong to. To facilitate the presentation, our following discussions only concern $\mathbb{E}\left(X_{ij}|Z_i\right)$s and $\hat{\mathbb{E}}\left(X_{ij}|Z_i\right)$s; $\mathbb{E}\left(Y_i|Z_i\right)$ and $\hat{\mathbb{E}}\left(Y_i|Z_i\right)$ can be argued in the same fashion.

Example 1: $Z_i \in \mathbb{R}$ and $\mathcal{F} \in \mathcal{S}^m$ (the mth order Sobolev ball). Estimating $\mathbb{E}\left(X_{ij}|Z_i\right)$s via (7) or (8) can be reduced to the smoothing spline procedure, which achieves the sharp rate, $n^{-\frac{2m}{2m+1}}$, on r_n^2. In this case, we require $\sqrt{n}n^{-\frac{2m}{2m+1}}h(s_j,\, s_0) = o(1)$.

Example 2: $Z_i \in \mathbb{R}$ and $\mathcal{F}$ is the class of linear combinations of bounded basis functions $\psi_l(\cdot)$s such that for $f \in \mathcal{F}$, $f(Z_i) = \sum_{l=1}^{d_1}\theta_l\psi_l(Z_i)$ and $\theta := (\theta_l)_{l=1}^{d_1}$ belongs to the l_0-"ball" of "radius" k. Suppose $d_1 \geq n$ and $d_1 \geq 4k$. Then the standard Lasso procedure would yield upper bounds with scaling $\frac{k\log d_1}{n}$ on the quantities in (13). The scaling $\frac{k\log d_1}{n}$ almost achieves the sharp rate, $\frac{k\log\frac{d_1}{k}}{n}$, on r_n^2. In this case, we require $\frac{k\log d_1}{\sqrt{n}}h(s_j,\, s_0) = o(1)$. If we use the recently proposed Slope (Su and Candés, 2016) instead of the standard Lasso, then the scaling $\frac{k\log\frac{d_1}{k}}{n}$ can be attained. In this case, we require $\frac{k\log\frac{d_1}{k}}{\sqrt{n}}h(s_j,\, s_0) = o(1)$.

Example 3: $Z_i \in \mathbb{R}^d$ and $\mathcal{F}$ is the class of $|S| := k$ sparse additive nonparametric functions in the sense that any member f in $\mathcal{F}$ has the following decomposition form $f(Z_i) = \sum_{l=1}^{d}f_l(Z_{il}) = \sum_{l\in S}f_l(Z_{il})$; moreover, f_l belongs to an RKHS of univariate functions. We may then apply program (7) in Raskutti, et al. (2012) to estimate $\mathbb{E}\left(Y_i|Z_i\right)$ and $\mathbb{E}\left(X_{ij}|Z_i\right)$s. If the underlying RKHS is $\mathcal{G}^m$, we would require $\sqrt{n}k\left(n^{-\frac{2m}{2m+1}} \vee \frac{\log d}{n}\right)h(s_j,\, s_0) = o(1)$.

4 Experiments

In this section, we evaluate the performance of our methods with simulation studies and a real data example.

For the simulation studies, to generate the full covariates X, we first generate X_0 from the $p-$dimensional normal distribution with mean 0 and variance $\Sigma_{X_0} = (\Sigma_{X_0,ij})_{i,j=1}^p$, which takes three different forms:

(S1) Independent: $\Sigma_{X_0} = I_p$;

(S2) AR(1): $\Sigma_{X_0,ij} = 0.5^{|i-j|}$;

(S3) Exchangeable/Compound Symmetric: $\Sigma_{X_0,ii} = 1$ and $\Sigma_{X_0,ij} = 0.5$ if $i \neq j$.

The covariates $\{Z_i\}_{i=1}^n$ are *i.i.d.* from $U[0, 2]$. To incorporate the dependence between X and Z, we set $X_{i1} = X_{0,i1}+3Z_i, X_{i2} = X_{0,i2}+3Z_i^2, X_{i3} = X_{0,i3}-3Z_i$ and $X_{ij} = X_{0,ij}, 1 \le i \le n, 4 \le j \le p$. The set of nonzero coefficients in β_0 is from a fixed realization of $s_0 = 3$ *i.i.d.* $U[0, 3]$. The active set is set to be $S_0 = \{1, 2, 3\}$. We consider two different non-linear functions g_0:

(G1) $g_0(z) = 1.5\sin(2\pi z)$;

(G2) $g_0(z) = z^{10}(1 - z)^4/B(11, 5) + 4z^4(1 - z)^{10}/B(5, 11)$

where $B(\cdot\,, \cdot)$ denotes the beta distribution. The error terms are generated from a standard normal distribution. We estimate $\hat{X}$ by regressing X on Z with the smoothing spline method and $(\hat{\beta}, \hat{g})$ are obtained from the procedure proposed in Müller and van de Geer (2015). As a result, the debiased estimator in our simulations concerns (6). We do not test the performance of (5) with simulation experiments but expect it to behave similarly as (6). Similar to Zhang and Cheng (2017), the estimated variance $\hat{\sigma}_\varepsilon^2$ is calculated as follows:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \left(Y_i - X_i\hat{\beta} - \hat{g}(Z_i)\right)^2}{n - \left\|\hat{\beta}\right\|_1}.$$

We set the tuning parameter $\mu = n^{-2/5}/10$ (Müller and van de Geer, 2015) and let λ and λ_j $(1 \le j \le p)$ be calculated from the 10-fold cross validation (van de Geer, et al., 2014). Across all the simulations, we set the sample size $n = 100$ and the number of variables $p = 500$. Results in sections 4.1 and 4.2 are based on 100 replications, while those in section 4.3 are based on 500 replications. Results in section 4.4 is based on real data analysis.

4.1 Component-wise confidence interval

Average coverage and average length of the intervals for individual coefficients corresponding to variables in either S_0 or S_0^c are considered. Denote CI_j as a two-sided confidence interval for β_j^0. In Table 1, we report the empirical versions of

Avgcov $S_0 = s_0^{-1}\sum_{j\in S_0}\mathbb{P}(\beta_{0j} \in \mathrm{CI}_j)$;
Avglength $S_0 = s_0^{-1}\sum_{j\in S_0}\mathrm{length}(\mathrm{CI}_j)$;
Avgcov $S_0^c = (p - s_0)^{-1}\sum_{j\in S_0^c}\mathbb{P}(0 \in \mathrm{CI}_j)$;
Avglength $S_0^c = (p - s_0)^{-1}\sum_{j\in S_0^c}\mathrm{length}(\mathrm{CI}_j)$.

The results in Table 1 agree with our theoretical predictions. The average coverage probabilities of confidence intervals for S_0^c are close to the nominal 95% level, while those for S_0 are slightly lower than 95%. The confidence intervals for S_0^c are comparably narrower than those for S_0. We also notice that as the columns in X_0 become more correlated (so the inverse Θ of the Hessian becomes less sparse), the coverage performance becomes worse. This finding confirms our earlier comment (in Section 2) that the sparsity condition on the off diagonal elements of Θ plays a crucial role in the effectiveness of the debiased approach as this condition makes remainder terms like $\left(\hat{\Theta}_j - \Theta_j\right)\frac{1}{\sqrt{n}}\tilde{X}^T\varepsilon$ small in the asymptotic expansion of $\sqrt{n}\left(\tilde{b}_j - \beta_{0j}\right)$.

4.2 Simultaneous confidence intervals

In Table 2, we present the coverage probabilities and interval widths for the simultaneous confidence intervals for $\beta_{0j}, 1 \le j \le p$. For each simulation run, we record whether the simultaneous confidence interval contains β_{0j} for $1 \le j \le p$ and the corresponding interval width. Again, it is not surprising that the coverage probability is affected by the amount of correlations between the columns in X_0. Overall, both studentized and non-studentized method provide satisfactory coverage probability. When Σ_{X_0} is the identity matrix, non-studentized method has better coverage; while when Σ_{X_0} takes the form of S2 or S3, the performance of the studentized method is better.

4.3 Support recovery

The major goal of this section is to identify signal locations of β_0 in a pre-specified set $G = \{1, 2, \ldots, p\}$, i.e. support recovery. Similarly as the procedure in Zhang and Cheng (2017), we take the signal set

$$\hat{\mathcal{S}}_0 = \{j \in \tilde{G} : |\tilde{b}_j| > \lambda_j^*\},$$

where $\lambda_j^* = \sqrt{2\hat{\omega}_{jj}\log(p)/n}$ and $\hat{\omega}_{jj} = \hat{\sigma}_\varepsilon^2\hat{\Theta}_j\frac{\hat{X}^T\hat{X}}{n}\hat{\Theta}_j^T$. Note that similar arguments as Proposition 3.1 of Zhang and Cheng (2017) implies this support recovery procedure is consistent. To assess the performance, we consider the following similarity measure

$$d(\hat{\mathcal{S}}_0, \mathcal{S}_0) = \frac{|\hat{\mathcal{S}}_0 \cap \mathcal{S}_0|}{\sqrt{|\hat{\mathcal{S}}_0| \cdot |\mathcal{S}_0|}}.$$

Table 3 summarizes the mean and standard deviation of $d(\hat{\mathcal{S}}_0, \mathcal{S}_0)$ as well as the number of false positives (FP) and false negatives (FN) normalized by

Table 1: Average coverage probabilities and lengths of confidence intervals at the 95% nominal level with 100 iterations; $n = 100, p = 500$; (i) Avgcov S_0, (ii) Avglength S_0, (iii), Avgcov S_0^c, (iv) Avglength S_0^c

	Active set $S_0 = \{1, 2, 3\}$; Error $\varepsilon \sim N(0, 1)$					
	S1 ,G1	S2, G1	S3, G1	S1, G2	S2, G2	S3, G2
i	0.896	0.857	0.693	0.887	0.823	0.683
ii	0.802	0.812	0.807	0.798	0.789	0.827
iii	0.953	0.955	0.963	0.953	0.955	0.963
iv	0.476	0.510	0.547	0.480	0.500	0.559

$\sqrt{|\hat{\mathcal{S}}_0| \cdot |\mathcal{S}_0|}$. When the amount of correlations between the columns in X_0 increases (as in S3), the false positive rates are comparably higher.

4.4 Real data set

In this section, we illustrate our method with the data from [2]. This data set contains the wage information of 534 workers and their years of experience, education, living region, gender, race, occupation and marriage status. This data set was studied in [25] that also concerns high dimensional partial linear models. We consider the following model:

$$Y_i = \sum_{j=1}^{14} X_{ij}\beta_{0j} + g_0(Z_i) + \varepsilon_i, \; i = 1, ..., 534, \qquad (15)$$

where Y_i is the ith worker's wage, Z_i is his/her year of experience, X_{ij}s are additional covariates and ε_is are i.i.d. errors. We exhibit brief descriptions of these covariates, their estimates and standard errors in Table 4. In view of Table 3 in [25], our estimates are closer to those from the unpenalized PLM. Nevertheless, the magnitude (in absolute value) and statistical significance of the estimate associated with "Sales" from our method are substantially larger than those from the unpenalized PLM. In addition, the signs of the estimates associated with "Hispanic", "Married", and "Clerical" are opposite to those from unpenalized PLM (although both the unpenalized PLM and our method suggest that these variables are far from being statistically significant). Our support recovery includes education, gender, union member, management and professional, which give a more refined subset of variables, compared to those from PLM-SCAD and PLM-LASSO methods in [25].

Table 2: Coverage probabilities and interval widths for the simultaneous confidence intervals based on the non-studentized (NST) and studentized (ST) test statistics with 100 iterations; $n = 100, p = 500$; (i) NST coverage, (ii) NST width , (iii) ST coverage, (iv) ST width

	Active set $S_0 = \{1, 2, 3\}$; Error $\varepsilon \sim N(0, 1)$					
	S1 ,G1	S2, G1	S3, G1	S1, G2	S2, G2	S3, G2
i	0.95	0.74	0.72	0.96	0.86	0.74
ii	1.05	1.09	1.12	1.06	1.06	1.12
iii	0.88	0.94	0.87	0.82	0.91	0.83
iv	0.88	0.96	1.04	0.87	0.93	1.04

Table 3: The mean and standard deviation (SD) of $d(\hat{\mathcal{S}}_0, \mathcal{S}_0)$, and the numbers of false positives (FP) and false negatives (FN) with 500 iterations; $n = 100, p = 500$; (i) Mean, (ii) SD, (iii) FP, (iv) FN

	Active set $S_0 = \{1, 2, 3\}$; Error $\varepsilon \sim N(0, 1)$					
	S1 ,G1	S2, G1	S3, G1	S1, G2	S2, G2	S3, G2
i	0.96	0.97	0.94	0.97	0.97	0.94
ii	0.07	0.06	0.08	0.07	0.06	0.08
iii	0.09	0.06	0.13	0.09	0.07	0.12
iv	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Real data example

Variable	Description	$\hat{b}_j(SE)$
X_1	Number of years of eduction	0.638(0.095)
X_2	1 = Southern region, 0 = other	-0.569(0.433)
X_3	1 = Female, 0 = male	-1.992(0.422)
X_4	1 = Union member, 0 = nonmember	1.379(0.525)
X_5	1 = White, 0 = other	0.726(0.582)
X_6	1 = Hispanic, 0 = other	0.365(0.994)
X_7	1 = Management, 0 = other	2.956(0.758)
X_8	1 = Sales, 0 = other	-1.058(0.823)
X_9	1 = Clerical, 0 = other	-0.135(0.663)
X_{10}	1 = Service, 0 = other	-0.702(0.659)
X_{11}	1 = Professional, 0 = other	1.762(0.694)
X_{12}	1 = Manufacturing, 0 = other	1.086(0.544)
X_{13}	1 = Construction, 0 = other	0.511(0.958)
X_{14}	1 = Married, 0 = other	0.103(0.408)

References

[1] Bartlett, P. and Mendelson, S. (2002). Gaussian and Rademacher complexities: Risk bounds and structural results. J. Mach. Learn. Res. 3: 463-482.

[2] Berndt, E. R. (1991). The Practice of Econometrics: Classical and Contemporary. Addison-Wesley, Reading, MA.

[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). Efficient and Adaptive Estimation for Semiparametric Models. Springer-Verlag, New York.

[4] Cheng G., Zhang H. and Shang Z. (2015). Sparse and efficient estimation for partial spline models with increasing dimension. Annals of the Institute of Statistical Mathematics. 67: 93-127.

[5] Donald, S. G., and Newey, W. (1994). Series estimation of semilinear models. Journal of Multivariate Analysis. 50: 30-40.

[6] Härdle, W., Liang, H., and Gao, J. (2000). Partially Linear Models. Physica-Verlag, Heidelberg.

[7] Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15: 2869-2909.

[8] Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Statist. 34: 2593-2656.

[9] Li, R., and Liang, H. (2008). Variable selection in semiparametric regression modeling. Ann. Statist. 36: 261-286.

[10] Liang, H., and Li R. (2009). Variable selection for partially linear models with measurement errors. Journal of the American Statistical Association. 104: 234-248.

[11] Loh, P., and Wainwright, M. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. Ann. Statist. 40: 1637-1664.

[12] Mendelson, S. (2002). Geometric parameters of kernel machines. In Proceedings of COLT. 29-43.

[13] Müller, P. and van de Geer, S. (2015). The partial linear model in high dimensions. Scandinavian Journal of Statistics. 42: 580-608.

[14] Negahban, S., P. Ravikumar, M. J. Wainwright, and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statistical Science. 27: 538-557. 2010 version: arXiv:1010.2731v1.

[15] Raskutti, G., Wainwright, J. M., and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. J. Mach. Learn. Res. 13: 389-427.

[16] Robinson, P. M. (1988). Root-n-consistent semiparametric regression. Econometrica. 56: 932-954.

[17] Sherwood, B. and Wang L. (2016). Partially linear additive quantile regression in ultra-high dimension. Ann. Statist. 44: 288-317.

[18] Su, W. and Candès E. (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. Ann. Statist. 44: 1038-1068.

[19] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B. 58: 267-288.

[20] van de Geer, S. (2000). Empirical Processes in M-Estimation. Cambridge University Press.

[21] van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42: 1166-1202.

[22] van der Vaart, A. W. and J. Wellner (1996). Weak Convergence and Empirical Processes. Springer-Verlag, New York, NY.

[23] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices, in Eldar, Y. and G. Kutyniok, Eds, Compressed Sensing: Theory and Applications. 210-268, Cambridge.

[24] Wainwright, J. M. (2015). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. In preparation. University of California, Berkeley.

[25] Xie, H. L. and Huang J. (2009). Scad-penalized regression in high-dimensional partially linear models. Ann. Statist. 37: 673-696.

[26] Yu, Z., Levine, M. and Cheng, G. (2017). Minimax optimal estimation in high dimensional semiparametric models. arXiv preprint.

[27] Zhang, C. H. and Zhang S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc. Ser. B. Stat. Methodol. 76: 217-242.

[28] Zhang, X. and Cheng, G. (2017). Simultaneous inference for high-dimensional linear models. Forthcoming in Journal of the American Statistical Association - Theory & Methods. 112, 757-768.

[29] Zhu, Y. (2017). Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients. Forthcoming in Ann. Statist. 45, 2274-2298.