The Annals of Statistics
2017, Vol. 45, No. 5, 2274–2298
DOI: 10.1214/16-AOS1528

NONASYMPTOTIC ANALYSIS OF SEMIPARAMETRIC REGRESSION MODELS WITH HIGH-DIMENSIONAL PARAMETRIC COEFFICIENTS[1]

BY YING ZHU

Michigan State University

We consider a two-step projection based Lasso procedure for estimating a partially linear regression model where the number of coefficients in the linear component can exceed the sample size and these coefficients belong to the l_q-"balls" for $q \in [0, 1]$. Our theoretical results regarding the properties of the estimators are nonasymptotic. In particular, we establish a new nonasymptotic "oracle" result: Although the error of the nonparametric projection *per se* (with respect to the prediction norm) has the scaling t_n in the first step, it only contributes a scaling t_n^2 in the l_2-error of the second-step estimator for the linear coefficients. This new "oracle" result holds for a large family of nonparametric least squares procedures and regularized nonparametric least squares procedures for the first-step estimation and the driver behind it lies in the projection strategy. We specialize our analysis to the estimation of a semiparametric sample selection model and provide a simple method with theoretical guarantees for choosing the regularization parameter in practice.

1. Introduction. Semiparametric regression modeling plays an important role in the statistics and econometric literature as it retains the flexibility of nonparametric models while avoiding the "curse of dimensionality." Härdle, Liang and Gao (2000), Ruppert, Wand and Carroll (2003) and Yatchew (2003) provide detailed discussions on a variety of semiparametric regression models. The leading example of semiparametric regression models is the partially linear regression introduced by Engle et al. (1986):

$$Y_i = X_i\beta^* + g(\gamma_i^*) + \eta_i, \tag{1}$$

where $\{\eta_i\}_{i=1}^n$ is a zero-mean unobserved random vector, $\{Y_i, X_i, \gamma_i^*\}_{i=1}^n$ are observed data, and $\mathbb{E}[\eta_i|\gamma_i^*, X_i] = 0$. In this paper, we focus on univariate γ_i^* only; $g(\cdot)$ is an unknown function of the single variable γ_i^*; β^* is a p-dimensional vector of unknown coefficients of main interests and its jth component is denoted by β_j^*. There is a rich literature on estimations of (1) where the dimensions of β^* is fixed or small relative to n and β^* is exactly sparse (i.e., it belongs to the l_0-"ball");

Received October 2015; revised November 2016.

[1]Supported by LAMS Research Fellowship and Dean's Research Fellowship at U.C. Berkeley.

MSC2010 subject classifications. Primary 62J02; secondary 62N01, 62N02, 62G08, 62J12.

Key words and phrases. High-dimensional statistics, Lasso, nonasymptotic analysis, partially linear models, sample selection.

see, for example, Chen (1988), Robinson (1988), Donald and Newey (1994), Yu and Ruppert (2002). Some variable selection procedures are also available: Bunea (2004) and Bunea and Wegkamp (2004) develop model selection criteria for (1) with i.i.d. data while Fan and Li (2004) propose a profile least squares procedure with longitudinal data; Liang and Li (2009) develop penalized least squares and penalized quantile regression for (1) with i.i.d. data where the covariates are contaminated with additive errors. All these aforementioned papers focus on the case where p is small relative to n and β^* belongs to the l_0-"ball."

Many important statistical models can be in fact transformed into (1). One instance concerns the sample selection model in econometrics:

$$Y_{1i} = 1\{W_i\theta^* + \epsilon_{1i} > 0\}, \tag{2}$$

$$Y_i = X_i\beta^* + \epsilon_{2i} \qquad \text{whenever } Y_{1i} = 1, \tag{3}$$

$$\mathbb{E}(\epsilon_{2i}|W_i, X_i, Y_{1i} = 1) = g(W_i\theta^*), \tag{4}$$

where β^* is a p-dimensional vector of unknown coefficients of interests and θ^* is a d-dimensional vector of unknown nuisance parameters; $g(\cdot)$, referred to as the "selection bias," is an unknown function of the single index $W_i\theta^*$; ϵ_{1i}s and ϵ_{2i}s are zero-mean unobserved random errors; W_is, X_is, Y_{1i}s and Y_is are observed data. For example, Y_{1i} indicates whether individual i worked or not whereas Y_i is actual hours i worked if $Y_{1i} = 1$. In social science, observational studies based on selected samples (i.e., units with $Y_{1i} = 1$) like the previous example are common. When a sample, intentionally or unintentionally, is based in part on values taken by a dependent variable, parameter estimates without corrective measures may be inconsistent. Selection may result from self-selection, where the outcome of interest is determined in part by the individual choice of whether or not to participate in the activity of interest. It can also result from endogenous stratification, where those who participate in the activity of interest are deliberately oversampled with an extreme case being sampling only participants.

In the classical low-dimensional sample selection models, parameter estimates obtained from OLS on (3) may be inconsistent unless corrective measures are taken [see, e.g., Gronau (1973); Heckman (1976); Ahn and Powell (1993); Newey (2009)]. Equation (4) is known as the "single-index" restriction used in Ahn and Powell (1993) and Newey (2009). In particular, it is implied by independence of the errors $(\epsilon_{1i}, \epsilon_{2i})$ and the covariates (W_i, X_i). Note that equations (2)–(4) imply (1) with $\gamma_i^* = W_i\theta^*$ when $Y_{1i} = 1$, and by construction, $\mathbb{E}[\eta_i|W_i, X_i, Y_{1i} = 1] = 0$. Consequently, a consistent estimator of β^* based on (1) and the subsample with $Y_{1i} = 1$ is not contaminated by the sample selectivity bias.

Given an estimate $\hat{\theta}$ of θ^*, a natural way to estimate the linear coefficients β^* consists of two steps. The first step (projection) uses nonparametric regression to obtain the partial residuals, a commonly used idea in the estimation of partially linear models [see, e.g., Robinson (1988); Donald and Newey (1994); Liang and

Li (2009)]. The second step is an l_1-penalized least squares estimator (the Lasso) [Tibshirani (1996)] using the partial residuals from the first step. Upon the availability of the estimate $\hat{\beta}$ for β^*, an estimator for the nonparametric component $g(\cdot)$ can be obtained by performing a nonparametric regression where β^* is surrogated with $\hat{\beta}$. This projection-based Lasso procedure is intuitive and can be easily implemented using built-in routines in standard software packages (e.g., *matlab*, *R*). It decomposes the joint estimation of the high-dimensional linear coefficients and the nonparametric component into sequential estimations with each searching over a much smaller parameter space. The first-step estimation is easily paralleable as it involves solving $p + 1$ independent subproblems and each subproblem can be in general solved with an efficient algorithm.

In this paper, we study the aforementioned procedures by allowing the dimension of β^*, p, to exceed n; moreover, β^* belongs to the l_q-"balls" for $q \in [0, 1]$. Throughout this paper, we will refer to β^* as the high-dimensional linear coefficients and g as the nonparametric component. Our analysis differs from the existing approaches in the following aspects. First, the theoretical guarantees in this paper are nonasymptotic in nature, while the results in previous literature such as Chen (1988), Robinson (1988), Donald and Newey (1994), Yu and Ruppert (2002), Liang and Li (2009) and Liang et al. (2010) are asymptotic. Complementary to the asymptotic "oracle" results from existing literature [e.g., Chen (1988); Liang and Li (2009); Zhu, Dong and Li (2013)] which says the coefficients in the linear component can be estimated asymptotically as well as if the unknown nonparametric component were known, we derive a new nonasymptotic "oracle" result: if the error of the nonparametric projection *per se* (with respect to the prediction norm) has the scaling t_n in the first step of our procedure, it only contributes a scaling t_n^2 in the l_2-error of the second-step estimator for β^*, where t_n is related to the "critical radius" for the local complexity of a function class—a notion used in nonparametric literature [e.g., van der Vaart and Wellner (1996); van de Geer (2000); Bartlett and Mendelson (2002); Koltchinskii (2006); Wainwright (2015)]. This new "oracle" result is derived for a large family of nonparametric least squares estimators and regularized nonparametric least squares estimators for the first step. The driver behind this "oracle" result lies in the projection strategy.

Second, our analysis allows β^* to belong to the l_q-"balls" for $q \in [0, 1]$, which covers a spectrum of sparsity cases (exact and approximate), while the analyses in previous literature such as Bunea (2004), Bunea and Wegkamp (2004), Liang and Li (2009) and Liang et al. (2010) focus on the case where β^* is exactly sparse (i.e., it belongs to the l_0-"ball"). This distinction together with the use of l_1-penalty in this paper makes our analysis a novel contribution to the existing literature on semiparametric regression. We further specialize our nonasymptotic analysis to the estimation of a leading case of the semiparametric model (2)–(4) where the coefficient vector θ^* in (2) has a dimension (d) which can also exceed n, and θ^* belongs to the l_q-"balls" for $q \in [0, 1]$. While allowing (4) to be nonparametric, we focus on the case where ϵ_{1i} in (2) has a standard normal distribution given that this

model is the most widely applied for studying sample selectivity in social science [e.g., Wooldridge (2002)]. To the best of our knowledge, these results are also new to econometrics.

In our problem with $p \geq n$ or even $p \gg n$, the $p \times p$ random matrix $\frac{\hat{v}^T \hat{v}}{n}$ has rank at most n [where $\hat{v}_j$ is the estimate of the true partial residuals $v_j = (X_{ij} - \mathbb{E}(X_{ij}|\gamma_i^*))_{i=1}^n$ for $j = 1, \ldots, p$]. This singularity issue plus the estimation errors in $(\hat{v}_j)_{j=1}^p$ from the p nonparametric regressions in the first step pose a substantial challenge in the nonasymptotic analysis of our two-step procedure. In the standard high-dimensional sparse linear regression where the covariates are perfectly observed and absent from estimation errors, lower restricted eigenvalue (LRE) conditions [see, e.g., Bickel, Ritov and Tsybakov (2009); Meinshausen and Yu (2009); Bühlmann and van de Geer (2011); Negahban et al. (2012)] are often used to establish the error bounds. However, in our setting, guarantees for the random matrix $\frac{\hat{v}^T \hat{v}}{n}$ to satisfy these existing high-level assumptions are not automatic and require careful verifications. We provide detailed analysis which shows that an LRE condition is satisfied by $\frac{1}{n}\hat{v}^T \hat{v}$ with high probability when it is satisfied by the population matrix $\mathbb{E}[v_i^T v_i]$.

Section 2 describes the estimation procedure and Section 3 provides preliminaries to the identification assumptions on the models of our interests. General upper bounds on the l_2-errors of the estimators for the high-dimensional linear coefficients in the standard partially linear model (1) are established in Section 4. Section 5.1 presents nonasymptotic bounds for the parametric component and nonparametric component, respectively, along with a result on variable selection concerning a leading case of the semiparametric model (2)–(4) when it is estimated by the procedure in Section 2. For this estimator, we also provide a practical method with theoretical guarantees for choosing the regularization parameter and evaluate it with Monte-Carlo simulations in Section 5.2. Section 6 discusses future directions of this research. Proofs of the main results are collected in Appendix A, with the remaining technical lemmas and proofs contained in Appendix S. Both Appendices are included in the supplementary materials [Zhu (2017)].

Notation. The l_q-norm of a p-dimensional vector Δ is denoted by $|\Delta|_q$, $1 \leq q \leq \infty$ where $|\Delta|_q := (\sum_{j=1}^p |\Delta_j|^q)^{1/q}$ when $1 \leq q < \infty$ and $|\Delta|_q := \max_{j=1,\ldots,p} |\Delta_j|$ when $q = \infty$. For a matrix $A \in \mathbb{R}^{p_1 \times p}$, write $|A|_\infty := \max_{i,j} |a_{ij}|$ to be the elementwise l_∞-norm of A. The l_2-operator norm, or spectral norm of the matrix A corresponds to its maximum singular value, defined as $\|A\|_2 := \sup_{\Delta \in S} |A\Delta|_2$, where $S = \{\Delta \in \mathbb{R}^p | |\Delta|_2 = 1\}$. The l_∞-matrix norm (maximum absolute row sum) of A is denoted by $\|A\|_\infty := \max_i \sum_j |A_{ij}|$ (note the difference between $|A|_\infty$ and $\|A\|_\infty$). For a square matrix $A \in \mathbb{R}^{p \times p}$, denote its minimum eigenvalue by $\lambda_{\min}(A)$. For a vector $\Delta \in \mathbb{R}^p$, let $J(\Delta) = \{j \in \{1, \ldots, p\} | \Delta_j \neq 0\}$ be its support, that is, the set of indices corresponding to its nonzero components Δ_j. The cardinality of a set $J \subseteq \{1, \ldots, p\}$ is denoted by $|J|$. Define $\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ that places a weight $\frac{1}{n}$ on each observation X_i for $i = 1, \ldots, n$,

and the associated $\mathcal{L}^2(\mathbb{P}_n)$-norm of the vector $\Delta := \{\Delta(X_i)\}_{i=1}^n$, denoted by $|\Delta|_n$, is given by $[\frac{1}{n}\sum_{i=1}^n(\Delta(X_i))^2]^{\frac{1}{2}}$. For functions $f(n)$ and $g(n)$, write $f(n) \gtrsim g(n)$ 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 $f(n) \gtrsim g(n)$ and $f(n) \precsim g(n)$ hold simultaneously. Also denote $\max\{a, b\}$ by $a \vee b$ and $\min\{a, b\}$ by $a \wedge b$.

2. Estimation procedures. We first present the general framework for estimating the high-dimensional linear coefficients β^* and the nonparametric component $g(\cdot)$ in (1). We then provide a specific example of the general procedure for estimating a leading case of the semiparametric model (2)–(4).

The two-step estimator of the high-dimensional linear coefficients. First-step estimation. To simplify the notation, we write, for $j = 1, \ldots, p$, $\mathbb{E}(X_{ij}|\gamma_i) := m_j(\gamma_i)$, $\hat{\mathbb{E}}(X_{ij}|\gamma_i) := \hat{m}_j(\gamma_i)$, $\mathbb{E}(Y_i|\gamma_i) := m_0(\gamma_i)$, and $\hat{\mathbb{E}}(Y_i|\gamma_i) := \hat{m}_0(\gamma_i)$. To estimate model (2)–(4) via (1), we assume that a surrogate $\hat{\gamma}_i = W_i\hat{\theta}$ for $\gamma_i^* = W_i\theta^*$ is available for now. For the standard partially linear model (1) when γ_i^* is an observed variable, simply define $\hat{\gamma}_i := \gamma_i^*$. For an estimator of $m_j(\gamma_i^*)$, we focus on the following least squares estimators or the regularized least squares estimators:

$$\hat{m}_j \in \arg\min_{\tilde{m}_j \in \mathcal{F}_j} \left\{ \frac{1}{2n} \sum_{i=1}^n \big(z_{ij} - \tilde{m}_j(\hat{\gamma}_i)\big)^2 \right\}, \tag{5}$$

$$\hat{m}_j \in \arg\min_{\tilde{m}_j \in \mathcal{F}_j} \left\{ \frac{1}{2n} \sum_{i=1}^n \big(z_{ij} - \tilde{m}_j(\hat{\gamma}_i)\big)^2 + \lambda_{nj,2}|\tilde{m}_j|_{\mathcal{F}_j}^2 \right\}, \tag{6}$$

where $|\cdot|_{\mathcal{F}_j}$ is a norm associated with the function class $\mathcal{F}_j$ and $\lambda_{nj,2} > 0$ is a regularization parameter and $z_{i0} = y_i$ and $z_{ij} = x_{ij}$ for each $j = 1, \ldots, p$. For the partially linear model (1) with observed γ_i^*, we want to point out that the procedures above will not work when γ_i^* has a dimension that is large relative to n unless $\mathcal{F}_j$ has an additive decomposition structure.

A nonparametric regression problem like (5)–(6) is a standard setup in many modern statistics books [e.g., van de Geer (2000); Wainwright (2015)]. Examples of (5) include the linear regression, sparse linear regressions, convex regressions and Lipschitz Isotonic regressions. Examples of (6) include sieves-based estimators and kernel ridge regressions (KRR). In this paper, we let $\mathcal{F}_j$ in (6) be a reproducing kernel Hilbert space equipped with the norm $|\cdot|_{\mathcal{F}_j}$ and the solution to (6) is known as the KRR estimate. By Lagrangian duality, for a properly chosen radius $\bar{R}_j > 0$, the minimization problem in (6) can be reformulated as

$$\min_{\tilde{m}_j \in \mathcal{F}_j} \frac{1}{n} \sum_{i=1}^n \big(z_{ij} - \tilde{m}_j(\hat{\gamma}_i)\big)^2 \qquad \text{such that } |\tilde{m}_j|_{\mathcal{F}_j} \leq \bar{R}_j. \tag{7}$$

Second-step estimation. An estimator of the high-dimensional linear coefficients β^* can be obtained by performing the following Lasso program:

$$\hat{\beta} \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} : \frac{1}{2n} |\hat{v}_0 - \hat{v}\beta|_2^2 + \lambda_{n,3} \sum_{j=1}^{p} \hat{\sigma}_{v_j} |\beta_j|, \tag{8}$$

where $\lambda_{n,3} \geq 0$ is some regularization parameter, $\hat{v} = (\hat{v}_1, \ldots, \hat{v}_p)$ with $\hat{v}_j = (X_{ij} - \hat{m}_j(\hat{\gamma}_i))_{i=1}^n$, and $\hat{v}_0 = (Y_i - \hat{m}_0(\hat{\gamma}_i))_{i=1}^n$. For simplicity, we assume $\hat{v}_j$ is normalized such that $\hat{\sigma}_{v_j} := \sqrt{\frac{1}{n}\sum_{i=1}^n \hat{v}_{ij}^2} = 1$ for each $j = 1, \ldots, p$ in our analysis. When performing the two-step procedure above to estimate (2)–(4) via (1), program (5) or (6) and program (8) only apply to the observations with $Y_{1i} = 1$; that is, $n = \sum_{i=1}^{n_0} Y_{1i}$ where n_0 is the number of the full sample observations for (2).

Estimator of the nonparametric component. Given the estimate $\hat{\beta}$ of β^*, an estimator for $g(\cdot)$ can be either (9) or (10) as below:

$$\hat{g} \in \arg\min_{\tilde{g} \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - x_i\hat{\beta} - \tilde{g}(\hat{\gamma}_i)\big)^2, \tag{9}$$

$$\hat{g} \in \arg\min_{\tilde{g} \in \mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - x_i\hat{\beta} - \tilde{g}(\hat{\gamma}_i)\big)^2 + \lambda_n^* |\tilde{g}|_{\mathcal{F}}^2 \right\}, \tag{10}$$

where $\lambda_n^* > 0$ is a regularization parameter. As in (5)–(6), we define $\hat{\gamma}_i := \gamma_i^*$ for the standard partially linear model (1) when γ_i^* is an observed variable.

An example. We specialize our general procedure above to estimate a leading case of the semiparametric model (2)–(4), where ϵ_{1i} in (2) has a standard normal distribution (binary probit model), and for every $j = 0, \ldots, p$, $m_j(\cdot)$ is assumed to belong to the class of Lipschitz functions. For an estimator of θ^*, we consider the *l_1-regularized conditional maximum likelihood estimator*:

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ -\frac{1}{n_0} \sum_{i=1}^{n_0} y_{1i}\phi_1(w_i\theta) + \frac{1}{n_0} \sum_{i=1}^{n_0} \phi_2(w_i\theta) + \lambda_{n_0,1}|\theta|_1 \right\}, \tag{11}$$

where $\lambda_{n_0,1} \geq 0$ is a regularization parameter, $\phi_1(w_i\theta) = \log \frac{\Phi(w_i\theta)}{1-\Phi(w_i\theta)}$, and $\phi_2(w_i\theta) = -\log[1 - \Phi(w_i\theta)]$ [$\Phi(\cdot)$ is the standard normal c.d.f.].

We now provide an estimator based on Lipschitz regression for the first step (nonparametric projection). Note that a function $f : [a, b] \to \mathbb{R}$ is *L-Lipschitz* if

$$\big|f(t) - f(t')\big| \leq L|t - t'| \tag{12}$$

for all $t, t' \in [a, b]$. When m_j satisfies the L-Lipschitz assumption, we can restrict $\tilde{m}_j$ in (5) to be in the class of L-Lipschitz functions. As a result, program (5) can be converted to an equivalent finite-dimensional problem by applying the constraint

(12) to each of the sampled points $\hat{\gamma}_i = w_i\hat{\theta}$ so that there must exist a real-valued vector $(\tilde{z}_{ij})_{i=1}^n$ which satisfies (14) below:

$$(\hat{z}_{ij})_{i=1}^n \in \arg\min_{(\tilde{z}_{ij})_{i=1}^n} \left\{ \frac{1}{2n}\sum_{i=1}^n (z_{ij} - \tilde{z}_{ij})^2 \right\}$$
$$\text{s.t.} \quad |\tilde{z}_{ij} - \tilde{z}_{i'j}| \le L|\hat{\gamma}_i - \hat{\gamma}_{i'}| \qquad \text{for all } i < i', i, i' = 1, \ldots, n. \tag{13}$$

Given an optimal solution $(\hat{z}_{ij})_{i=1}^n$, a Lipschitz function $\hat{m}_j$ can be constructed by interpolating linearly between $\hat{z}_{ij}$s and the resulting function $\hat{m}_j$ is an estimate of m_j. Moreover, one can easily see that $\hat{m}_j(\hat{\gamma}_i) = \hat{z}_{ij}$. When $m_j(\cdot)$ is assumed to be a monotonic Lipschitz function, additional monotonicity constraints can be exploited together with the Lipschitz constraints in the convex program above. Kakade et al. (2011) provides a computationally efficient algorithm with provable guarantees for this type of minimization problems. Cross-validation methods can be used to select the Lipschitz constant L.

Given $\hat{\gamma}_i = w_i\hat{\theta}$ and $\hat{m}_j(\hat{\gamma}_i)$, we can then use (8) to estimate β^*. The resulting estimator together with $\hat{\theta}$ can be plugged into (9) to estimate $g(\cdot)$.

3. Preliminaries. We first formalize a notion of "sparsity" by the l_q-"balls" which is used in Ye and Zhang (2010), Raskutti, Wainwright and Yu (2011) and Negahban et al. (2012).

DEFINITION 1 (l_q-"sparsity"). The l_q-"balls" of "radius" R_q for $q \in [0, 1]$ are defined by

$$\mathcal{B}_q^p(R_q) := \left\{ \beta \in \mathbb{R}^p \,|\, |\beta|_q^q = \sum_{j=1}^p |\beta_j|^q \le R_q \right\} \qquad \text{for } q \in (0, 1],$$

$$\mathcal{B}_0^p(R_0) := \left\{ \beta \in \mathbb{R}^p \,|\, |\beta|_0 = \sum_{j=1}^p 1\{\beta_j \ne 0\} \le R_0 \right\} \qquad \text{for } q = 0.$$

REMARK. For example, the exact sparsity on β corresponds to the case $q = 0$ with $R_0 = k$, which says that β has at most k nonzero components. In the more general setting $q \in [0, 1]$, membership in $\mathcal{B}_q^p(R_q)$ has various interpretations and one of them involves how quickly the ordered $|\beta^{(j)}|$s decay where $|\beta^{(j)}| \le C^* j^{-\zeta}$ for some positive constants (C^*, ζ).

The following definition is in regard to the so-called "Lower Restricted Eigenvalue" (LRE) of $\mathbb{E}[v_i^T v_i]$, related to the identification of our models.

DEFINITION 2 (LRE). For a subset $S_{\underline{\tau}} \subseteq \{1, 2, \ldots, p\}$ and all nonzero $\Delta \in \mathbb{C}(S_{\underline{\tau}}) \cap \mathbb{S}_\delta$ where

$$\mathbb{C}(S_{\underline{\tau}}) := \left\{ \Delta \in \mathbb{R}^p : |\Delta_{S_{\underline{\tau}}^c}|_1 \le 3|\Delta_{S_{\underline{\tau}}}|_1 + 4\left|\beta^*_{S_{\underline{\tau}}^c}\right|_1 \right\},$$

$$S_{\underline{\tau}} := \{j \in \{1, 2, \ldots, p\} : |\beta_j^*| > \underline{\tau}\},$$

(with $\Delta_{S_{\underline{\tau}}}$ denoting the vector in $\mathbb{R}^p$ that has the same coordinates as Δ on $S_{\underline{\tau}}$ and zero coordinates on the complement $S_{\underline{\tau}}^c$ of $S_{\underline{\tau}}$), and

$$\mathbb{S}_\delta := \{\Delta \in \mathbb{R}^p : |\Delta|_2 \geq \delta\},$$

we say the LRE condition holds if there exists a $\kappa_L > 0$ *such that the matrix* $\Sigma_v = \mathbb{E}[v_i^T v_i]$ satisfies

$$c^* \kappa_L \leq \frac{\Delta^T \mathbb{E}[v_i^T v_i] \Delta}{|\Delta|_2^2}, \tag{14}$$

where $v_i = (X_{ij} - \mathbb{E}(X_{ij}|\gamma_i^*))_{j=1}^p$ and $c^* > 0$ is a universal constant. The choices of δ and $\underline{\tau}$ will be specified in Section 4.

REMARK. As an example, note if $\lambda_{\min}(\Sigma_v) > 0$, the LRE condition holds with respect to $\kappa_L = \lambda_{\min}(\Sigma_v)$. Suppose $\beta^* \in \mathbb{R}^p$ belongs to the l_{q_2}-"balls" $\mathcal{B}_{q_2}^p(R_{q_2})$ for a "radius" R_{q_2} and $q_2 \in [0, 1]$. When β^* is exactly sparse (namely, $q_2 = 0$), we can take $\delta = 0$ and choose $S_{\underline{\tau}} = J(\beta^*)$ [where $J(\beta^*)$ denotes the support of β^*], which reduces the set $\mathbb{C}(S_{\underline{\tau}}) \cap \mathbb{S}_\delta$ to the following cone:

$$\mathbb{C}(J(\beta^*)) := \{\Delta \in \mathbb{R}^p : |\Delta_{J(\beta^*)^c}|_1 \leq 3|\Delta_{J(\beta^*)}|_1\}.$$

Let us consider a simpler case where v is observed. The sample analog of $\frac{\Delta^T \Sigma_v \Delta}{|\Delta|_2^2}$ being bounded away from 0 over the cone $\mathbb{C}(J(\beta^*))$ is the so-called *lower restricted eigenvalue* (LRE) condition on the Gram matrix $\frac{v^T v}{n}$ where $v = (v_1, \ldots, v_p) \in \mathbb{R}^{n \times p}$ [see, e.g., Bickel, Ritov and Tsybakov (2009); Meinshausen and Yu (2009); Bühlmann and van de Geer (2011); and Negahban et al. (2012)]. When β^* is approximately sparse (namely, $q_2 \in (0, 1]$), in sharp contrast to the case of exact sparsity, the set $\mathbb{C}(S_{\underline{\tau}})$ is no longer a cone but rather contains a ball centered at the origin. As a consequence, it is never possible to ensure that $\frac{|v\Delta|_2^2}{n}$ is bounded from below for all vectors Δ in the set $\mathbb{C}(S_{\underline{\tau}})$ [see Negahban et al. (2012) for a geometric illustration of this issue]. For this reason, it is crucial to further restrict the set $\mathbb{C}(S_{\underline{\tau}})$ for $q_2 \in (0, 1]$ by intersecting it with the set $\mathbb{S}_\delta := \{\Delta \in \mathbb{R}^p : |\Delta|_2 \geq \delta\}$. Provided that the parameters δ and $\underline{\tau}$ are suitably chosen, the intersection $\mathbb{C}(S_{\underline{\tau}}) \cap \mathbb{S}_\delta$ excludes many "flat" directions (with eigenvalues of 0) in the space for the case of $q_2 \in (0, 1]$. To the best of our knowledge, the necessity of this additional set $\mathbb{S}_\delta$, essential for the approximately sparse cases with $q_2 \in (0, 1]$, is first recognized explicitly in Negahban et al. (2012). To derive a general upper bound on the l_2-error of the procedure introduced in Section 2 for β^*, we use an idea similar to Negahban et al. (2012) in our analysis.

In the problem of our interest, v is unknown and needs to be estimated. We provide results that imply the LRE condition holds for $\frac{\hat{v}^T \hat{v}}{n}$ with high probability

when $\kappa_L > 0$. These results provide finite-sample guarantees of the population LRE condition when the unknown residual v is replaced with its estimate $\hat{v}$ and the expectation is replaced with a sample average.

For (2)–(4), the condition $\kappa_L > 0$ is a high-dimensional generalization of the identification assumption used in the low-dimensional sample selection model literature [e.g., Ahn and Powell (1993); Newey (2009)].

4. General error bounds on $|\hat{\beta} - \beta^*|_2$ for the partially linear model. This section provides general upper bounds on $|\hat{\beta} - \beta^*|_2$ for the standard partially linear model (1) where γ_i^* is an observed variable and the estimator $\hat{\beta}$ for β^* is obtained by the two-step procedure described in Section 2. For notational simplicity, in the theoretical results presented below, we assume the regime of interest is $p \geq (n \vee 2)$; the modification to allow $p < (n \vee 2)$ is trivial. Moreover, we assume that $n \gtrsim \log p$. Also, as a general rule for this paper, all the $c \in (0, \infty)$ constants denote positive universal constants that are independent of n, p, and R_{q_2} (and also n_0, d and R_{q_1} in Section 5). The specific values of these constants may change from place to place. We impose the following sampling assumption on the model of our interests.

ASSUMPTION 4.1. The data are i.i.d. with finite second moments.

Recall from programs (5)–(6), $\tilde{m}_j(\cdot) \in \mathcal{F}_j$. We also assume $m_j(\cdot) \in \mathcal{F}_j$. Define the shifted version of the function class $\mathcal{F}_j$:

$$\bar{\mathcal{F}}_j := \{f = f' - f'' : f', f'' \in \mathcal{F}_j\}.$$

The following regularity assumptions are imposed to obtain the theoretical results in this section.

ASSUMPTION 4.2. For any $j = 0, \ldots, p$, $\bar{\mathcal{F}}_j$ is a star-shaped function class; that is, for any $f \in \bar{\mathcal{F}}_j$ and $\alpha \in [0, 1]$, $\alpha f \in \bar{\mathcal{F}}_j$.

REMARK. The star-shaped condition is often seen in literature of nonparametric statistics [see e.g., Wainwright (2015)]. It is relatively mild; for instance, it is satisfied whenever the set $\bar{\mathcal{F}}_j$ is convex and contains the function $f = 0$. It is also satisfied, for example, when the underlying $\mathcal{F}_j$ is the class of $k-$sparse linear combinations of basis functions $\psi_l(\cdot)$s; that is, for $f' \in \mathcal{F}_j$, $f'(\gamma_i^*) = \sum_{l=1}^m \pi_l^* \psi_l(\gamma_i^*)$ and $\pi^* := (\pi_l^*)_{l=1}^m$ belongs to the l_0-"ball" of "radius" k.

For any radius $\tilde{r}_{nj} > 0$, define the conditional *local complexity*:

$$\mathcal{G}_n(\tilde{r}_{nj}; \mathcal{F}_j) := \mathbb{E}_\xi \left[\sup_{f \in \Omega(\tilde{r}_{nj}; \mathcal{F}_j)} \left| \frac{1}{n} \sum_{i=1}^n \xi_i f(\gamma_i^*) \right| \middle| \{\gamma_i^*\}_{i=1}^n \right], \tag{15}$$

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

$$\Omega(\tilde{r}_{nj}; \mathcal{F}_j) = \{f \in \bar{\mathcal{F}}_j : |f|_n \le \tilde{r}_{nj}\},$$

where $|f|_n := \sqrt{\frac{1}{n}\sum_{i=1}^n [f(\gamma_i^*)]^2}$. For any star-shaped class $\bar{\mathcal{F}}_j$ (Assumption 4.2), it can be shown that the function $t \mapsto \frac{\mathcal{G}_n(t;\mathcal{F}_j)}{t}$ is nonincreasing on the interval $(0, \infty)$ (Lemma S.15 in the supplementary materials). Therefore, we can always find some large enough $\tilde{r}_{nj} > 0$ that satisfies the *critical inequality*:

$$\mathcal{G}_n(\tilde{r}_{nj}; \mathcal{F}_j) \le \frac{\tilde{r}_{nj}^2}{2\sigma^\dagger}, \tag{16}$$

where $\sigma^\dagger$ is some sub-Gaussian parameter to be defined shortly; moreover, (16) has a smallest positive solution. In practice, determining the exact value of this smallest positive solution can be difficult so obtaining reasonable upper bounds on it is more useful. For this, we describe a method from existing literature. It is known that the Dudley's entropy integral can be used in bounding the complexity $\mathcal{G}_n(\tilde{r}_{nj}; \mathcal{F}_j)$ in (15) [see, e.g., van de Geer (2000); Bartlett and Mendelson (2002); Koltchinskii (2006); Wainwright (2015)]. Let $N_n(t; \Omega(\tilde{r}_{nj}; \mathcal{F}_j))$ denote the t-covering number of the set $\Omega(\tilde{r}_{nj}; \mathcal{F}_j)$ in the $\mathcal{L}^2(\mathbb{P}_n)$ norm. Then the smallest positive solution to (16) is bounded above by any $\tilde{r}_{nj} \in (0, \sigma^\dagger]$ such that

$$\frac{c'}{\sqrt{n}} \int_{\frac{\tilde{r}_{nj}^2}{4\sqrt{2}\sigma^\dagger}}^{\tilde{r}_{nj}} \sqrt{\log N_n\big(t; \Omega(\tilde{r}_{nj}; \mathcal{F}_j)\big)}\, dt \le \frac{\tilde{r}_{nj}^2}{4\sigma^\dagger}. \tag{17}$$

For popular function classes, the covering number $N_n(t; \Omega(\tilde{r}_{nj}; \mathcal{F}_j))$ is readily available so we can compute $\tilde{r}_{nj}$ based on (17). This method is known to yield $\tilde{r}_{nj}$ with sharp scaling in a wide range of statistical problems, one of which is illustrated in Section 5.1.

ASSUMPTION 4.3. Define $v_{i0} = Y_i - \mathbb{E}(Y_i|\gamma_i^*)$ and $v = (v_1, \ldots, v_p) \in \mathbb{R}^{n\times p}$ with $v_{ij} = X_{ij} - \mathbb{E}(X_{ij}|\gamma_i^*)$, $j = 1, \ldots, p$. (i) For any unit vector $\rho \in \mathbb{R}^p$, the random variable $\rho^T v_i^T$ is sub-Gaussian with parameter at most σ_v where v_i is the ith row of v; for $j = 0, \ldots, p$, $v_j \in \mathbb{R}^n$ is sub-Gaussian with parameter at most σ_{v_j}; (ii) the random variable η_i is sub-Gaussian with parameter at most σ_η; $\sigma_v^* = \max_{j\in\{0,\ldots,p\}} \sigma_{v_j}$ and $\sigma = \sigma_v^* \vee \sigma_\eta$.

REMARK. In the literature of nonparametric estimation, common measures of function complexities associated with sub-Gaussian variables can be controlled with standard techniques as in van der Vaart and Wellner (1996) and van de Geer (2000). There are some cases where other concentration results [e.g., Ledoux (1995/97); Bobkov and Ledoux (2000)] may provide sharper tail probabilities when we relax the identicalness of $\{\eta_i\}_{i=1}^n$ and $\{v_{ij}\}_{i=1}^n$ for each $j = 0, \ldots, p$.

These cases include: v_j for $j = 0, \ldots, p$ and η (i) have *strongly log-concave* distributions (defined below) for some $\varphi_{v_j} > 0$ and $\varphi_\eta > 0$; or, (ii) are bounded vectors such that for every $i = 1, \ldots, n$, v_{ij} and η_i are supported on the interval (a'_{v_j}, a''_{v_j}) with $B_{v_j} := a''_{v_j} - a'_{v_j}$, and on (a'_η, a''_η) with $B_\eta := a''_\eta - a'_\eta$.

DEFINITION 3 (Strongly log-concave distributions). A distribution with density $\mathtt{p}$ (with respect to the Lebesgue measure) is a strongly log-concave distribution if the density $\mathtt{p}(x) = \exp(-\psi(x))$ where the function $\psi : \mathbb{R}^n \to \mathbb{R}$ is strongly convex, meaning there is some $\varphi > 0$ such that

$$\lambda\psi(x) + (1-\lambda)\psi(y) - \psi\big(\lambda x + (1-\lambda)y\big) \geq \frac{\varphi}{2}\lambda(1-\lambda)|x-y|_2^2$$

for all $\lambda \in [0,1]$, and $x, y \in \mathbb{R}^n$.

REMARK. Any Gaussian distribution with covariance matrix $\Sigma \succ 0$ is strongly log-concave with parameter $\varphi = \lambda_{\min}(\Sigma^{-1})$. Moreover, there are various non-Gaussian distributions that are also strongly log-concave.

With all the previous "ingredients" in hand, we introduce the following quantities which are related to the sources of statistical errors in $|\hat{\beta} - \beta^*|_2$. Let

$$\mathcal{T}_0 = c_0 \max\{\mathcal{T}_2, \mathcal{T}_2^\star, \mathcal{T}_3\};$$

$$\mathcal{T}_2 = \big(|\beta^*|_1 \vee 1\big) \max_{j\in\{0,\ldots,p\}} t_{nj}^2 \qquad \text{for any } t_{nj} \geq r_{nj},$$

where $r_{nj} > 0$ is the smallest radius satisfying (16) with $\sigma^\dagger = \sigma_v^*$;

$$\mathcal{T}_2^\star = \max_{j\in\{1,\ldots,p\}} r_{nj}'^2,$$

where $r'_{nj} > 0$ is the smallest radius satisfying (16) with $\sigma^\dagger = \sigma_\eta$;

$$\mathcal{T}_3 = \max_{j\in\{1,\ldots,p\}} \sigma_{v_j}\sigma_\eta \sqrt{\frac{\log p}{n}}.$$

The following assumption concerns the LRE κ_L defined in (14).

ASSUMPTION 4.4. The coefficient vector $\beta^* \in \mathbb{R}^p$ belongs to the l_{q_2}-"balls" $\mathcal{B}_{q_2}^p(R_{q_2})$ for a "radius" R_{q_2} and $q_2 \in [0,1]$. The condition $\kappa_L > 0$ holds over the restricted set $\mathbb{C}(J(\beta^*))$ with $\underline{\tau} = 0$ and $\delta = 0$ for the exact sparsity case ($q_2 = 0$ with $R_{q_2} = k_2$), and over $\mathbb{C}(S_{\underline{\tau}}) \cap \mathbb{S}_\delta$ with $\underline{\tau} = \frac{\mathcal{T}_0}{\kappa_L}$ and $\delta = c_3 \kappa_L^{-1+\frac{q_2}{2}} R_{q_2}^{\frac{1}{2}} \mathcal{T}_0^{1-\frac{q_2}{2}}$ for the approximate sparsity case ($q_2 \in (0,1]$), respectively.

The following theorem (Theorem 4.1) provides a general upper bound on the error $|\hat{\beta} - \beta^*|_2$ when the first-step estimation concerns a program as in (5).

THEOREM 4.1. *For model* (1) *where* γ_i^* *is an observed variable, consider the procedure based on* (5) *with* $\hat{\gamma}_i := \gamma_i^*$ *and* (8). *Let Assumptions* 4.1–4.4 *hold and* $\lambda_{n,3} = \mathcal{T}_0$ *in* (8). *Assume*

$$R_{q_2}\mathcal{T}_0^{-q_2}\max\left\{\kappa_L\left(\frac{\sigma_v^4}{\kappa_L^2}\vee 1\right)\frac{\log p}{n}, \bar{t}_n^2\right\} \precsim \kappa_L^{1-q_2}, \tag{18}$$

where $\bar{t}_n = \max_{j\in\{0,\dots,p\}} t_{nj}$. *Then, for some universal constants* $c > c_3 > 0$ *with* c_3 *given in Assumption* 4.4,

$$|\hat{\beta} - \beta^*|_2 \le \frac{cR_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}} \tag{19}$$

with probability at least $1 - c_0'\exp(-c_1'\log p) - c_2'\exp(-c_3'\frac{n\bar{t}_n^2}{\sigma^2} + c_4'\log p)$.

REMARK (on Theorem 4.1). When v_js and η have *strongly log-concave* distributions with parameters $\varphi_{v_j} > 0$ and $\varphi_\eta > 0$, respectively, $c_3'\sigma^{-2}$ in the probability guarantee for Theorem 4.1 can be replaced by $\frac{1}{2}\min_{j\in\{0,\dots,p\}}\varphi_{v_j}\wedge\varphi_\eta$; when v_{ij}s and η_is are supported on (a_{v_j}', a_{v_j}'') with $B_{v_j} := a_{v_j}'' - a_{v_j}'$, and on (a_η', a_η'') with $B_\eta := a_\eta'' - a_\eta'$, respectively, $c_3'\sigma^{-2}$ can be replaced by $\frac{1}{4}(\max_{j\in\{0,\dots,p\}} B_{v_j}^2 \vee B_\eta^2)^{-1}$. Note that in (18)–(19), we may take $R_{q_2} = |\beta^*|_{q_2}^{q_2}$ for $q_2 \in (0,1]$ and $R_{q_2} = |\beta^*|_0$ for $q_2 = 0$.

Condition (18) in Theorem 4.1 ensures that with high probability, $\frac{\hat{v}^T\hat{v}}{n}$ satisfies the LRE condition over the restricted sets specified in Assumption 4.4—a finite-sample guarantee of the population identification condition. From Theorem 4.1, it can be seen that the general upper bound on $|\hat{\beta} - \beta^*|_2$ depends on two sources of errors. The terms t_{nj} (in $\mathcal{T}_2$) and r_{nj}' (in $\mathcal{T}_2^\star$) are related to the statistical errors of the nonparametric projection step and $\mathcal{T}_3$ is related to the statistical error of the Lasso estimation. The factor $|\beta^*|_1$ in $\mathcal{T}_2$ is related to the fact that the estimator uses $\hat{v}$ to surrogate for the unknown v. Other surrogate-type Lasso estimators such as the one in Rosenbaum and Tsybakov (2013) also involve the factor $|\beta^*|_1$ in the choice of their regularization parameter and error bounds.

When setting $t_{nj} = r_{nj}$ in $\mathcal{T}_2$, we note the statistical error contributed by the nonparametric projection step is $\max_j r_{nj}^2$ instead of the optimal rate $\max_j r_{nj}$ that one would expect from the nonparametric regressions as (5). Note that $\max_j r_{nj}^2 < \max_j r_{nj}$ if $\max_j r_{nj} < 1$; similarly, in terms of $\mathcal{T}_2^\star$, $\max_j r_{nj}'^2 < \max_j r_{nj}'$ if $\max_j r_{nj}' < 1$. As long as

$$\left\{\left[(|\beta^*|_1 \vee 1)\max_{j\in\{0,\dots,p\}} r_{nj}^2\right] \vee \left[\max_{j\in\{1,\dots,p\}} r_{nj}'^2\right]\right\} \precsim \mathcal{T}_3, \tag{20}$$

the scaling of the upper bound on $|\hat{\beta} - \beta^*|_2$ is as good as if the unknown nonparametric component were known. This result establishes a new nonasymptotic

"oracle" result. One of the drivers behind this oracle result lies on the terms:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{v}_{ij'}\left[\hat{m}_j(\gamma_i^*)-m_j(\gamma_i^*)\right]\right|,\qquad \left|\frac{1}{n}\sum_{i=1}^{n}\eta_i\left[\hat{m}_j(\gamma_i^*)-m_j(\gamma_i^*)\right]\right|. \tag{21}$$

As an example for the standard partially linear model (1), suppose that $\mathcal{F}_j$ is the class of linear combinations of basis functions $\psi_l(\cdot)$s for all $j=0,...,p$ such that for $f\in\mathcal{F}_j$, $f(\gamma_i^*)=\sum_{l=1}^{m}\pi_l^*\psi_l(\gamma_i^*)$ and $\pi^*:=(\pi_l^*)_{l=1}^m$ belongs to the l_1-ball of radius R. Then, under an LRE assumption and a normalization on the matrix of covariates formed by the basis functions, if $\sigma_v^*\asymp 1$, $\sigma_\eta\asymp 1$, and $m\geq p$, applying the Lasso procedure in the nonparametric projection step would yield upper bounds with scaling $R(\frac{\log m}{n})^{\frac{1}{2}}$ on the quantities in (21). These scaling attain the sharp rates on r_{nj}^2 and $r_{nj}^{'2}$. If π^* belongs to the l_0-"ball" of "radius" k, then the standard Lasso procedure would yield upper bounds with scaling $\frac{k\log m}{n}$ on the quantities in (21). These scaling almost achieve the sharp rates $\frac{k\log\frac{m}{k}}{n}$ on r_{nj}^2 and $r_{nj}^{'2}$. We point out that while the preceding examples concern functions of the univariate γ_i^*, similar ideas can apply to multivariate functions, where sparsity assumptions become even more crucial for producing manageable classes of models.

Further note that, since Theorem 4.1 holds for any $t_{nj}\geq r_{nj}$, the choice of t_{nj} incurs a trade-off between $\mathcal{T}_2$ and $c_2'\exp(-c_3'\frac{nt_n^2}{\sigma^2}+c_4'\log p)$. This is a general phenomenon for these tail bounds. When (20) holds, t_{nj} can be chosen in the way $\mathcal{T}_2\asymp\mathcal{T}_3$ so that the probability guarantee may be improved relative to the choice $t_{nj}=r_{nj}$.

The following theorem (Theorem 4.2) provides a general upper bound on the error $|\hat{\beta}-\beta^*|_2$ when the first-step estimation concerns a regularized program as in (6) where $\mathcal{F}_j$ is a reproducing kernel Hilbert space for $j=0,\ldots,p$. For Theorem 4.2, let the *local complexity* measure $\mathcal{G}_n(\tilde{r}_{nj};\mathcal{F}_j)$ be defined over the set

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

recalling $\bar{R}_j$ in (7), let $\mathcal{T}_2=(|\beta^*|_1\vee 1)\max_{j\in\{0,\ldots,p\}}\{(\bar{R}_j\vee 1)t_{nj}^2\}$ where $t_{nj}\geq(r_{nj}\vee c''\sqrt{\frac{\sigma^2}{n}})$ for some universal positive constant c'' that is sufficiently large; $\mathcal{T}_2^\star=\max_{j\in\{1,\ldots,p\}}\{(\bar{R}_j\vee 1)r_{nj}^{'2}\}$. The definitions for r_{nj}, $r_{nj}^{'2}$ and $\mathcal{T}_3$ remain the same as those for Theorem 4.1.

THEOREM 4.2. *For model* (1) *where* γ_i^* *is an observed variable, consider the procedure based on* (6) *with* $\hat{\gamma}_i:=\gamma_i^*$ *and* (8). *For* $j=0,\ldots,p$, *let* $\lambda_{nj,2}=(2+\varsigma)t_{nj}^2$ *in* (6) *for any constant* $\varsigma\geq 0$ *and* $\lambda_{n,3}=\mathcal{T}_0$ *in* (8); *also suppose Assumptions* 4.1–4.4 *hold,* $|m_j|_{\mathcal{F}_j}\leq 1$, *and condition* (18) *is satisfied with* $\tilde{t}_n^2:=\max_{j\in\{0,\ldots,p\}}\{(\bar{R}_j\vee 1)t_{nj}^2\}$. *Then the upper bound* (19) *holds* (*where* $\mathcal{T}_2$ *and* $\mathcal{T}_2^\star$ *correspond to the ones defined for Theorem* 4.2) *with probability at least* $1-c_0'\exp(-c_1'\log p)-c_2''\exp(-c_3''\frac{n\max_{j\in\{0,\ldots,p\}}t_{nj}^2}{\sigma^2}+c_4''\log p)$.

REMARK. Theorem 4.2 has similar implications as Theorem 4.1. In the context where $\mathcal{F}_j$ is a reproducing kernel Hilbert space for $j = 0, \ldots, p$, it is well understood that computing the kernel ridge regression estimate (6) can be based upon the empirical kernel matrix. Moreover, Mendelson (2002) shows that $\mathcal{G}_n(\tilde{r}_{nj}; \mathcal{F}_j) \asymp \sqrt{\frac{1}{n}\sum_{i=1}^n \min\{\tilde{r}_{nj}^2, \tilde{\mu}_i\}}$ where $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_n \geq 0$ are the eigenvalues of the underlying kernel matrix for the estimate. Consequently, we can seek an upper bound on the smallest positive solution to (16) by solving for $\tilde{r}_{nj}$ via $\sqrt{\frac{1}{n}\sum_{i=1}^n \min\{\tilde{r}_{nj}^2, \tilde{\mu}_i\}} \leq \frac{\tilde{r}_{nj}^2}{2\sigma^\dagger}$. This method is known to yield $\tilde{r}_{nj}$ with sharp scaling for various choices of kernels.

5. A leading case of the semiparametric sample selection model. In this section, we specialize our analysis to the estimation of a leading case of model (2)–(4). In the following, $n = \sum_{i=1}^{n_0} Y_{1i}$ denotes the number of the selected observations for (3) and n_0 is the number of the full sample observations for (2). Without loss of generality, we assume that the selected sample consists of the first n out of n_0 observations. For notational simplicity, in the theoretical results presented in this section, we assume the regime of interest is $p \geq (n \vee 2)$ and $d \geq (n_0 \vee 2)$; the modification to allow $p < (n \vee 2)$ and/or $d < (n_0 \vee 2)$ is trivial. Moreover, we assume that $n \gtrsim \log p$ and $n_0 \gtrsim \log d$.

5.1. *Error bounds and variable selection.* We begin with a lemma on the radii r_{nj} and r'_{nj} in Section 4; this lemma is based on (17) and results on the metric entropy of Lipschitz functions.

LEMMA 5.1. *Suppose the data are i.i.d. and the matrix W consists of bounded elements; for $j = 0, \ldots, p$, suppose $m_j(\cdot)$ belongs to the class of L-Lipschitz functions and $\tilde{m}_j$ in (5) is constrained to be in the same class. Then* $\max_{j\in\{0,\ldots,p\}} r_{nj} \leq c_0'(\frac{|\theta^*|_1 L\sigma_v^{*2}}{n})^{\frac{1}{3}}$ *and* $\max_{j\in\{1,\ldots,p\}} r'_{nj} \leq c_0''(\frac{|\theta^*|_1 L\sigma_\eta^2}{n})^{\frac{1}{3}}$.

The resulting estimator obtained using the framework described in Section 2 for model (2)–(4) is referred to as $\hat{\beta} := \hat{\beta}_{\mathrm{HSEL}}$ (a candidate for $\hat{\beta}_{\mathrm{HSEL}}$ is provided in the example in Section 2); $\hat{\beta}_{\mathrm{HSEL}}$ together with $\hat{\theta}$ can be plugged into (9) to estimate $g(\cdot)$. We now provide an upper bound for $|\hat{\beta}_{\mathrm{HSEL}} - \beta^*|_2$ based on Lemma 5.1.

THEOREM 5.1. *Suppose Assumptions 4.3–4.4 hold with $\gamma_i^* = W_i\theta^*$ and $\mathcal{T}_0$ defined in (22). Also let the conditions in Lemma 5.1 hold and $\mathcal{U} \geq \frac{1}{n}\sum_{i=1}^n L^2(W_i\hat{\theta} - W_i\theta^*)^2$ with probability at least $1 - \alpha_1$ and $\sqrt{\mathcal{U}} \precsim \sigma_v^*$. Define $\mathcal{T}_1 = M\sqrt{\mathcal{U}}$ with $M = [\sigma_v^*(|\beta^*|_1 \vee 1)] \vee \sigma_\eta$ and the definitions for $\mathcal{T}_2$, $\mathcal{T}_2^\star$ and $\mathcal{T}_3$ remain the same as those for Theorem 4.1. Assume $t_{nj} \geq c_0'(\frac{|\theta^*|_1 L\sigma_v^{*2}}{n})^{\frac{1}{3}}$ in $\mathcal{T}_2$ is*

chosen such that

$$\mathcal{T}_0 \asymp \mathcal{T}_2 \asymp \max\left\{\mathcal{T}_1, \mathcal{T}_3, \left(|\beta^*|_1 \vee 1\right)\left(\frac{|\theta^*|_1 L\sigma_v^{*2}}{n}\right)^{\frac{2}{3}}, \left(\frac{|\theta^*|_1 L\sigma_\eta^2}{n}\right)^{\frac{2}{3}}\right\}, \tag{22}$$

$$\sigma^2 \log p \precsim n \max_{j\in\{0,\dots,p\}} t_{nj}^2 \qquad \text{where } \sigma = \sigma_v^* \vee \sigma_\eta. \tag{23}$$

Suppose $\lambda_{n,3} = \mathcal{T}_0$ *and*

$$R_{q_2}\mathcal{T}_0^{-q_2}\max\left\{\kappa_L\left(\frac{\sigma_v^4}{\kappa_L^2}\vee 1\right)\frac{\log p}{n}, \mathcal{T}, \bar{t}_n^2\right\} \precsim \kappa_L^{1-q_2}$$

with $\mathcal{T} = \max_{j\in\{1,\dots,p\}}\sigma_{v_j}\sqrt{\mathcal{U}}$ *and* $\bar{t}_n = \max_{j\in\{0,\dots,p\}} t_{nj}$. *Then, with probability at least* $1-\alpha_1 - c_2\exp(-c_3\log p)$, *we have*

$$\left|\hat{\beta}_{\mathrm{HSEL}} - \beta^*\right|_2 \le \frac{c\sqrt{R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}}\mathcal{T}_0^{1-\frac{q_2}{2}}.$$

Lemma 5.2 below provides an upper bound on $\sqrt{\frac{1}{n}\sum_{i=1}^n (W_i\hat{\theta} - W_i\theta^*)^2}$ *in Theorem 5.1 when* ϵ_{1i} *in (2) has a standard normal distribution and* θ^* *is estimated by (11).*

ASSUMPTION 5.1. For any unit vector $\rho \in \mathbb{R}^d$, the random variable $\rho^T W_i^T$ (for $i = 1,\dots,n_0$) has a sub-Gaussian parameter at most σ_W where W_i is the ith row of $W \in \mathbb{R}^{n_0\times d}$; $W_j \in \mathbb{R}^{n_0\times 1}$ (for $j = 1,\dots,d$) has a sub-Gaussian parameter at most σ_{W_j}; $\sigma_W^* = \max_{j\in\{1,\dots,d\}}\sigma_{W_j}$.

ASSUMPTION 5.2. The coefficient vector $\theta^* \in \mathcal{B}_{q_1}^d(R_{q_1})$ for $q_1 \in [0,1]$ with "radius" R_{q_1}. For $\Sigma_W = \mathbb{E}[W_i^T W_i]$, $0 < \kappa_L^W \precsim \frac{\Delta^T\Sigma_W\Delta}{|\Delta|_2^2}$ for all $\Delta \in \mathbb{C}(J(\theta^*)) \setminus \{\mathbf{0}\}$ (namely, $\underline{\tau}_1 = 0$, $\delta_1 = 0$) for the exact sparsity case of θ^* with $q_1 = 0$, $R_{q_1} = k_1$, and for all nonzero $\Delta \in \mathbb{C}(S_{\underline{\tau}_1}) \cap \mathbb{S}_{\delta_1}$ where $\underline{\tau}_1 = \frac{c'\sigma_W^*}{\kappa_L^W}\sqrt{\frac{\log d}{n_0}}$ and $\delta_1 = c''(\kappa_L^W)^{-1+\frac{q_1}{2}}R_{q_1}^{\frac{1}{2}}(\sigma_W^*\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ for the approximate sparsity case of θ^* with $q_1 \in (0,1]$; $\frac{\Delta^T\Sigma_W\Delta}{|\Delta|_2^2} \le \kappa_U^W < \infty$ for all $\Delta \in \mathbb{C}(S_{\underline{\tau}_1}) \setminus \{\mathbf{0}\}$.

LEMMA 5.2. *Suppose the data* $\{Y_{1i}, W_i\}_{i=1}^{n_0}$ *are i.i.d. and Assumptions 5.1–5.2 hold;* $\mathbb{P}(Y_{1i} = 1|W_i)$ *is bounded away from* 0 *and* 1; $n_0 \ge b_0 R_{q_1}^{\frac{2}{2-q_1}}\log d$ *and* $b_0' R_{q_1}\frac{\log d}{n}(\sqrt{\frac{\log d}{n_0}})^{-q_1} \le 1$ *for some sufficiently large constants* $b_0, b_0' > 0$ *that only depend on* $\kappa_L^W, \kappa_U^W, \sigma_W$ *and* σ_W^*. *If* $\hat{\theta}$ *solves program (11) with* $\lambda_{n_0,1} = c'\sigma_W^*\sqrt{\frac{\log d}{n_0}}$,

then, with probability at least $1-\alpha = 1 - c_2' \exp(-c_3' \log d)$,

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} [W_i(\hat{\theta} - \theta^*)]^2 \right\}^{\frac{1}{2}} \le c_1 \frac{R_{q_1}^{\frac{1}{2}} \sqrt{\kappa_U^W}}{(\kappa_L^W)^{1-\frac{q_1}{2}}} \left(\sigma_W^* \sqrt{\frac{\log d}{n_0}} \right)^{1-\frac{q_1}{2}}.$$

REMARK. If $\kappa_L^W \succsim 1$, $\kappa_U^W \precsim 1$, $\sigma_W^* \precsim 1$, $\sigma_v^* \precsim 1$ and $L \precsim 1$, with Lemma 5.2, we have $\mathcal{T} \precsim R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ and $\mathcal{T}_1 \precsim M R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ in Theorem 5.1 where $M \asymp (|\beta^*|_1 \vee 1) \vee \sigma_\eta$; moreover, for $q_1 = 0$ with $R_{q_1} = k_1$, if $n \asymp n_0$ and $|\theta_j^*| \le \overline{b} < \infty$ for all $j \in J(\theta^*)$, then

$$\max\left\{ (|\beta^*|_1 \vee 1) \left(\frac{|\theta^*|_1}{n} \right)^{\frac{2}{3}}, \left(\frac{|\theta^*|_1 \sigma_\eta^2}{n} \right)^{\frac{2}{3}} \right\} \precsim M \sqrt{\frac{k_1 \log d}{n_0}}$$

so $\mathcal{T}_0 = c_0[(M\sqrt{\frac{k_1 \log d}{n_0}}) \vee (\sigma_\eta \sqrt{\frac{\log p}{n}})]$, and consequently,

$$|\hat{\beta}_{\mathrm{HSEL}} - \beta^*|_2 \le \frac{c\sqrt{R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}} \left[\left(M \sqrt{\frac{k_1 \log d}{n_0}} \right) \vee \left(\sqrt{\frac{\sigma_\eta^2 \log p}{n}} \right) \right]^{1-\frac{q_2}{2}}$$

with probability at least $1 - c_2 \exp(-c_3 \log p) - c_2' \exp(-c_3' \log d)$. Suppose $|\beta^*|_1 \succsim 1$; let us illustrate various choices of $t_{nj}^2 \ge r_{nj}^2$ with the exact sparsity case of θ^*. Setting

$$t_{nj}^2 \asymp \max\left\{ \sqrt{\frac{\sigma_\eta^2 \log p}{n|\beta^*|_1^2}}, \sqrt{\frac{\sigma_\eta^2 k_1 \log d}{n_0 |\beta^*|_1^2}}, \sqrt{\frac{k_1 \log d}{n_0}} \right\} \succsim \left(\frac{k_1}{n} \right)^{\frac{2}{3}} \succsim r_{nj}^2$$

makes $\mathcal{T}_2 \asymp \sqrt{\frac{\sigma_\eta^2 \log p}{n}} \vee [(|\beta^*|_1 \vee \sigma_\eta) \sqrt{\frac{k_1 \log d}{n_0}}]$ so that (22) holds. Under this choice of t_{nj}^2, condition (23) in Theorem 5.1 requires that $\max\{\sqrt{\frac{n\sigma_\eta^2 \log p}{|\beta^*|_1^2}}, \sqrt{\frac{n^2 \sigma_\eta^2 k_1 \log d}{n_0 |\beta^*|_1^2}}, \sqrt{\frac{n^2 k_1 \log d}{n_0}}\} \succsim \sigma^2 \log p$ in order to ensure $\exp(-c_3' \frac{n\tilde{t}_\eta^2}{\sigma^2} + c_4' \log p) \precsim \exp(-c_3 \log p)$.

More generally, when $\mathcal{F}_j$ in (5) belongs to a Hölder class of order $\nu \ge 1$, we can again apply (17) and results on the metric entropy of νth order smoothness classes to show that $r_{nj}^2 \precsim (\frac{k_1}{n})^{\frac{2\nu}{2\nu+1}}$ for all $j = 0, \ldots, p$ and $r_{nj}'^2 \precsim (\frac{k_1 \sigma_\eta^2}{n})^{\frac{2\nu}{2\nu+1}}$ for all $j = 1, \ldots, p$. Choosing

$$t_{nj}^2 \asymp \max\left\{ \sqrt{\frac{\sigma_\eta^2 \log p}{n|\beta^*|_1^2}}, \sqrt{\frac{\sigma_\eta^2 k_1 \log d}{n_0 |\beta^*|_1^2}}, \sqrt{\frac{k_1 \log d}{n_0}} \right\}$$

yields

$$|\hat{\beta}_{\mathrm{HSEL}} - \beta^*|_2 \le \frac{c\sqrt{R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}} \left[\sqrt{\frac{\sigma_\eta^2 \log p}{n}} \vee \left((|\beta^*|_1 \vee \sigma_\eta) \sqrt{\frac{k_1 \log d}{n_0}} \right) \right]^{1-\frac{q_2}{2}}$$

with probability at least $1 - c_2 \exp(-c_3 \log p) - c_2' \exp(-c_3' \log d)$.

Variable-selection for exactly sparse β^*. The following theorem (Theorem 5.2) establishes a result on the variable selection property of $\hat{\beta}_{\text{HSEL}}$ when β^* is exactly sparse ($q_2 = 0$). Theorem 5.2 requires the following assumption.

ASSUMPTION 5.3. $\|\mathbb{E}[v_{i,J(\beta^*)^c}^T v_{i,J(\beta^*)}][\mathbb{E}(v_{i,J(\beta^*)}^T v_{i,J(\beta^*)})]^{-1}\|_\infty \leq 1 - \phi$ for some constant $\phi \in (0, 1]$.

REMARK. Assumption 5.3 is a population "incoherence condition" similar to Wainwright (2009). The "incoherence condition" is a refined version of the "irrepresentable condition" in Zhao and Yu (2006) and the "neighborhood stability condition" by Meinshausen and Bühlmann (2006). Bühlmann and van de Geer (2011) shows this type of conditions is sufficient and "essentially necessary" for the Lasso to achieve consistent variable selection.

THEOREM 5.2. *Let Assumptions 5.3 and the conditions in Theorem 5.1 hold*; $\frac{k_2^3 \log p}{n} \precsim b^*$ *and* $[(k_2^2 \mathcal{T}) \vee (k_2 \max_{j \in \{0,\ldots,p\}} t_{nj}^2)] \precsim b_0^*$ *for some* $b_0^*, b^* > 0$ *depending only on* $\lambda_{\min}(\mathbb{E}[v_{i,J(\beta^*)}^T v_{i,J(\beta^*)}])$, ϕ, *and* $\max_{j \in \{1,\ldots,p\}} \sigma_{v_j}$. *Suppose* $\lambda_{n,3} = \frac{16-2\phi}{\phi} \mathcal{T}_0$. *Then, with probability at least* $1 - \alpha_1 - c_2 \exp(-c_3 \log p)$, *we have*: (a) *the support* $J(\hat{\beta}_{\text{HSEL}}) \subseteq J(\beta^*)$; (b) *if* $\min_{j \in J(\beta^*)} |\beta_j^*| > \bar{B}$, *where*

$$\bar{B} := \frac{c''(16 - 2\phi)\sqrt{k_2}\mathcal{T}_0}{\phi \lambda_{\min}(\mathbb{E}[v_{i,J(\beta^*)}^T v_{i,J(\beta^*)}])}$$

then $J(\hat{\beta}_{\text{HSEL}}) \supseteq J(\beta^*)$, *and hence* $J(\hat{\beta}_{\text{HSEL}}) = J(\beta^*)$.

Bounds on the nonparametric selection bias function. For the case where $g(\cdot)$ belongs to the class of Lipschitz functions,[2] we consider the plug-in nonparametric least squares estimator (9) (e.g., a candidate may be the Lipschitz regression) by surrogating θ^* with $\hat{\theta}$ and β^* with $\hat{\beta}_{\text{HSEL}}$.

ASSUMPTION 5.4. For any unit vector $\rho \in \mathbb{R}^p$, the random variable $\rho^T X_i^T$ (for $i = 1, \ldots, n$) is sub-Gaussian with parameter at most σ_X where X_i is the ith row of $X \in \mathbb{R}^{n \times p}$; for all $\Delta \in \mathbb{C}(S_{\underline{\tau}}) \setminus \{\mathbf{0}\}$ with $\underline{\tau}$ defined in Assumption 4.4, the matrix $\Sigma_X = \mathbb{E}[X_i^T X_i]$ satisfies $\frac{\Delta^T \Sigma_X \Delta}{|\Delta|_2^2} \leq \kappa_U^X < \infty$; $b_0'' R_{q_2} \frac{\log p}{n} \mathcal{T}_0^{-q_2} \leq 1$ for some sufficiently large constant $b_0'' > 0$ that only depends on κ_U^X, σ_X and κ_L; $g(\cdot)$

[2]When ϵ_{1i} and ϵ_{2i} in (2)–(3) are bivariate normal with $\text{var}(\epsilon_{1i}) = \text{var}(\epsilon_{2i}) = 1$, the selection bias function characterized by the Inverse Mills Ratio is a 1-Lipschitz function.

belongs to the class of $\bar{L}$-Lipschitz functions and $\tilde{g}$ in (9) is constrained to be in the same class.

THEOREM 5.3. *Define* $\mathcal{T}_0^g = c' \max_{l\in\{1,2,3,4\}}\{\mathcal{T}_l^g\}$ *where*

$$\mathcal{T}_1^g = \left(\frac{|\theta^*|_1 \sigma_\eta^2 \bar{L}}{n}\right)^{\frac{1}{3}}, \qquad \mathcal{T}_2^g = \frac{\sqrt{\kappa_U^X R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}} \mathcal{T}_0^{1-\frac{q_2}{2}},$$

$$\mathcal{T}_3^g = \bar{L} \frac{R_{q_1}^{\frac{1}{2}} \sqrt{\kappa_U^W}}{(\kappa_L^W)^{1-\frac{q_1}{2}}} \left(\sigma_W^* \sqrt{\frac{\log d}{n_0}}\right)^{1-\frac{q_1}{2}},$$

$$\mathcal{T}_4^g = \sqrt{\bar{L}(\kappa_L^W)^{q_1-1} \sigma_\eta R_{q_1}} \left(\frac{\sigma_W^{*2} \log d}{n}\right)^{\frac{1}{4}} \left(\frac{\sigma_W^{*2} \log d}{n_0}\right)^{\frac{1-q_1}{4}}.$$

Let the conditions in Theorem 5.1, Lemma 5.2, and Assumption 5.4 hold. Then, for the estimator $\hat{g}(\cdot)$ *obtained by (9),*

$$\left\{\frac{1}{n}\sum_{i=1}^n [\hat{g}(W_i\hat{\theta}) - g(W_i\theta^*)]^2\right\}^{\frac{1}{2}} \le \mathcal{T}_0^g$$

with probability at least

$$1 - c_2 \exp(-c_3 \log p) - c_2' \exp(-c_3' \log d) - c_4 \exp\left\{-c_5 \frac{n(\mathcal{T}_0^g)^2}{\sigma_\eta^2}\right\}.$$

REMARK. Note that $\mathcal{T}_1^g$ is expected from the fact that $g(\cdot)$ belongs to the class of $\bar{L}$-Lipschitz functions; $\mathcal{T}_2^g$ and $\mathcal{T}_3^g$ are expected from the statistical errors of $\hat{\beta}_{\mathrm{HSEL}}$ and $\hat{\theta}$ that we plug into the nonparametric regression (9); $\mathcal{T}_4^g$ comes from controlling the random variable $\frac{1}{n}\sum_{i=1}^n \eta_i [W_i\hat{\theta} - W_i\theta^*]$ using the fact that $\mathbb{E}[\eta_i | W_i, Y_{1i} = 1] = 0$ implied by the single index restriction (4).

For the standard partially linear model (1) with the observed variable γ_i^*, if we define the local complexity via $\mathcal{F}$ in (9) and the associated critical radius $r_n'^g$ in a similar fashion as (15) and (16) with $\sigma^\dagger = \sigma_\eta$, respectively, we have

$$\left\{\frac{1}{n}\sum_{i=1}^n [\hat{g}(\gamma_i^*) - g(\gamma_i^*)]^2\right\}^{\frac{1}{2}} \le \mathcal{T}_0^g := c'' \max\left\{r_n'^g, \frac{\sqrt{\kappa_U^X R_{q_2}}}{\kappa_L^{1-\frac{q_2}{2}}} \mathcal{T}_0^{1-\frac{q_2}{2}}\right\}$$

with probability at least $1 - c_2'' \exp(-c_3'' \log p) - c_4' \exp\{-c_5' \frac{n(\mathcal{T}_0^g)^2}{\sigma_\eta^2}\}$, where $\mathcal{T}_0$ is defined in Theorem 4.1.

5.2. *Monte-Carlo simulations*.

Choosing the regularization parameters. We now turn to evaluating the performance of $\hat{\beta}_{\mathrm{HSEL}}$ using Monte-Carlo simulations. In practice, the choice of the regularization parameter $\lambda_{n_0,1}$ in (11) is straightforward. On the other hand, the choice of our last-stage regularization parameter $\lambda_{n,3}$ in (8) depends on $|\beta^*|_1$. Guided by the theoretical specification for $\lambda_{n,3}$ in Theorem 5.1, we propose a simple algorithm for its practical choice, which consists of two main steps: By over-penalizing, the first step sets $\lambda_{n,3}$ to a choice that has a size no smaller than its optimal value and returns an initial estimator $\hat{\beta}^{(1)}_{\mathrm{HSEL}}$ such that $(|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1 \vee 1) \asymp (|\beta^*|_1 \vee 1)$ with probability at least $1-c_2\exp(-c_3\log p)-c_2'\exp(-c_3'\log d)$; the second step tunes the amount of regularization and sharpens the rate of convergence using the initial estimator $|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1$ returned by Step 1.

This algorithm also allows us to account for the size of the noise term η and the analysis involves a variance-based control which reduces the sub-Gaussian parameter σ_η to $\sqrt{\mathrm{var}(\eta_i)}$. In what follows, $\sigma_\eta := \sqrt{\mathrm{var}(\eta_i)}$. We introduce the quantity $\hat{\mathcal{T}}_1 = \sqrt{\frac{(\hat{k}_1 \vee 1)\log d}{n_0}}$ where $\hat{k}_1 = |J(\hat{\theta})|$ and initialize $|\hat{\beta}^{(0)}_{\mathrm{HSEL}}|_1 := (\frac{n_0}{\log d})^{\frac{1}{4}}$, $\hat{\sigma}^{(0)}_\eta := (\frac{n_0}{\log d})^{\frac{1}{4}} \wedge (\frac{n}{\log p})^{\frac{1}{4}}$. The algorithm is described in the following.

Algorithm for setting the regularization parameter $\lambda_{n,3}$.

1. *Over-Penalization*: Perform (8) with
$$\lambda_{n,3} = \lambda^{(0)}_{n,3} = \hat{\mathcal{T}}^{(0)}_0 = \left[(|\hat{\beta}^{(0)}_{\mathrm{HSEL}}|_1 \hat{\mathcal{T}}_1) \vee \left(\hat{\sigma}^{(0)}_\eta \sqrt{\frac{\log p}{n}} \right) \right]$$
and obtain $\hat{\beta}^{(1)}_{\mathrm{HSEL}}$; form $\hat{\sigma}^{(1)}_\eta = \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{v}_{i0} - \hat{v}_i \hat{\beta}^{(1)}_{\mathrm{HSEL}})^2}$.
2. *Adjusted-Penalization*: Perform (8) with
$$\lambda_{n,3} = \lambda^{(1)}_{n,3} = \hat{\mathcal{T}}^{(1)}_0 = c\left[(\hat{M}^{(1)} \hat{\mathcal{T}}_1) \vee \left(\hat{\sigma}^{(1)}_\eta \sqrt{\frac{\log p}{n}} \right) \right] \tag{24}$$
[where $\hat{M}^{(1)} = (|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1 \vee 1) \vee \hat{\sigma}^{(1)}_\eta$] and obtain $\hat{\beta}^{(2)}_{\mathrm{HSEL}}$.
3. *(Optional)*: Form $\hat{\sigma}^{(2)}_\eta = \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{v}_{i0} - \hat{v}_i \hat{\beta}^{(2)}_{\mathrm{HSEL}})^2}$. By replacing $|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1$ with $|\hat{\beta}^{(2)}_{\mathrm{HSEL}}|_1$ and $\hat{\sigma}^{(1)}_\eta$ with $\hat{\sigma}^{(2)}_\eta$ in (24), additional adjustment can be applied.

REMARK. Note that the algorithm above works if $|\beta^*|_1 \precsim (\frac{n_0}{\log d})^{\frac{1}{4}}$ and $\sigma_\eta \precsim (\frac{n_0}{\log d})^{\frac{1}{4}} \wedge (\frac{n}{\log p})^{\frac{1}{4}}$. The growth rate condition $(\frac{n_0}{\log d})^{\frac{1}{4}}$ on $|\beta^*|_1$ and σ_η are motivated by looking at the largest possible scalings (on σ_η and $|\beta^*|_1$) that lead to $\mathcal{T}_1 \asymp 1$ when $q_1 = 1$ and $\sqrt{|\theta^*|_1} \gtrsim 1$ (cf. $\mathcal{T}_1$ in the remark following Lemma 5.2). The other growth rate condition $(\frac{n}{\log p})^{\frac{1}{4}}$ on σ_η is stronger than what is needed for

$\mathcal{T}_3 \asymp 1$ and can be indeed increased to $(\frac{n}{\log p})^{\varepsilon^*}$ for $\varepsilon^* \in (\frac{1}{4}, \frac{1}{2})$ at the expense of introducing a possibly larger estimation error in $|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1$ and $\hat{\sigma}^{(1)}_\eta$. Given that in problems with i.i.d. data the variance of noise is often assumed to be a constant which does not grow with n, using $(\frac{n}{\log p})^{\frac{1}{4}}$ in $\hat{\sigma}^{(0)}_\eta$ for the initialization of the algorithm is in fact a conservative choice.

A theoretical guarantee for the two-step algorithm above is provided by Theorem A.1 in Section A.8 of the supplementary materials. In particular, it shows that under certain conditions, with probability at least $1 - c_2\exp(-c_3\log p) - c_2'\exp(-c_3'\log d)$, $(|\hat{\beta}^{(1)}_{\mathrm{HSEL}}|_1 \vee 1) \asymp (|\beta^*|_1 \vee 1)$; moreover, in the second step of the algorithm, $\lambda_{n,3} = \lambda^{(1)}_{n,3} = \hat{\mathcal{T}}^{(1)}_0 \asymp \mathcal{T}^{(1)}_0$ achieve the size needed for Theorem 5.1 and

$$|\hat{\beta}^{(2)}_{\mathrm{HSEL}} - \beta^*|_2 \le \frac{c' R_{q_2}^{\frac{1}{2}}}{\kappa_L^{1-\frac{q_2}{2}}} \big(\mathcal{T}^{(1)}_0\big)^{1-\frac{q_2}{2}},$$

with probability at least $1 - c_2\exp(-c_3\log p) - c_2'\exp(-c_3'\log d)$, where

$$\mathcal{T}^{(1)}_0 = c\max\left\{M\sqrt{\frac{(|S_{\underline{\tau}_1}| \vee 1)\log d}{n_0}}, \sqrt{\frac{\sigma_\dagger^2 \log p}{n}}\right\} \tag{25}$$

with $\sigma_\dagger^2 = \sigma_\eta^2 \vee \mathcal{T}_\eta$ and $M = (|\beta^*|_1 \vee 1) \vee \sigma_\dagger$; $\mathcal{T}_\eta \ge |(\hat{\sigma}^{(1)}_\eta)^2 - \sigma_\eta^2|$ with probability at least $1 - c_2\exp(-c_3\log p) - c_2'\exp(-c_3'\log d)$ and $\mathcal{T}_\eta$ is defined in (26) in the supplement. If $\sigma_\eta \neq 0$, Theorem A.1 requires $\mathcal{T}_\eta \le \frac{1}{2}\sigma_\eta^2$ which imposes growth conditions on n and n_0; consequently, $\sigma_\dagger^2 = \sigma_\eta^2 \asymp (\hat{\sigma}^{(1)}_\eta)^2$ and $\lambda_{n,3} = \lambda^{(1)}_{n,3} = \hat{\mathcal{T}}^{(1)}_0 \asymp \mathcal{T}^{(1)}_0$ have the optimal scaling. In the case $\sigma_\eta = 0$, $\sigma_\dagger^2 = \mathcal{T}_\eta$ as $(\hat{\sigma}^{(1)}_\eta)^2$ still involves a nonnegative estimation error bounded above by $\mathcal{T}_\eta$.

The quantity $|S_{\underline{\tau}_1}|$ denotes the cardinality of the thresholded set

$$S_{\underline{\tau}_1} := \{j \in \{1, 2, \ldots, d\} : |\theta^*_j| > \underline{\tau}_1\}$$

with $\underline{\tau}_1$ defined in Lemma A.9[3] in the supplement (also Assumption 5.2 in Section 5.1). When $|\theta^*_{S^c_{\underline{\tau}_1}}|_1$ is sufficiently small as in many important problems [i.e., under the condition $|\theta^*_{S^c_{\underline{\tau}_1}}|_1 \le c_2\frac{(|S_{\underline{\tau}_1}|\vee 1)}{\kappa_L^W}\sqrt{\frac{\log d}{n_0}}$ in Lemma A.9], the estimation error bound for $|\hat{\theta} - \theta^*|_2$ dominates the approximation error bound; this fact together with the assumptions $\kappa_L^W \gtrsim 1$, $\kappa_U^W \precsim 1$, and $\sigma_W^* \precsim 1$ in Lemma A.9 yield $\sqrt{\frac{1}{n}\sum_{i=1}^n (W_i\hat{\theta} - W_i\theta^*)^2} \precsim |\hat{\theta} - \theta^*|_2 \precsim \sqrt{\frac{(|S_{\underline{\tau}_1}|\vee 1)\log d}{n_0}}$ with probability at

[3]In Lemma A.9, we use the notation $|S_{\underline{\tau}_1}| = k_1$, which is more general than the exact sparsity parameter k_1 defined for Lemma 5.2.

least $1 - c_2' \exp(-c_3' \log d)$. Note that the term $\sqrt{\frac{|S_{\underline{\tau}_1}| \log d}{n_0}}$ is no greater than the term $R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ in Lemma 5.2. This is because in deriving a general bound on $|\hat{\theta} - \theta^*|_2$ with $\theta^* \in \mathcal{B}_{q_1}^d(R_{q_1})$ for Lemma 5.2 in Section 5.1, we utilize $|S_{\underline{\tau}_1}| \leq \underline{\tau}_1^{-q_1} R_{q_1}$ for $q_1 \in (0, 1]$ [for $q_1 = 0$, we simply have $|S_{\underline{\tau}_1}| = |J(\theta^*)|$]; consequently, $\sqrt{\frac{|S_{\underline{\tau}_1}| \log d}{n_0}} \leq \sqrt{\frac{\underline{\tau}_1^{-q_1} R_{q_1} \log d}{n_0}} \asymp R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$. To summarize the previous observations, $\sqrt{\frac{(|S_{\underline{\tau}_1}| \vee 1) \log d}{n_0}}$ in $\mathcal{T}_0^{(1)}$ (25) provides a sharper upper bound for $\sqrt{\frac{1}{n} \sum_{i=1}^n (W_i \hat{\theta} - W_i \theta^*)^2}$ than $R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ in Lemma 5.2 when $|S_{\underline{\tau}_1}| \geq 1$ and $|\theta^*_{S^c_{\underline{\tau}_1}}|_1 \leq c_2 \frac{(|S_{\underline{\tau}_1}| \vee 1)}{\kappa_L^W} \sqrt{\frac{\log d}{n_0}}$ (without this additional condition, the bound in Lemma 5.2 is indeed sharp).

In practice, basing on $\sqrt{\frac{(|S_{\underline{\tau}_1}| \vee 1) \log d}{n_0}}$ to set $\lambda_{n,3}$ is also easier to implement than basing on $R_{q_1}^{\frac{1}{2}} (\sqrt{\frac{\log d}{n_0}})^{1-\frac{q_1}{2}}$ as we can surrogate $|S_{\underline{\tau}_1}|$ with $\hat{k}_1 = |J(\hat{\theta})|$. Under a "separation condition" on $\min_{j \in S_{\underline{\tau}_1}} |\theta_j^*|$ and a "bounded sparse eigenvalue" condition on $\frac{W^T W}{n_0}$, together with the condition $|\theta^*_{S^c_{\underline{\tau}_1}}|_1 \leq c_2 \frac{(|S_{\underline{\tau}_1}| \vee 1)}{\kappa_L^W} \sqrt{\frac{\log d}{n_0}}$, we have $(\hat{k}_1 \vee 1) \asymp (|S_{\underline{\tau}_1}| \vee 1)$ (see Assumption A.2 and Lemma S.7 in the supplement).

Simulation designs. We simulate data based on model (2)–(3) where $W \in \mathbb{R}^{n_0 \times d}$ is a matrix consisted of independent uniform random variables on $[-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2}]$ with variance 1 and X takes on the first p columns of W. The i.i.d. errors $\epsilon_{1i} \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n_0$ where n_0 denotes the number of observations generated for equation (2). Conditional on the observations with $Y_{1i} = 1$, the i.i.d. errors $(\epsilon_{1i}, \epsilon_{2i})$ have the following joint normal distribution:

$$(\epsilon_{1i}, \epsilon_{2i}) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & 1 \end{pmatrix}\right), \tag{26}$$

where $\varrho \in \{0, 0.9\}$. Given the structure of $(\epsilon_{1i}, \epsilon_{2i})$ in (26), $g(\cdot) = \varrho \frac{\phi(\cdot)}{\Phi(\cdot)}$ in (4), the scaled inverse mills-ratio [$\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal c.d.f. and p.d.f., respectively].

We set $d = 400$, $p = 200$, and $n_0 = 300$. On average $n = 150$ observations will be used for estimating the main equation (3). This setup represents a scenario where the number of covariates in (2) and (3), respectively, exceeds the number of observations used to estimate the corresponding equation. We consider two sparse designs where $\theta_l^* = 0.5$ for $l = 1, 2, 3, 201$ and $\beta_1^* = \beta_{200}^* = 1$ for both designs. For $l \neq 1, 2, 3, 201$ and $j \neq 1, 200$, Design A sets $\theta_l^* = 0$ and $\beta_j^* = 0$ (exact sparsity) while Design B sets $\theta_l^* = \frac{0.1}{l}$ and $\beta_j^* = \frac{0.1}{j}$ (approximate sparsity). This approxi-

TABLE 1
Simulation results for $d = 400$, $p = 200$, *and* $n_0 = 300$

	a	b	c	d	e	f	g
Design A, $\varrho = 0$	0.234	0.723	0.767	0	0.373	2	4.2
Design A, $\varrho = 0.9$	0.230	0.686	0.773	0	0.398	2	4.2
Design B, $\varrho = 0$	0.230	0.734	0.768	3×10^{-6}	0.374	3.1	4.2
Design B, $\varrho = 0.9$	0.222	0.668	0.782	-3×10^{-6}	0.412	2.4	4.2

mate sparsity design is motivated by one of the interpretations for membership in $\mathcal{B}_q^p(R_q)$ discussed in the remark following Definition 1.

Following the process described above, 100 sets of data are generated. We choose $\lambda_{n_0,1} = \sqrt{\operatorname{var}(W_{ij})\frac{\log d}{n_0}} = \sqrt{\frac{\log d}{n_0}}$ and choose $\lambda_{n,3}$ according to the algorithm described earlier with three iterations and the choice $c = 0.5$ in (24) for the second and third iterations.

Table 1 reports the following: (a) the mean of $\lambda_{n,3}$ (from the third iteration) over the 100 trials; (b) $\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_1^t$; (c) $\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_{200}^t$; (d) $\frac{1}{198}\sum_{j\neq 1,200}\frac{1}{100}\sum_{t=1}^{100}\hat{\beta}_j^t$; (e) $\frac{1}{100}\sum_{t=1}^{100}|\hat{\beta}^t - \beta^*|_2$; (f) $\frac{1}{100}\sum_{t=1}^{100}\sum_{j=1}^{200}1\{\hat{\beta}_j^t \neq 0\}$; (g) $\frac{1}{100}\sum_{t=1}^{100}\sum_{j=1}^{400}1\{\hat{\theta}_j^t \neq 0\}$. The results in *columns b–g* show that our estimator in conjunction with the algorithm for setting $\lambda_{n,3}$ performs well for these sparse designs.

6. Future work. Here, we discuss two future directions for this work. First, it is worth noting that while perfect variable selection of $\hat{\beta}_{\mathrm{HSEL}}$ is a desirable property in the sense that it allows us to conduct post-selection inference by performing low-dimensional procedures on the selected model, we recognize that the conditions required in Theorem 5.2 might be hard to achieve in practice. Therefore, it is useful to build a bias-corrected post procedure which uses $\hat{\beta}_{\mathrm{HSEL}}$ as an initial estimate to construct confidence intervals for individual coefficients and linear combinations of several of them [similar to Zhang and Zhang (2014)].

Second, it may be worthwhile to extend our analysis to allow for non-sub-Gaussian errors η in (1). There are a couple of ways to relax the sub-Gaussian condition on the error terms. For example, the square-root Lasso and the pivotal Dantzig selector in literature evoke a bound for moderate deviations of self-normalized sums of random variables which does not require sub-Gaussian tails. However, compared to the standard Lasso, the square-root Lasso or the pivotal Dantzig selector involves a more sophisticated optimization algorithm computation-wise. Another paper by Minsker (2015) that uses a "trick" is also able to avoid imposing a sub-Gaussian condition on the error terms when deriving the nonasymptotic bounds for the standard Lasso. It is possible to apply these techniques in our problem, albeit doing so would distract the main focus of this paper; therefore, we leave these extensions to future research.

Acknowledgments. I am grateful to the Associate Editor, the anonymous referee, Professor Runze Li, and the co-editor Professor Tailen Hsing for providing valuable feedback. I thank my Ph.D. committee members Professor Martin Wainwright and Professor James Powell for useful suggestions and discussions. I also thank Professor Jeffrey Wooldridge for useful comments. All errors are my own.

SUPPLEMENTARY MATERIAL

Supplementary materials for "Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients" (DOI: 10.1214/16-AOS1528SUPP; .pdf). This supplement contains two Appendices. Appendix A provides the proofs for the main results and Appendix S provides the remaining technical lemmas and proofs.

REFERENCES

AHN, H. and POWELL, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econometrics* **58** 3–29. MR1230978

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026

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

BOBKOV, S. G. and LEDOUX, M. (2000). From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.* **10** 1028–1052.

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32** 898–927.

BUNEA, F. and WEGKAMP, M. H. (2004). Two-stage model selection procedures in partially linear regression. *Canad. J. Statist.* **32** 105–118. MR2064395

CHEN, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

DONALD, S. G. and NEWEY, W. K. (1994). Series estimation of semilinear models. *J. Multivariate Anal.* **50** 30–40.

ENGLE, R., GRANGER, C., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FAN, J. and LI, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99** 710–723. MR2090905

GRONAU, R. (1973). The effects of children on the housewife's value of time. *J. Polit. Econ.* **81** S168–S199.

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

HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* **5** 475–492.

KAKADE, S., KALAI, A. T., KANADE, V. and SHAMIR, O. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems* **23** 927–935.

KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

LEDOUX, M. (1995/97). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Stat.* **1** 63–87. MR1399224

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

LIANG, H., LIU, X., LI, R. and TSAI, C.-L. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38** 3811–3836.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Proceedings of COLT* 29–43.

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

NEWEY, W. K. (2009). Two-step series estimation of sample selection models. *Econom. J.* **12** S217–S229. MR2543488

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762

ROSENBAUM, M. and TSYBAKOV, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes—A Festschrift in Honor of Jon* (A. Wellner, M. Banerjee et al., eds.). *IMS Collections* **9** 276–290. IMS, Beachwood, OH.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

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

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

WAINWRIGHT, J. M. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Univ. California, Berkeley. In preparation.

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

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

YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

ZHU, Y. (2017). Supplement to "Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients." DOI:10.1214/16-AOS1528SUPP.

ZHU, L.. DONG, Y. and LI, R. (2013). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statist. Sinica* **23** 1235–1255. MR3114712

DEPARTMENT OF ECONOMICS
MICHIGAN STATE UNIVERSITY
486 W CIRCLE DR RM 110
EAST LANSING, MICHIGAN 48824
USA
E-MAIL: yzhu@msu.edu