Debiased Inference on Treatment Effect in a High Dimensional Model

Jingshen Wang, Xuming He, and Gongjun Xu

Department of Statistics, University of Michigan–Ann Arbor

Abstract

This article concerns the potential bias in statistical inference on treatment effects when a large number of covariates are present in a linear or partially linear model. While the estimation bias in an under-fitted model is well understood, we address a lesser known bias that arises from an over-fitted model. The over-fitting bias can be eliminated through data splitting at the cost of statistical efficiency, and we show that smoothing over random data splits can be pursued to mitigate the efficiency loss. We also discuss some of the existing methods for debiased inference and provide insights into their intrinsic bias-variance trade-off, which leads to an improvement in bias controls. Under appropriate conditions we show that the proposed estimators for the treatment effects are asymptotically normal and their variances can be well estimated. We discuss the pros and cons of various methods both theoretically and empirically, and show that the proposed methods are valuable options in post-selection inference.

Keywords: Data splitting; De-sparsified Lasso; Post selection inference.

1 Introduction

In the modern era when data collection has become easier, we are often challenged by high dimensional data with many different characteristics per subject. This article considers the problem of statistical inference on the treatment effect in the presence of high dimensional covariates.

Suppose that we have a random sample of n observations from the units indexed by $i = 1, \cdots, n$. For each unit, let Y_i be the outcome and D_i be the treatment variable. In addition, each unit has a vector of features, referred to as potential confounders denoted by W_i. We consider the parameter of interest α in a model of the form

$$Y_i = \alpha D_i + g(W_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | D_i, W_i) = 0, \quad i =, 1 \cdots, n, \tag{1.1}$$

where $g(\cdot)$ is an unknown real-valued function and the ε_i's are independent random errors. When the dimension of the potential confounders is small relative to n, model (1.1) has been discussed in the literature of treatment effect estimation; see Robinson (1988), Härdle et al. (2012) and, or more recently Cattaneo et al. (2016). In this article, we adopt a framework similar to that of Belloni et al. (2014). Formally, we assume that $g(W_i)$ can be well approximated by a sparse linear combination of the vector $X_i = P(W_i) \in \mathbb{R}^p$, where $P(W_i)$ is a known transformation of W_i, and then Model (1.1) can be written as

$$Y_i = \alpha D_i + X_i^{\mathrm{T}} \beta + R_{ni} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i | D_i, W_i) = 0, \quad i = 1, \cdots, n, \tag{1.2}$$

where the R_{ni}'s are approximation errors, which will be assumed to be sufficiently small, and X_i is referred to as the covariates in the subsequent analysis.

When the dimension p is greater than n, inference about α cannot be made without regularization or model selection. A major assumption we make in this paper is the sparsity in β. Formally, we require $M_0 = \mathrm{supp}(\beta) = \{j \in \{1, \ldots, p\} : \beta_j \neq 0\}$ has $s_0 \ll n$ elements. Without loss of generality, we assume in the theoretical treatment that the response variable and the covariates are all centered so that

no intercept is included in the model. In the high dimensional regime, when the approximation errors are small, inference on the treatment effect α is frequently carried out in two ways: inference after a sufficiently small model is selected, or debiased inference directly on a regularization method.

In the first part of the paper, we focus on the approach where inference is carried out on a relatively small model selected by Lasso (Tibshirani, 1996) or the smoothly clipped absolute deviation penalized maximum likelihood (Fan and Li, 2001), among many others. When perfect model selection is attained, the resulting estimate of the treatment effect achieves the oracle property, and post selection inference is asymptotically valid (Minnier et al., 2011). However, perfect model selection often relies on some unrealistically strong assumptions and inference procedures based on the belief of having an oracle estimator may result in substantial biases (Belloni et al., 2014), and see also Example 1.

Based on a selected model $\widehat{M}$, a common practice is to refit with the ordinary least squares (OLS) estimator and then perform inference on α. Since the model $\widehat{M}$ is randomly chosen, there are two possible sources of bias in the OLS estimator. The first is the under-fitting bias when an active covariate is missing in the selected model. To a large extent, the under-fitting bias can be reduced by choosing a larger model that has a high probability of $M_0 \subset \widehat{M}$. However, even if the model selection procedure retains all relevant variables, we demonstrate that the OLS estimator suffers from what we will call "over-fitting bias" when irrelevant variables are selected due to spurious correlation. The over-fitting bias is negligible in low dimensional problems, but becomes evident when p is large. This issue is not as much discussed in the literature but is recognized in Fan et al. (2012), Hong et al. (2018) and Chernozhukov et al. (2018) in a related context. An easy solution to avoid this over-fitting bias is an old idea of data-splitting. A main contribution of the paper is to introduce and examine the method of repeated data splitting, which helps minimize the efficiency loss due to data splitting or cross-estimation.

The repeated data splitting approach is similar in spirit to the bagging of Breiman

(1996). Data splitting has been used by other researchers for debiased inference. Fan et al. (2012) used the approach of data splitting and aggregation for estimating the variance of the noise, and Chernozhukov et al. (2018), Robins et al. (2017) and Wager et al. (2016) adopted the same approach to estimate the treatment effect. A key difference with our work is that these methods use non-overlapping sub-samples for parameter estimation so that the variance of the aggregated estimator is easier to handle, but the "splitting-and-aggregation" strategy is not pursued to its full potential for variance reduction. We refer such procedure as cross-estimation. Our proposed approach results in better efficiency by allowing repeated data splitting with overlapping sub-samples for estimation.

In the second part of the paper, we discuss another line of work for inference that relies on "de-sparsifying" via a two-stage selection procedure, which has been studied in van de Geer et al. (2014), and Zhang and Zhang (2014) for the high-dimensional models. We show that the de-sparsified Lasso and the post-double-selection method of Belloni et al. (2014) are asymptotically similar, and they achieve bias reduction by essentially allowing all the covariates, including the inactive ones in Model (1.2), to be used to adjust for the treatment variable first, but these approaches can lead to substantially reduced variability in the post-adjusted treatment variable. Consequentially, there can be significant efficiency loss in the estimation of α as compared to a one-stage selection procedure without adjusting for the treatment variable D_i.

While the post-double-selection estimator reduces the under-fitting bias, it does not completely avoid the risk of over-fitting. Therefore, building upon the post-double-selection estimator of Belloni et al. (2014), we discuss a projection-assisted approach to reduce the risks of the under- and over-fitting biases simultaneously. As each method has its own strength, we provide both theoretical and numerical comparisons for the those debiased inference methods.

The rest of the paper is structured as follows. In Section 2, we use a motivating example to illustrate the bias issue for inference on α by refitting the OLS to

a selected model. In Section 3, we propose repeated data splitting to eliminate the over-fitting bias. In Section 4, we discuss the relationship between the de-sparsified Lasso and the post-double-selection, and propose a new projection-assisted approach to further reduce the over-fitting bias in the post-double-selection estimator. We also identify the conditions under which the proposed estimators of the treatment effect are asymptotically normal. In Section 5, we give theoretical and numerical comparisons for several methods of debiased inference. In Section 6, we conduct simulation experiments to show the finite-sample performances of the proposed methods in comparison with several others. In Section 7, we illustrate how our proposed methods can be applied to the NCHS Vital Statistics Natality Birth Data to assess the effect of smoking on birth weight. Finally, we conclude the paper in Section 8 with some additional remarks.

2 Bias after model selection

In this section, we first formalize the notations used in the paper, and then discuss the bias issue of the OLS estimator in a selected model, followed by a special case study with a binary treatment variable $D_i \in \{0, 1\}$.

2.1 Notations

Suppose that the choice of X_i have been made, whether they are simply equal to W_i or some basis functions used to approximate $g(W)$. For $i = 1, \cdots, n$, define $Z_i = (D_i, X_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+1}$, $\mathcal{X}_i = (Y_i, D_i, X_i)$, and $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_i\}_{i=1}^n$. Also let $\boldsymbol{Z} = (Z_1^{\mathrm{T}}, \cdots, Z_n^{\mathrm{T}}) \in \mathbb{R}^{n\times(p+1)}$, $\boldsymbol{X} = (X_1^{\mathrm{T}}, \cdots, X_n^{\mathrm{T}}) \in \mathbb{R}^{n\times p}$, $D = (D_1, \cdots, D_n)^{\mathrm{T}}$, and $R_n = (R_{n1}, \cdots, R_{nn})^{\mathrm{T}}$. Suppose M is a subset of $\{1, \cdots, p\}$, and for any p-dimensional vector a, define a_M to be the sub-vector of a indexed by M, and a_{-M} to be the sub-vector of a indexed by $M^c = \{1, \cdots, p\}\backslash M$. Let $\boldsymbol{X}_M = \{X_{\cdot j}, j \in M\}$, where $X_{\cdot j}$ is the jth column of $\boldsymbol{X}$, for $j = 1, \cdots, n$, and $\boldsymbol{Z}_M = (D, \boldsymbol{X}_M)$. Let $\boldsymbol{P}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^{\mathrm{T}}$, $\boldsymbol{P}_M^* = \boldsymbol{Z}_M(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}$ be the projection matrices sending vectors in $\mathbb{R}^n$ onto the

space spanned by $\boldsymbol{X}_M$ and $\boldsymbol{Z}_M$, respectively. Also let $\boldsymbol{Q}_M = \boldsymbol{I} - \boldsymbol{P}_M$, where $\boldsymbol{I}$ is an n-dimensional identity matrix. Let the index matrix $\widetilde{\boldsymbol{I}}_M \in R^{(|M|+1)\times(p+1)}$ be such that $\widetilde{\boldsymbol{I}}_M Z_i = Z_{i,M}$. Let $e_1 = (1, 0, \cdots, 0)^{\mathrm{T}}$, whose dimension is context-specific. Furthermore, let $\widehat{\Sigma} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}/n$ be the sample covariance matrix, and $\Sigma = \mathbb{E}(Z_i Z_i^{\mathrm{T}})$ be the population covariance of the covariates, and similarly let $\Sigma_X = \mathbb{E}(X_i X_i^{\mathrm{T}})$, and $\Sigma_{DX} = \mathbb{E}(D_i X_i)$. Define Σ_M as the sub-matrix of the population covariance matrix indexed by set M, i.e. $\Sigma_M = \mathbb{E}(Z_{i,M} Z_{i,M}^T)$. We use the notation $x \lesssim_P y$ to denote $x = O_P(y)$. We use $\rightsquigarrow$ to denote the convergence in distribution. By 1_T we denote the indicator function of an event T.

2.2 Over-fitting and under-fitting bias

Based on a properly chosen data-dependent model $\widehat{M}$, the OLS estimator is

$$(\widehat{\alpha}_{\mathrm{OLS}}, \widehat{\beta}_{\mathrm{OLS}}^{\mathrm{T}})^{\mathrm{T}} = \arg\min \left\{ \sum_{i=1}^{n} (Y_i - \alpha D_i - X_i^{\mathrm{T}}\beta)^2 : \alpha \in \mathbb{R}, \beta \in \mathbb{R}^p, \beta_{\widehat{M}^c} = 0 \right\}. \quad (2.1)$$

The performance of $\widehat{\alpha}_{\mathrm{OLS}}$ is evaluated by Belloni and Chernozhukov (2013), which showed that this estimator has at least the same rate of convergence as Lasso, and has a smaller bias. To heuristically illustrate the impact of the random model $\widehat{M}$ on the estimate of α, we decompose $\widehat{\alpha}_{\mathrm{OLS}}$ as

$$\begin{aligned}\sqrt{n}(\widehat{\alpha}_{\mathrm{OLS}} - \alpha) = & \underbrace{e_1^{\mathrm{T}} \left(\frac{1}{n} \boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}} \boldsymbol{Z}_{\widehat{M}} \right)^{-1} \frac{1}{\sqrt{n}} \boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}} \varepsilon}_{:=b_{n1} \text{ (over-fitting)}} \\ & + \left(\frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}}) D \right)^{-1} \underbrace{\frac{1}{\sqrt{n}} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}})(\boldsymbol{X}\beta + R_n)}_{:=b_{n2}(\text{ under-fitting})}. \end{aligned} \quad (2.2)$$

Details about this decomposition is provided in Section S.4.1 of the Supplementary Materials. The first term b_{n1} labeled "over-fitting" is really due to the correlation between $\boldsymbol{Z}_{\widehat{M}}$ and ε. When $\widehat{M}$ is not data-dependent, b_{n1} has mean zero. Otherwise, we have in general $\mathbb{E}(\varepsilon | \boldsymbol{Z}_{\widehat{M}}) \neq 0$. In this case the bias of $\widehat{\alpha}_{\mathrm{OLS}}$ as an estimator of α is in the same order of $1/\sqrt{n}$, which would result in biased inference.

When the approximation error R_n is small, the contributor to the "under-fitting" bias, b_{n2}, vanishes to zero if $M_0 \subseteq \widehat{M}$. Wasserman and Roeder (2009), for example, provides sufficient conditions under which $\mathbb{P}(M_0 \subseteq \widehat{M}) \to 1$, as $n \to \infty$, holds for Lasso. Those conditions are much weaker than the conditions needed for the perfect model selection in the sense of $\mathbb{P}(M_0 = \widehat{M}) \to 1$. Therefore, when the estimation efficiency is not a major concern, selecting a larger model seems to be a simple remedy to avoid the under-fitting bias. Additional methods to reduce the under-fitting bias will be discussed in Section 4. Next, we illustrate the over- and under-fitting biases through an example with a sparse β.

Example 1 (A numerical study with the adaptive Lasso)**.** *We start with a simple simulation study where the adaptive Lasso is used for variable selection. Implementation details are provided in Section S.11 of the Supplementary Materials. We refer to this estimator as Alasso+OLS estimator. The data are generated from model* (1.2) *with* $R_n = 0$, $\alpha = 3$, $\beta = (1, 1, 0.5, 0.5, 0 \cdots, 0)^{\mathrm{T}} \in \mathbb{R}^{p\times 1}$, *and* $(n, p) = (100, 500)$. *We first generate a random matrix* $\widetilde{\boldsymbol{Z}} \in \mathbb{R}^{n\times(p+1)}$ *where each row is randomly drawn from* $N(0, \Sigma)$, *with* $\Sigma_{ij} = 0.9^{|i-j|}$, $(1 \le i, j \le p+1)$. *Then let* $D_i = 1(\widetilde{Z}_{i1} > 0)$ *and* $X_{ij} = \widetilde{Z}_{ij}$ *be the covariates, for* $i = 1, \ldots, n$, $j = 2, \ldots, p+1$. *If a model selection procedure is the oracle, then*

$$\mathbb{P}(\widehat{M} = M_0) \to 1, \quad \sigma_{oracle}^{-1}\sqrt{n}(\widehat{\alpha}_{OLS} - \alpha) \rightsquigarrow N(0, 1),$$

where $\sigma_{oracle}^2 = \sigma_\varepsilon^2(\Sigma_{M_0}^{-1})_{11}$, $\sigma_\varepsilon^2 = Var(\varepsilon_i)$, *and* $(\Sigma_{M_0}^{-1})_{11}$ *denotes the first diagonal element of* $\Sigma_{M_0}^{-1}$. *As the tuning parameter* λ *decreases from* $\exp(-3)$ *to* $\exp(-2)$, *we keep track of the selected model* $\widehat{M}$ *and report the standardized bias of* $\widehat{\alpha}_{OLS}$ *from the selected model* $\widehat{M}$. *In this setting,* α *is often refereed to as the average treatment effect (ATE).*

The numerical results presented in Figure 1 are evaluated though 1000 Monte Carlo samples. From Figure 1(a), we see that when λ is greater than $\exp(1)$ and some active covariates are often missed in the refitting step, leading to clear under-

fitting bias. When the tuning parameter decreases from exp(2) to exp(1), the under-fitting bias decreases quickly as more covariates are used in the ordinary least squares estimates. However, as λ decreases further to include more and more covariates in the selected model, the bias does not vanish but begins to increase in the opposite direction. By the nature of model selection, the over-selected variables are most likely highly correlated with Y in each sample. Since they account for the variability in Y in the data, the estimated coefficient on D is attenuated. In this particular example, the over-fitting bias can be as significant as the under-fitting bias, and will lead to invalid statistical inference.

From Figure 1(b), we observe clearly that perfect model selection cannot be achieved with high probability, but as λ decreases towards $\exp(-3)$, the under-fitting probability decreases rapidly toward 0; and in most of the Monte Carlo samples, the selected model $\widehat{M}$ contains M_0. If we use a small λ in the adaptive Lasso, the main issue to be concerned with is indeed the over-fitting bias for the estimation of α.

In Section S.2 of the Supplementary Materials, we give a simple example to illustrate the connection between the over-fitting bias and a better-known concept of spurious correlation.

3 Repeated data-splitting

The over-fitting bias can be avoided by the idea of data splitting. Data splitting divides a sample of size n into two parts: the model building part of size n_1 and the estimation part of size $n_2 = n - n_1$. The first part of the data is then used for model selection and the remaining part is used for estimation based on the selected model. When β is sparse and by selecting a larger model in the first part, we expect the OLS estimator from the second part of the data to be free of significant bias. Rinaldo et al. (2018) considered data splitting for debiased inference. However, it is also clear that data splitting enables debiased inference after model selection at a cost. As only part of the sample can be used in the estimation step, which means a loss

of efficiency even if a perfect model has been selected. We consider using repeated splits and then averaging the estimates of α over those splits. This strategy, similar to bagging or bootstrap aggregating proposed in Breiman (1996), is a machine learning ensemble meta-algorithm and can help improve the stability and accuracy over a single split or a small number of splits. Similar ideas based on bagging are considered in Meinshausen and Bühlmann (2010) and Meinshausen et al. (2009) for the recovery of sparse representations. We consider two data splitting schemes, repeated random splitting (R-Split), which is detailed in the paper, and bootstrap-induced splitting (B-Split), which is discussed in the Supplementary Materials.

3.1 R-Split

Based on repeated random data splitting, the estimation and inference procedure for the treatment effect α can be described as follows (Algorithm 1). First, we set an upper bound of the selected model size to ensure the existence of the OLS estimator in any given subsample. Next, the choice of model size is subjective but needs to be large enough for the under-fitting bias to be negligible. In our empirical work, we use Lasso for model selection, and choose the model size from cross-validation with an upper bound n_2 minus a small number to determine the level of penalization; we note that this can be done in standard software for regularized regression, such as the package `glmnet` in *R*.

Algorithm 1 R-Split

For $b \leftarrow 1$ to B do

Step 1. Randomly split the data $\{(Y_i, D_i, X_i)\}_{i=1}^n$ into group T_1 of size n_1 and group T_2 of size $n_2 = n - n_1$, and let $v_{bi} = \mathbf{1}_{(i \in T_2)}$, for $i = 1, \cdots, n$.

Step 2. Select a model $\widehat{M}_b$ to predict Y based on T_1.

Step 3. Refit the model with the data in T_2 to get

$(\widehat{\alpha}_b, \widehat{\beta}_b^{\mathrm{T}}) = \arg\min \sum_{j \in T_2} (Y_j - \alpha D_j - X_{j,\widehat{M}_b}^{\mathrm{T}} \beta)^2$,

The final "smoothed" estimate is $\widetilde{\alpha} = \frac{1}{B} \sum_{b=1}^B \widehat{\alpha}_b$.

In Algorithm 1, any reasonable model selection procedures may be used in Step 2. Our empirical studies suggest that the variance of the aggregated estimator is a non-increasing function of B, and the decay slows down if B grows larger than 1,000. Therefore, we recommend using $B = 1,000$ as a good balance between computational load and statistical inference accuracy. In the theoretical investigations, we consider B to be infinitely large.

Let $\mathcal{V}_{n_2} = \{V = (V_1, \cdots, V_n) \in \mathbb{R}^n : V_i \in \{0, 1\}, \sum_{i=1}^n V_i = n_2\}$ be the space of n-tuples with the l_1 norm equal to n_2. The data splitting weight $v_b = (v_{b1}, \cdots, v_{bn})$ given in Step 1 takes value in $\mathcal{V}_{n_2}$ with equal probability $\mathbb{P}(V = v_b) = 1/\binom{n}{n_2}$. For a single split, the selected model can be viewed as a function of the data $\boldsymbol{\mathcal{X}} = \{Y_i, D_i, X_i\}_{i=1}^n$ and the random weight $V \in \mathcal{V}_{n_2}$, i.e. $\widehat{M} = M(\boldsymbol{\mathcal{X}}, V)$. The proposed R-Split estimator can then be defined as the expectation of $\widehat{\alpha}_b$ given the data, that is, $\widetilde{\alpha} = \mathbb{E}(\widehat{\alpha}_b|\boldsymbol{\mathcal{X}})$.

Following a strategy proposed in Efron (2014) and the bias corrected version of Wager et al. (2014), we can estimate the variance of the smoothed estimator through the nonparametric delta method. The estimated variance takes the following form with the derivation provided in Section S.8 of the Supplementary Materials:

$$\widehat{\sigma}_n^2 = n\sum_{j=1}^n \left(\frac{n-1}{n-n_2}\widehat{S}_j\right)^2 - \frac{n_2 n^2}{B^2(n-n_2)}\sum_{b=1}^B (\widehat{\alpha}_b - \widetilde{\alpha})^2, \tag{3.1}$$

where $\widehat{S}_j = \frac{1}{B}\sum_{b=1}^B (v_{bj} - \frac{1}{B}\sum_{k=1}^B v_{kj})\widehat{\alpha}_b$. In Section 3.3, we prove under certain conditions, the smoothed estimator $\widetilde{\alpha}$ converges to a normal distribution. We can then construct an approximate $(1-q)$ level confidence interval for α via the normal approximation.

3.2 Theoretical investigation of R-Split

In this section, we study the theoretical properties of the smoothed estimator. For a fixed model M and a weight $V \in \mathcal{V}_{n_2}$, define the covariance matrix in the given sub-sample as $\widehat{\Sigma}_{V,M} = n^{-1}\sum_{i=1}^n V_i Z_{i,M} Z_{i,M}^{\mathrm{T}}$, with the notations that $Z_{i,M} = (D_i, X_{i,M}^{\mathrm{T}})^{\mathrm{T}}$. Let $\boldsymbol{Z}_V = (D_V, \boldsymbol{X}_V)$ be the design matrix with rows $\{Z_i : V_i = 1, i = 1, \cdots, n\}$

and $g_V(\boldsymbol{W}) = \{g(W_i) : V_i = 1, i = 1, \cdots, n\}$. Define the projection matrix in the given subsample as $\boldsymbol{P}_{V,M} = \boldsymbol{X}_{V,M}(\boldsymbol{X}_{V,M}^{\mathrm{T}}\boldsymbol{X}_{V,M})^{-1}\boldsymbol{X}_{V,M}^{\mathrm{T}}$. Furthermore, let $\breve{V} = (\breve{V}_1, \cdots, \breve{V}_n) \in \mathcal{V}_{n_2}$ be from another split independent of $V = (V_1, \cdots, V_n)$. Suppose $\breve{M} = M(\boldsymbol{\mathcal{X}}, \breve{V})$ is the selected model from $\breve{V}$, and $\widehat{M} = M(\boldsymbol{\mathcal{X}}, V)$ denotes the selected model from V, and let

$$\widehat{h}_{i,n} = \left\{ \mathbb{E}\left(V_i e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widetilde{\boldsymbol{I}}_{\widehat{M}} \middle| \mathcal{X} \right) - \mathbb{E}\left(V_i e_1^{\mathrm{T}} \widehat{\Sigma}_{\breve{V},\breve{M}}^{-1} \widetilde{\boldsymbol{I}}_{\breve{M}} \middle| \boldsymbol{\mathcal{X}} \right) \right\} Z_i \varepsilon_i,$$

where the expectations are taken with respect to V and $\breve{V}$ conditional on the data. It is helpful to explain the difference between the two expectations in the above definition. Note that $\breve{V}$ and V have the same distributions, and the first expectation

$$\mathbb{E}\left(V_i e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widetilde{\boldsymbol{I}}_{\widehat{M}} | \boldsymbol{\mathcal{X}} \right) = \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widetilde{\boldsymbol{I}}_{\widehat{M}} | \boldsymbol{\mathcal{X}}, V_i = 1 \right) \mathbb{P}(V_i = 1),$$

so the difference of the two expectations in the definition of $\widehat{h}_{i,n}$ is the difference in the means due to leaving the i-th observation out for the model selection step in obtaining $\widehat{M}$ but not always so in obtaining $\breve{M}$. With a change of possibly one out of n observations, the distributions of the quantities involved and their means typically change in the order of $1/n$ for most model selection methods. Assumption 3 below formalizes this for technical convenience.

Assumption 1. *Data generating process. (a). Suppose* $\{(Y_i, D_i, X_i)^{\mathrm{T}}\}_{i=1}^n$ *is a random sample, and the covariates* $Z_i = (D_i, X_i)$ *have zero mean and bounded support* $\|Z_i\|_\infty \leq C$ *for some constant* C. *(b). The error variable* ε_i *is sub-Gaussian with* $\mathbb{E}(\varepsilon_i | Z_i) = 0$ *and* $\mathbb{E}(\varepsilon_i^2 | Z_i) = \sigma_\varepsilon^2$, *for* $i = 1, \cdots, n$.

Assumption 2. *The split ratio* $r_v = n_2/n$ *is a constant in* $(0, 1)$. *The selected model sizes in all split are bounded by* s *with* $s = o(n)$.

Assumption 3. *For any candidate model* M *of size up to* s *and any* $V \in \mathcal{V}_{n_2}$, *the matrix* $\widehat{\Sigma}_{V,M}$ *is non-singular. The quantities* $\widehat{h}_{i,n}$*'s satisfy* $\sum_{i=1}^n \widehat{h}_{i,n}/\sqrt{n} = o_P(1)$.

Assumption 4. *There exists a random vector $\eta_n \in \mathbb{R}^{p+1}$ which is independent of ε, and $\|\eta_n\|_\infty$ is bounded in probability, and satisfies*

$$\left\| r_v \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widetilde{\boldsymbol{I}}_{\widehat{M}} \middle| \boldsymbol{\mathcal{X}} \right) - \eta_n \right\|_1 = o_P\left(1/\sqrt{\log p}\right)$$

Assumption 5. *There is negligible amount of under-fitting bias after averaging over all splits in the sense that*

$$\mathbb{E}\left((D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{V,\widehat{M}}) D_V/n)^{-1} \cdot D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{V,\widehat{M}}) g_V(\boldsymbol{W})/\sqrt{n} | \boldsymbol{\mathcal{X}} \right) = o_P(1).$$

Theorem 1 (Asymptotic normality of R-Split estimator)**.** *Under Assumptions 1-5, the smoothed estimator from R-Split has the following representation*

$$\sqrt{n}(\widetilde{\alpha} - \alpha) = \eta_n^{\mathrm{T}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i Z_i + o_p(1). \tag{3.2}$$

Therefore, by letting $\widetilde{\sigma}_n = \sigma_\varepsilon \left(\eta_n^{\mathrm{T}} \widehat{\Sigma}_n \eta_n \right)^{1/2}$, we have

$$\widetilde{\sigma}_n^{-1} \sqrt{n}(\widetilde{\alpha} - \alpha) \rightsquigarrow N(0,1). \tag{3.3}$$

Assumption 1 requires bounded covariates to simplify our theoretical proofs but it can be relaxed to include sub-Gaussian covariates. Assumption 2 plays a limit on the sparsity level of the model. This assumption for data splitting is weaker than the ultra-sparsity assumption needed for the post-double-selection or the de-sparsified Lasso. Assumption 3 has been discussed following the definitions of $\widehat{h}_{i,n}$ and $h_{i,n}$. Assumption 4 says that the conditional expectation of matrix $\widehat{\Sigma}_{\widehat{M}}^{-1}$ for the randomly selected model $\widehat{M}$ is asymptotically independent of the noise, regardless of which point in the sample space is conditioned on. The error rate of $1/\sqrt{\log p}$ is a weak requirement for the assumed data generating process. Assumption 5 is a high-level condition to ensure that the under-fitting bias to be small. We also note that under our model assumption $g(W) = \boldsymbol{X}_{M_0}\beta_{M_0} + R_n$, Assumption 5 implicitly requires the approximation errors R_n to be small; see Section S.6.1 of the Supplementary Materials for more details. The proof of the theorem and a cleaner version of the

influence function for the R-Split estimator are given in Section S.4.1 and Section S.4.2 of the Supplementary Materials, respectively.

As we mentioned in Section 1, cross-estimation is a limited version of repeated data splitting. Suppose $\widetilde{\alpha}_{\text{cv}}$ is the estimator obtained from two-fold (or any finitely many-fold) cross estimation, we show in Section S.4.3 of the Supplementary Materials that R-Split is more efficient than cross-estimation, that is,

$$\text{Var}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)) > \text{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha)). \tag{3.4}$$

4 A revisit to debiased inference

In this section, we start with a review of two existing methods in the high dimensional debiased inference literature. The first is the post-double-selection estimator of Belloni et al. (2014), which aims to reduce the under-fitting bias by a two-stage selection. The second is the de-sparsified Lasso of van de Geer et al. (2014) and Zhang and Zhang (2014), which removes the penalization bias of Lasso by using an estimate of the inverse population covariance matrix. In the first subsection, we highlight the connection between these two methods, and provide a comparison between their asymptotic variances. In the second subsection, we propose an improvement to the post-double-selection method by removing moderating covariates first through a linear projection to further reduce the over-fitting bias.

4.1 Connection between the post-double-selection and the de-sparsified Lasso

To estimate α without bias one must suppress the effects of extraneous variables that influence both D and Y. When $p < n$, we can do so by projecting Y and D on the the space spanned by $\boldsymbol{X}$:

$$(\boldsymbol{I} - \boldsymbol{P})Y = \alpha(\boldsymbol{I} - \boldsymbol{P})D + (\boldsymbol{I} - \boldsymbol{P})\varepsilon,$$

where $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$. Then the estimate of α is the marginal regression coefficient by regressing $(\boldsymbol{I}-\boldsymbol{P})Y$ on $(\boldsymbol{I}-\boldsymbol{P})D$:

$$\widehat{\alpha}_{\text{full}} = (\widehat{D}^{\mathrm{T}}D)^{-1}\widehat{D}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\text{full}}), \tag{4.1}$$

where $\widehat{D} = (\boldsymbol{I}-\boldsymbol{P})D$, and $\widehat{\beta}_{\text{full}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}Y$. For the cases with $p \gg n$, the sample covariance matrix is singular, and the de-sparsified Lasso and the post-double-selection offer two strategies to remove the confounding effects from $\boldsymbol{X}$.

The post-double-selection estimator of Belloni et al. (2014) goes as follows. First, a set of control variables, indexed by $\widehat{M}_D$, that are useful for predicting D is selected. Second, the variables indexed by $\widehat{M}_Y$ are selected to predict Y. Then, α is estimated by refitting the model $\widehat{M} = \widehat{M}_D \cup \widehat{M}_Y$ with the OLS. The post-double-selection estimator can be written as

$$\widehat{\alpha}_{\text{double}} = (\widehat{D}_{\widehat{M}}^{\mathrm{T}}D)^{-1}\widehat{D}_{\widehat{M}}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\widehat{M}}), \tag{4.2}$$

where $\widehat{D}_{\widehat{M}} = D - \boldsymbol{X}\widehat{\gamma} = (I - \boldsymbol{P}_{\widehat{M}})D$ is the residual of D after controlling for the effect in $\boldsymbol{X}_{\widehat{M}}$, and $\widehat{\gamma} \in \mathbb{R}^p$ is a sparse vector with $\widehat{\gamma}_{\widehat{M}} = (\boldsymbol{X}_{\widehat{M}}^{\mathrm{T}}\boldsymbol{X}_{\widehat{M}})^{-1}\boldsymbol{X}_{\widehat{M}}D$ and $\widehat{\gamma}_{-\widehat{M}} = 0$. As shown in Section S.5.3.1 of the Supplementary Materials, we have

$$\breve{\sigma}_n^{-1}\sqrt{n}(\widehat{\alpha}_{\text{double}} - \alpha) \rightsquigarrow N(0,1), \quad \breve{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{1}{\|D - \boldsymbol{X}\widehat{\gamma}\|_2^2/n} + o_P(1). \tag{4.3}$$

The de-sparsified Lasso estimator of α removes the penalization bias by finding an estimate $\widehat{\boldsymbol{\Theta}}$ of the inverse of the population covariance matrix. If we focus only on the estimation of α, we simply need $e_1^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}$. One way to get there is to let $e_1^{\mathrm{T}}\widehat{\boldsymbol{\Theta}} = \widehat{\nu}^{-2}(1, -\widehat{\gamma}_{\text{lasso}}^{\mathrm{T}})$, where $\widehat{\nu}^2 = \widehat{D}_{\text{lasso}}^{\mathrm{T}}D/n$,

$$\widehat{D}_{\text{lasso}} = D - \boldsymbol{X}\widehat{\gamma}_{\text{lasso}}, \quad \text{and } \widehat{\gamma}_{\text{lasso}} = \arg\min_{\gamma\in\mathbb{R}^p} \frac{1}{n}\sum\nolimits_{i=1}^{n}(D_i - X_i^{\mathrm{T}}\gamma)^2 + \lambda_d\|\gamma\|_1$$

for some tuning constant λ_d. As shown in Section S.5.3.1 of the Supplementary Materials, the de-sparsified Lasso estimator can be written as

$$\widehat{\alpha}_{\text{desparse}} = (\widehat{D}_{\text{lasso}}^{\mathrm{T}}D)^{-1}\widehat{D}_{\text{lasso}}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\text{lasso}}). \tag{4.4}$$

Under certain regularity conditions, as in Remark 2.1 of van de Geer et al. (2014), we have

$$\ddot{\sigma}_n^{-1}\sqrt{n}(\widehat{\alpha}_{\text{desparse}} - \alpha) \rightsquigarrow N(0,1), \quad \ddot{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{\|D - \boldsymbol{X}\widehat{\gamma}_{\text{lasso}}\|_2^2/n}{(\|D - \boldsymbol{X}\widehat{\gamma}_{\text{lasso}}\|_2^2/n + \lambda_d\|\widehat{\gamma}_{\text{lasso}}\|_1)^2}. \tag{4.5}$$

With a suitable choice λ_d in the order of $\sqrt{\log p/n}$, and with ultra-sparsity of $s_0 = o(\sqrt{n}/\log p)$, we have $\lambda_d\|\widehat{\gamma}_{\text{lasso}}\|_1 = o(1)$. Then, the variance $\ddot{\sigma}_n^2$ can be compared with that of the post-double-selection estimator in (4.3).

It follows from (4.2) and (4.4) that the post-double-selection estimator and the de-sparsified Lasso estimator are similar, except that the residuals of D (after adjusting for $\boldsymbol{X}$) are obtained differently. Following Belloni et al. (2014), we find it helpful to view γ as a regression coefficient of the following model

$$D = \boldsymbol{X}\gamma + \nu, \quad \mathbb{E}(\nu|\boldsymbol{X}) = 0, \quad \text{Cov}(\nu) = \sigma_\nu^2\boldsymbol{I}, \tag{4.6}$$

for some constant σ_ν^2. A good estimation of γ helps reduce the under-fitting bias b_{n2} in (2.2). Moreover, in a special case that $p < n$, $\lambda_d = 0$ and $\widehat{M} = \{1, \cdots, p\}$, the de-sparsified Lasso and the post-double-selection are equivalent to (4.1), which is the full model OLS estimator. Without loss of generality, we refer to the method that selects the variables to predict D as a two-stage selection estimator. Usually, the two-stage selection estimator requires ultra-sparsity to achieve asymptotic normality of the estimator; see Jankova et al. (2018) for more discussion.

Though a good estimate of γ helps reduce the bias after model selection, it may increase the variability, and vice versa. To see this, we note that if $\lambda_d\|\widehat{\gamma}_{\text{lasso}}\|_1$ in (4.5) is of $o(1)$ and the $\widehat{\gamma}_{\text{lasso}} \approx \widehat{\gamma}$ under the ultra-sparsity, the de-sparsified Lasso estimator of α is first-order equivalent to the post-double-selection. However, if we use a larger penalty term so that $\lambda_d\|\widehat{\gamma}_{\text{lasso}}\|_1$ is no longer negligible, the de-sparsified Lasso estimator of α will have a smaller variance. On the other hand, if $\widehat{D}_{\text{lasso}}$ does not remove the part of $\boldsymbol{X}$ that correlates with D, the de-sparsified Lasso will then have a bias. This bias-vaiance trade-off plays an important role in assessing the quality of inference from the the two-stage selection estimators.

To further address the bias issue in the two-stage selection method, we propose to add a projection onto double-selection (PODS) as an enhancement of the post-double-selection method of Belloni et al. (2014).

4.2 Projection onto double-selection (PODS)

In the post-double-selection method, the selected set of covariates $\widehat{M}$ aims to include those variables that are correlated with either Y or D to reduce the under-fitting bias, but it potentially increases the risk of over-fitting. As we observe from the simulation study in Section 2.2, the over-fitting bias tends to be an increasing function of the selected model size. We find that a simple remedy based on linear projections can help, with which the covariates that have spurious correlation with D are less likely to enter $\widehat{M}$ and thus the risk of over-fitting is reduced.

In the post-double-selection, suppose for the moment that $\widehat{M}_D \cap \widehat{M}_Y = \emptyset$ and $M_0 = \emptyset$, then the over-fitting term b_{n1} can be decomposed

$$b_{n1} = \frac{1}{k_{n1}} \underbrace{\frac{1}{\sqrt{n}} D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})\varepsilon}_{(I)} + \frac{1}{k_{n1}} \underbrace{\frac{1}{n} D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})\boldsymbol{X}_{\widehat{M}_Y}}_{(II)} \cdot (\boldsymbol{X}_{\widehat{M}_Y}^{\mathrm{T}} \boldsymbol{X}_{\widehat{M}_Y}/n)^{-1} \cdot \underbrace{\frac{1}{\sqrt{n}} \boldsymbol{X}_{\widehat{M}_Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})\varepsilon}_{(III)}, \quad (4.7)$$

where $k_{n1} = D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}})D/n$ is a scaler. In (4.7), (I) is a random variable of zero mean since $\widehat{M}_D$ is selected independent of ε; the product of (II) and (III) captures the main effect of the over-fitting bias and is generally not centered around 0. A careful examination of the bias decomposition suggests that, if D is uncorrelated with the over selected variables in $\widehat{M}_Y$, the over-fitting bias can be reduced to a smaller scale. This motivates our proposed method of projection onto double-selection (PODS).

A formal algorithm of PODS is given in Algorithm 2, where we do not specify the model selection procedure, which is similar to the post-double-selection. In our empirical studies, we use marginal screening (in an example provided in Section

S.5.1 of the Supplementary Materials), Lasso, or iterated Lasso, which is a tuning free method discussed in Belloni et al. (2014). Recall for a fixed model M, we define $\boldsymbol{P}^*_M = \boldsymbol{Z}_M(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}$.

Algorithm 2 PODS

Step 1. Select a set of variables $\widehat{M}_D$ for regressing D on $\boldsymbol{X}$.

Step 2. Construct the post-projection variables:

$$Y^* = (\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})Y, \quad \boldsymbol{X}^* = (\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})\boldsymbol{X}_{-\widehat{M}_D}.$$

Step 3. Select a model $\widehat{M}^*_Y$ for regressing Y^* on $\boldsymbol{X}^*$.

Step 4. Regress Y on D and $\boldsymbol{X}_{\widehat{M}^*} = \boldsymbol{X}_{\widehat{M}_D \cup \widehat{M}^*_Y}$ to get $\widehat{\alpha}$, which is the estimated coefficient of D.

In Step 1, we select a set of variables $\widehat{M}_D$ which are associated with D. In Step 2, to remove the components associated with D, we project $(Y, \boldsymbol{X})$ onto a space which is orthogonal to the space spanned by D and $\boldsymbol{X}_{\widehat{M}_D}$. By doing this, the additional variables selected in Step 3 are expected to have low correlation with D, and then the over-fitting bias can be controlled.

To better understand the difference between PODS and the post-double-selection, we provide a simple example that shows the difference between the distributions of $\widehat{M}^*_Y$ and $\widehat{M}_Y$ in Section S.5.1 of the Supplementary Materials. It is worth noting that the linear projection approach is also adopted in the correlated projection screening (CPS) method proposed by Lan et al. (2016). But CPS does not select the controls for predicting Y, and α is estimated via refitting the model $\widehat{M}_D$. Without including control variables in $\widehat{M}^*_Y$, the estimator of α can be less efficient than PODS.

Next, we provide a theoretical investigation for PODS. Some additional notations are introduced for convenience. For a model M, define the sample partial covariance

$$\widehat{\rho}_{D,j}(M) = \widehat{\rho}_{D,j} - \widehat{\Sigma}_{D,M}\widehat{\Sigma}_M^{-1}\widehat{\Sigma}_{M,j},$$

between D and X_j with $j \notin M$, where $\widehat{\rho}_{D,j} = D^{\mathrm{T}}X_j/n$, $\widehat{\Sigma}_{D,M} = D^{\mathrm{T}}\boldsymbol{X}_M/n$, $\widehat{\Sigma}_M = \boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M/n$, and $\widehat{\Sigma}_{M,j} = \boldsymbol{X}_M^{\mathrm{T}}X_j/n$. If the covariates have zero mean, $\widehat{\rho}_{D,j}(M)$ is

similar to the sample covariance between D and X_j conditional on X_M. Additionally, let $\widetilde{M}_D = \widehat{M}_D \cup (M_0 \cap \widehat{M}_Y^*)$, let $g(\boldsymbol{W}) = (g(W_1), \cdots, g(W_n))^{\mathrm{T}}$ be the vector of the nonparametric functions g for n individuals, and define the minimal s-sparse eigenvalue of a semi-positive definite matrix A as

$$\lambda_{\min,s}(A) = \min_{1 \le \|\nu\|_0 \le s} \frac{\nu^{\mathrm{T}} A \nu}{\nu^{\mathrm{T}} \nu}.$$

Assumption 6. *The selected model from PODS satisfies* $\max_{j \in \widehat{M}_Y^* \setminus M_0} |\widehat{\rho}_{D,j}(\widetilde{M}_D)| = O_P(\sqrt{\log p/n})$.

Assumption 7. *The cardinality of* $\widehat{M}_Y^*$ *is of the same order as* s_0*, which satisfies* $s_0 \log p = o(\sqrt{n})$.

Assumption 8. *There exists a positive constant* κ_2 *such that* $\lim_{n\to\infty} \mathbb{P}(\lambda_{\min,s_d+s_0}(X^{\mathrm{T}}X/n) \ge \kappa_2) = 1$*, where* s_d *is the cardinality of* $\widehat{M}_D$.

Assumption 9. *The under-fitting bias is small in the sense:* $D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\mathbf{W}) = o_p(\sqrt{n})$.

Theorem 2 (Asymptotic normality of PODS)**.** *Under Assumption 1 and Assumptions 6-9, we have*

$$\sqrt{n}(\widehat{\alpha} - \alpha) = \left(\frac{1}{n} D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D\right)^{-1} \frac{1}{\sqrt{n}} D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\varepsilon + o_P(1)$$

and $\breve{\sigma}_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha) \rightsquigarrow N(0,1)$*, where* $\breve{\sigma}_n^2 = \sigma_\varepsilon^2/(D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D/n)$.

Assumption 6 requires that the maximum sample partial covariance between D and the over-selected variables be of the order $\sqrt{\log p/n}$ after controlling for the effect in $\widehat{M}_D \cup M_0$. This condition is rather mild since $\widehat{M}_Y^*$ is selected after removing the effect of D and $\boldsymbol{X}_{\widehat{M}_D}$. Assumption 7 restricts the sparsity level of β and the selected model size. Assumption 8 is quite plausible for many designs of interest. For example as shown in Rudelson and Zhou (2012), when the X_i's are i.i.d. bounded centered random vectors, then the sample covariance has minimal $(s \log n)$-sparse

eigenvalues that are bounded above by a positive constant with probability goes to 1. This assumption says that, unlike the treatment in Belloni et al. (2014), PODS no longer requires (4.6) to be true or γ to be ultra-sparse. Assumption 9 assumes a negligible under-fitting bias. In a boarder context, Chernozhukov et al. (2018) assumed a similar condition.

We show in Section S.5.3 of the Supplementary Materials that under some additional assumptions, the asymptotic variance of $\widehat{\alpha}$ from PODS can be consistently estimated by $\widetilde{\sigma}_n^2 = \frac{\widehat{\sigma}_\varepsilon^2}{\|D - \boldsymbol{X}\widehat{\gamma}^*\|_2^2/n}$, where $\widehat{\sigma}_\varepsilon^2 = Y^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}^*})Y \cdot n/(n - |\widehat{M}^*| - 1)$, $\widehat{\gamma}^* \in \mathbb{R}^p$ is a sparse vector with $\widehat{\gamma}^*_{\widehat{M}^*} = (\boldsymbol{X}^{\mathrm{T}}_{\widehat{M}^*}\boldsymbol{X}_{\widehat{M}^*})^{-1}\boldsymbol{X}^{\mathrm{T}}_{\widehat{M}^*}D$ and $\widehat{\gamma}^*_{-\widehat{M}^*} = 0$. Note that PODS is an enhancement of the post-double-selection to further reduce the over-fitting bias by modifying the distribution of $\widehat{M}^*$. As a result, the asymptotic expression in Theorem 2 and the variance estimation for PODS also apply to the post-double-selection estimator.

5 Comparison between the one-stage and the two-stage selection methods

In Sections 3 and 4, we have considered one-stage and two-stage selection methods for debiased inference. The purpose of this section is to compare the statistical efficiencies between one- and two-stage selection methods for making inference on the parameter α. Since various two-stage selection methods have similar asymptotic representations, we use PODS as a representative in the discussion.

To compare the asymptotic behavior of R-Split with PODS more explicitly, we provide alternative asymptotic variance expressions of R-Split and PODS estimators under additional assumptions. First, in R-Split, we assume that on average, the maximum "correlation" between D and $\boldsymbol{X}$ after controlling for the effects in $\boldsymbol{X}_{\widehat{M}}$ is

bounded above by $\sqrt{\log p}$ in probability, or more formally,

$$\left\| \mathbb{E}\left\{ \frac{D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})\boldsymbol{X}_V/n}{D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V/n} \middle| \boldsymbol{\mathcal{X}} \right\} \right\|_\infty = O_P(\sqrt{\log p}). \tag{5.1}$$

Second, let the maximal s-sparse eigenvalue of a semi-positive definite matrix A as

$$\lambda_{\max,s}(A) = \max_{1\le\|\nu\|_0\le s} \frac{\nu^{\mathrm{T}} A \nu}{\nu^{\mathrm{T}}\nu},$$

and we assume that there exists some constant $K_0 > 0$ such that $\mathbb{P}(\limsup_{n\to\infty} \lambda_{\max,s}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n) \le K_0) = 1$, $\lambda_{\max,s}(\Sigma) \le K_0$, and the maximum eigenvalue of $\widehat{\Sigma}$ is bounded by $\log p$ in probability. Under Assumptions 1, 4, and 7, and the two additional assumptions stated above, we show in Section S.4.4 of the Supplementary Materials that the asymptotic variance of R-Split estimator satisfies

$$\widetilde{\sigma}_n^2 \le \sigma_\varepsilon^2 \mathbb{E}\left\{ (\Sigma_{\widehat{M}}^{-1})_{11} \middle| \boldsymbol{\mathcal{X}} \right\} + o_P(1), \tag{5.2}$$

where $(\Sigma_{\widehat{M}}^{-1})_{11}$ is the first component on the diagonal of $\Sigma_{\widehat{M}}^{-1}$.

As for PODS, under Model (4.6), with the assumption that γ is ultra-sparse and that the selected model $\widehat{M}^*$ satisfies $\|n^{1/4}(I - \boldsymbol{P}_{\widehat{M}^*})D\|_2 = o_P(1)$, the asymptotic variance of PODS estimator equals

$$\breve{\sigma}_n^2 = \sigma_\varepsilon^2(\Sigma^{-1})_{11} + o_P(1) = \sigma_\varepsilon^2/\sigma_\nu^2 + o_P(1). \tag{5.3}$$

Together with the theoretical results in Theorem 2.3 of van de Geer et al. (2014) and Theorem 2 of Belloni et al. (2014), PODS, the-desparisified Lasso and the post-double-selection estimators reach the semi-parametric efficiency bound for estimating α under homoscedasticity (see Robinson, 1988). However, when σ_ν is small, (5.3) indicates that the two-stage selection method is not very efficient. From the comparison between (5.2) and (5.3), we find that unless $\mathbb{E}\{\|\Sigma_{D,\widehat{M}^c} - \Sigma_{D,\widehat{M}}\Sigma_{\widehat{M}}^{-1}\Sigma_{\widehat{M},\widehat{M}^c}\|_2^2|\boldsymbol{\mathcal{X}}\} \approx 0$, the R-Split estimator has the smaller asymptotic variance than the two-stage selection estimators, which is not surprising since R-Split aims to work with a sparse model while the two-stage selection estimators are about bias-correction based on all the covariates.

To provide some numerical evidence for the comparison between one- and two-stage selection estimators, consider the following model

$$Y_i = \alpha D_i + \varepsilon_i, \tag{5.4}$$
$$D_i = \gamma_1 X_{i1} + \nu_i,$$

where $(\varepsilon_i, \nu_i/\sigma_\nu) \sim N(0, \boldsymbol{I}_2)$ with $\sigma_\nu = \sqrt{\text{Var}(\nu)}$ and $\boldsymbol{I}_2$ the 2 by 2 identity matrix, and X_{i1} is the first component of X_i, for $i = 1, \cdots, n$. Let $p = 500$, $\alpha = 1$, $\gamma_1 = 1$ and $(D_i, X_i) \sim N(0, \Sigma)$ independent of (ε_i, ν_i), and set σ_ν^2 as an increasing sequence from 0 to 1. The implementation details of various methods under comparison are provided in Section 6. For $n = 100$ and $n = 400$, we report $\sqrt{n}$ times bias and n times variance evaluated from Monte Carlo samples, and the results are provided in Figure 2. The variance of the oracle estimator is provided as a benchmark.

The results in Figure 2 indicate that R-Split is not as efficient as the oracle estimator, but has smaller variance than PODS and the de-sparsified Lasso. While the performance of R-Split is not sensitive to the change in σ_ν^2, the variances of PODS and the de-sparsified Lasso increase rapidly as σ_ν^2 becomes smaller. Furthermore, in the de-sparsified Lasso, we observe that although the penalization helps reduce the estimation variability, it increases the bias. The numerical results are in-line with our investigation about the bias-variance trade-off in Section 4.1.

Although R-Split tends to have better estimation efficiency, the fact that only a fraction of the sample is used for model selection increases the risk of under-fitting. While the concern of the under-fitting bias can be lessened via the use of the two-stage selection, the combination of R-Split and PODS or the post-double-selection may be used as an alternative approach. We summarize the combined approach by using R-Split in Algorithm S.7.1 that has been provided in Section S.7 of the Supplementary Materials. However, the combined approach inherits the inflated variance problem from the two-stage selection when D is highly correlated with some of the covariates. A similar idea of combining data splitting with the two-stage selection method has been studied in Chernozhukov et al. (2018), but their proposal

uses non-overlapping subsamples for parameter estimation so that the variance of the aggregated estimator can be easily estimated. Consequently, the combination of R-Split and two-stage selection can have smaller variance than cross-estimation. In Section 6, we further illustrate this point in a simulation study.

6 Simulation study

This section reports finite sample performances of the proposed methods in comparison with several others through Monte Carlo simulations. From an empirical study provided in Section S.3.1 of the Supplementary Materials, we find R-Split and B-Split have similar performances whenever $r_v \in [0.6, 0.7]$, which is typically a favorable range for r_v. In the simulation study, we focus on the performance of R-Split with $r_v = 0.7$.

6.1 Simulation designs

We compare the performances of the proposed methods with several others in two different simulation settings where β_0 is one of the following vectors,

$$\text{sparse: } (1,1,1,1,0,\cdots,0), \quad \text{dense: } (1,1/\sqrt{2},\cdots,1/\sqrt{p}),$$
$$\text{moderately sparse: } (\underbrace{5,\cdots,5}_{10},\underbrace{1,\cdots,1}_{10},0,\cdots,0),$$

and γ_0 is either $(0,0,0,0,1,1,1,1,0,\cdots,0)$ or dense as specified later.

Stetting 1. Similar to the classical model used in van de Geer et al. (2014), we have $Y_i = a + \alpha D_i + X_i^{\mathrm{T}}\beta + \varepsilon_i$ for $i = 1,\cdots,n$, where $(D_i, X_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+1} \sim N(0,\Sigma)$, $\varepsilon_i \sim N(0,1)$ are white noise, $a = 1$ is the intercept, $\alpha = 1.5$, $\beta = c_y\beta_0 \in \mathbb{R}^p$ with the constant $c_y \in \mathbb{R}$ chosen to achieve $R^2 = 0.8$, and Σ has one of the following forms:

$$\text{Independent:} \quad \Sigma = \boldsymbol{I}_p, \quad \text{Toeplitz:} \quad \Sigma_{jk} = 0.9^{|j-k|},$$
$$\text{Equal correlation:} \quad \Sigma_{jk} = 0.9^{\mathbf{1}(j\neq k)} \text{ or } 0.3^{\mathbf{1}(j\neq k)},$$

where Σ_{jk} is the (j,k)-th element of the matrix Σ for $j = 1, \cdots, p+1$ and $k = 1, \cdots, p+1$.

Setting 2. Consider the two-stage model used in Belloni et al. (2014), with $Y_i = a_y + \alpha D_i + X_i^{\mathrm{T}}\beta + \varepsilon_i$, and $D_i = a_d + X_i^{\mathrm{T}}\gamma + \nu_i$, for $i = 1, \cdots, n$, where $(\nu_i, \varepsilon_i) \sim N(0, I_2)$ are 2-dimensional white noise, $a_y = 1$ and $a_d = 0.5$ are the intercepts, $(D_i, X_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{p+1} \sim N(0, \Sigma)$ with $\Sigma_{jk} = 0.9^{|i-j|}$, $\alpha = 1.5$, $\beta = c_y\beta_0 \in \mathbb{R}^p$ and $\gamma = c_d\gamma_0 \in \mathbb{R}^p$, with the constants c_y and c_d chosen for designed signal-to-noise ratios of both components in the model as detailed in Table 2.

We include the following methods in the comparisons.

- "Oracle" refers to the oracle estimator based on the true model, and is used when β is sufficiently sparse.
- "Double" represents the post-double-selection of Belloni et al. (2014) and is implemented using the R packages `hdm`.
- "Double-2CV" represents the double-machine with two fold cross-estimation of Chernozhukov et al. (2018): for each fold, we select the model using the package `hdm`, and estimate the treatment effect and its variance from the remaining data.
- "PODS" refers to the proposed method PODS with the model selected from the function `rlasso` in the R package `hdm`, which is the same function for model selection used by the post-double-selection in `hdm`.
- "R-Split" refers to the proposed smoothed estimators from R-Split with $B = 1,000$. We select the model by the adaptive Lasso via package `glmnet`. The tuning parameter λ is selected by cross-validation with the `lamdba.min` option, while the maximum model size (`dfmax` in `glmnet`) is at most $n_2 - 6$. Since R-Split requires a large model to avoid the under-fitting, we also specify

a minimum model size $\widehat{s}_{\min}$ given in Table 1-2. The implementation details of the adaptive Lasso is provided in Section S.11 of the Supplementary Materials.

- "PODS-Split" is the combined approach we discussed at the end of Section 5. Its implementation is similar to R-Split, except that the minimum and maximum model sizes equal $\widehat{s}_{\min}/2$ and $n_2/2-3$ in each stage of model selection.

- "De-sparsified" represents the de-sparsified Lasso of van de Geer et al. (2014) and Zhang and Zhang (2014), and is implemented using the R package `hdi`.

- "Alasso+OLS" refers to the method of ordinary least squares applied to a model selected by Adaptive Lasso. The confidence intervals are constructed based on normal approximations.

The performance measures used in this section include $\sqrt{n}$ times bias, n times mean squared error, coverage probability and average length of the confidence intervals of the treatment effect α. The details about the dimension and the covariance structure of the covariates are provided in the captions of the accompanying tables.

6.2 Results

In this subsection, we provide the finite sample comparisons in our simulation studies through Tables 1 and 2, one for each setting.

Table 1 for Setting 1 shows that R-Split is an overall leader for sparse models in terms of bias, efficiency, and validity of inference, but provably due to under-fitting bias, the estimator can underperform for dense or sometimes moderately sparse models. In those cases, PODS-Split does well by reducing the bias and delivering confidence intervals with the desired coverage. PODS helps reduce the bias of post-double-selection estimators. The refitted estimator from Alasso+OLS is centered away from α, and the asymptotic approximation provides a very poor guide to the finite-sample distribution of this estimator. The post-double-selection with 2-fold cross estimation avoids the over-fitting bias, but is not as efficient as R-Split

or PODS-Split. The de-sparsified Lasso estimator often has smaller variance than others, but it is not as satisfactory in terms of the coverage of the resulting interval estimates, mainly due to bias, which is in line with our analysis in Section 4.1.

From the results in Table 2 for Setting 2, we see the same message that R-Split does well for sparse models, and equally noteworthy is that R-Split has substantially smaller variances than the two-stage selection methods whenever R_d^2 is high, that is, when the treatment D_i is well correlated with some of the covariates. On the other hand, when both γ and β dense, all the methods perform poorly in the coverage of the interval estimates. Overall, the relative performance of each method depends on the sparsity of the underlying model, but repeated data splitting and PODS are two promising additions to the toolkit of debiased inference on the treatment effect in a high dimensional setting.

7 Real data example

In this section we illustrate the use of the proposed methods by examining the effect of mother's smoking on infant birth weight. Lumley et al. (2000) confirmed the existence of a causal relationship between smoking cessation during pregnancy and birth weight in randomized trials. Here we use the regression analysis on an observational study to adjust for the potential confounders, as done in Nijiati et al. (2008).

To study the effect of smoking on infant birth weight, we use the 2015-2016 Natality data from the National Vital Statistics System of Centers for Disease Control and Prevention. To illustrate the utility of the proposed methods, we consider only live, singleton births to Asian mothers between the age of 18 and 45, with no more than 2 years of college education in the United State. This results in a data set of 59,250 births in 2015, and 58,785 births in 2016 with fully observed variables in this study, and each data set contains 217 main-effects variables. To avoid handpicking important interaction terms to be included in the model, we introduce all possible 12,543 interaction terms and then screen out the unimportant ones by model selection. The

screening procedure is carried out on the 2015 data so that the selected set of variables are independent of the 2016 data. As an implementation detail, we replace several continuous variables (mother's age, height, weight gain during pregnancy, and pre-pregnancy weight) with their spline basis functions. After Lasso screening (with the tuning parameter chosen by cross validation with the `lamdba.min` option of package `glmnet`), we control for the father's age and race, infant's sex, plurality, infant's birth defects, infant's Apgar score, the obstetric estimate of gestation, induction of labor, admission to NICU, mother's pre-pregnancy weight, mother's weight gain during pregnancy, mother's height, and several variables that indicate complications during pregnancy, and some interaction terms between these selected features. In total we keep $p = 630$ variables.

With the year 2016 data, since the sample size $58,785$ is much larger than p, we use the OLS estimate of the treatment effect from this full sample as a benchmark in the investigation. From fitting the full sample with a linear regression model, 47.56% of the variance of the infant birth weight can be explained by the selected 630 variables. The results regarding the infant birth weight by using the full sample show evidence that, on average, women who were self-reported smokers delivered infants weighting 80.33g less than the others. Our goal is to compare the performances of the existing methods for estimating the treatment effect based on randomly drawn subsamples of size n_{sub} from the 2016 data. Since only 2.06% of mothers were reported to have smoked during pregnancy, to have a more balanced group, we first draw $n_{\text{sub}}/2$ observations from the mothers who smoke during pregnancy, and draw another $n_{\text{sub}}/2$ observations from the remaining sample.

Since the performances of the de-sparsified Lasso and the post-double-selection are similar to PODS in this particular study, we include the results only for PODS, R-Split, PODS-Split and Alasso+OLS in Figure 3. We observe that R-Split and PODS with R-Split have relatively small mean squared errors, and the confidence intervals obtained via the non-parametric delta method achieve near nominal coverage probabilities. PODS gets reasonable coverage and improves with the sample size.

The "Alasso+OLS" estimator has low coverage due to bias after model selection, and the asymptotic approximation provides a very poor guide to the finite-sample distribution. Overall, the use of R-Split, whether used alone or together with PODS, would help the inference in this study at sample sizes below 300.

8 Concluding remarks

This paper addresses the issue of bias after model selection and its impact on statistical inference on treatment effects from a linear or partially linear model in a high dimensional setting. We consider the method of repeated data splitting to remove the over-fitting bias without much sacrifice in efficiency. We revisit some of the well-known two-stage selection estimators and discuss a delicate bias-variance trade-off with those methods. As made clear in the paper, there are pros and cons in each method. While the method of repeated data splitting eliminates the over-fitting bias and helps minimize the efficiency loss, it is subject to the risk of under-fitting, especially in a non-sparse model. The two-stage selection methods reduce the under-fitting bias but at the cost of efficiency loss when the treatment variable is correlated with some of the inactive covariates in the model. In the latter cases, we propose a new variant, PODS, that aims to suppress the over- and under-fitting biases simultaneously. Our theoretical and empirical investigations show that the proposed methods improve the validity of inference on the treatment effect in a high dimensional regression model.

In a special case where $D_i \in \{0, 1\}$ represents the treatment indicator, we may study the average treatment effect following the Neyman-Rubin's causal model; see Neyman (1923) and Rubin (1974). We refer to Section S.9 of the Supplementary Materials for further details.

Supplementary Materials

Section S.1 gives some useful lemmas needed in the proofs. Section S.2 gives an additional example to illustrate the over-fitting bias issue. Section S.3 details bootstrap-induced splits (B-Split), which is an alternative to R-Split, and offers a finite sample comparison between R-Split and B-Split. Section S.4 contains the proofs and derivations needed in Sections 3.2 and 5. Section S.5 contains the proofs for PODS discussed in Sections 4.2 and 5. Section S.6 gives sufficient conditions for Assumptions 5 and 9. Section S.7 contains the implementation details for the combined approach of R-Split and the post-double-selection method mentioned at the end of Section 5. Section S.8 contains the derivation of variance estimation in R-Split via the non-parametric delta method used in Section 3.1. In Section S.9, we discuss how regression adjustment for the average treatment effect estimation can be handled in our framework. In Section S.10, we provide an additional set of results for the analysis in Section 7 based on polynomial basis function expansions instead of spline basis expansions. Finally, Section S.11 contains the implementation details of the adaptive Lasso used in Example 1 of Section 2.2 and the simulation studies of Sections 6.

Acknowledgments

The authors acknowledge the support of NSF Awards DMS-1607840, DMS-1712717, SES-1659328, and National Natural Science Foundation of China Project 11690012. The authors thank Matias Cattaneo and Jonathan Taylor for helpful discussions, and anonymous referees and the Associate Editor for their constructive comments.

References

Alexandre Belloni and Victor Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

Matias D Cattaneo, Michael Jansson, and Whitney K Newey. Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, 34(2):1–25, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Bradley Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456): 1348–1360, 2001.

Jianqing Fan, Shaojun Guo, and Ning Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65, 2012.

Wolfgang Härdle, Hua Liang, and Jiti Gao. *Partially linear models*. Springer Science & Business Media, 2012.

Liang Hong, Todd A Kuffner, and Ryan Martin. On overfitting and post-selection uncertainty assessments. *Biometrika*, 105(1):221–224, 2018.

Jana Jankova, Sara Van De Geer, et al. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

Wei Lan, Ping-Shou Zhong, Runze Li, Hansheng Wang, and Chih-Ling Tsai. Testing a single regression coefficient in high dimensional linear models. *Journal of econometrics*, 195(1):154–168, 2016.

Judith Lumley, Sandy Oliver, and Elizabeth Waters. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database Syst Rev*, 2:CD001055, 2000.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488): 1671–1681, 2009.

Jessica Minnier, Lu Tian, and Tianxi Cai. A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association*, 106(496):1371–1382, 2011.

Jerzy S Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

Keyoumu Nijiati, Kenichi Satoh, Keiko Otani, Yukie Kimata, and Megu Ohtaki. Regression analysis of maternal smoking effect on birth weight. *Hiroshima journal of medical sciences*, 57(2):61–67, 2008.

Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv:1611.05401*, 2018.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56 (4):931–954, 1988.

Donald B Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651, 2014.

Stefan Wager, Wenfei Du, Jonathan Taylor, and Robert J Tibshirani. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

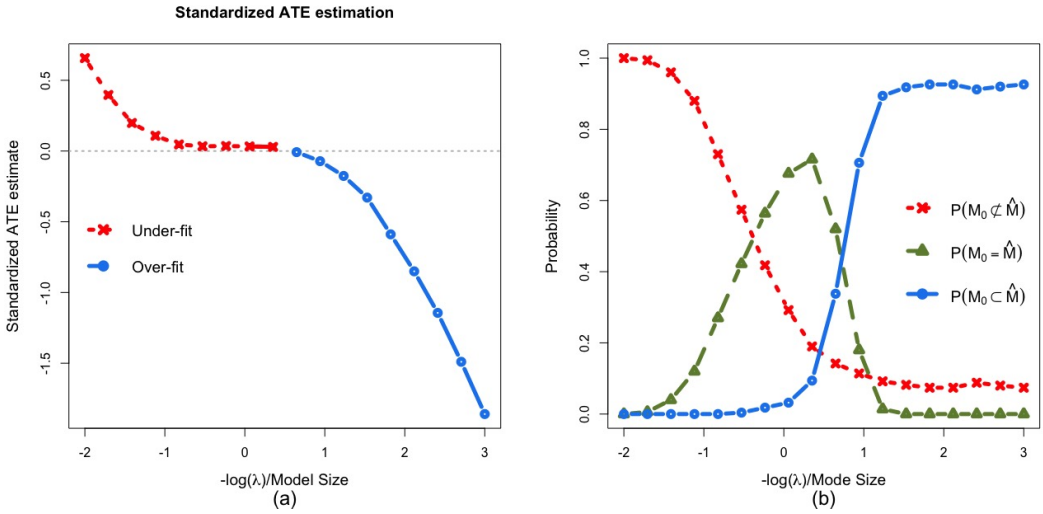

Figure 1: (a) The left panel shows standardized bias of Alasso+OLS estimator as the tuning parameter λ varies from $\exp(2)$ to $\exp(-3)$. The horizontal axis is $-\log(\lambda)$ as a measure of model size. (b) The right panel shows the probabilities of under-fitting $M_0 \not\subset \widehat{M}$, perfect selection $M_0 = \widehat{M}$, and no under-fitting $M_0 \subset \widehat{M}$ in Example 1.

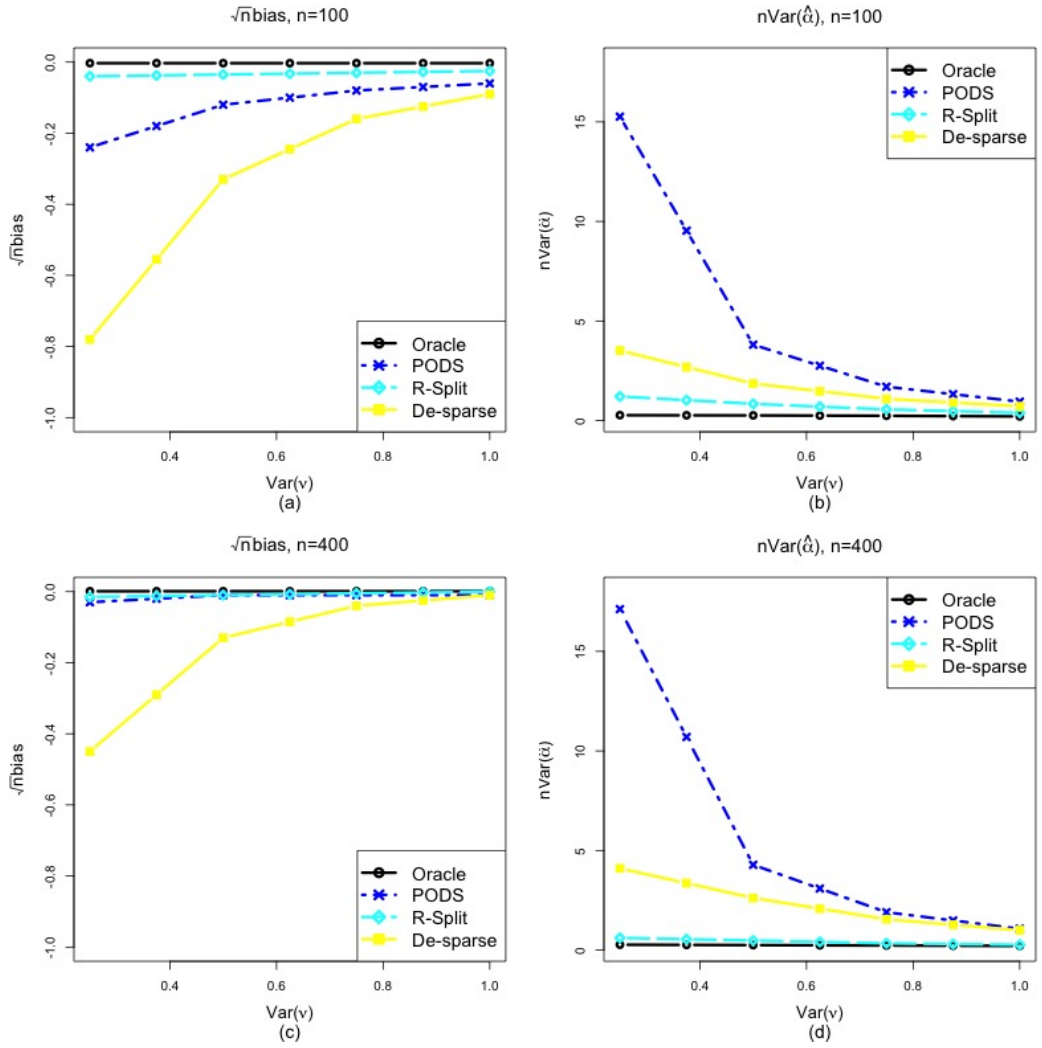

Figure 2: Finite sample comparison between R-Split and the two-stage selection methods based on Model (5.4). The data generating process is described in Section 5 and $\Sigma_{jk} = 0.9^{|j-k|}$ is the (j,k)-th element of Σ, for $j,k = 1, \cdots, p+1$. Panels (a) and (c) show the $\sqrt{n}$ times the bias of the α estimates. Panels (b) and (d) show n times the variance of the α estimates.

Table 1: Performance summaries for various methods under Setting 1 with $(n, p) = (100, 500)$.

	Oracle	Double	Double-2CV	PODS	R-Split	PODS-Split	De-sparsified	Alasso+OLS
	β is sparse, independent predictors, $\widehat{s}_{\min} = 6$							
$\sqrt{n}$Bias	0.05(0.05)	−0.17(0.05)	0.03(0.09)	0.04(0.05)	0.03(0.05)	0.03(0.05)	−0.28(0.06)	−0.37(0.05)
nMSE	1.07(0.07)	1.17(0.08)	4.16(1.66)	1.14(0.08)	1.20(0.08)	1.23(0.08)	1.72(0.12)	1.57(0.10)
Cover	0.95(0.01)	0.92(0.01)	0.91(0.01)	0.93(0.01)	0.96(0.01)	0.96(0.01)	0.93(0.01)	0.84(0.02)
Length	0.20(0.00)	0.20(0.00)	0.25(0.00)	0.20(0.00)	0.22(0.00)	0.22(0.00)	0.24(0.00)	0.17(0.00)
	β is sparse, $\Sigma_{ij} = 0.3^{\mathbf{1}(i\neq j)}$, $\widehat{s}_{\min} = 6$.							
$\sqrt{n}$Bias	0.02(0.05)	−0.62(0.08)	0.43(0.09)	0.03(0.09)	0.12(0.06)	0.19(0.07)	−0.15(0.07)	−2.72(0.07)
nMSE	1.36(0.09)	3.59(0.25)	3.87(0.25)	3.83(0.29)	2.12(0.14)	2.19(0.14)	2.40(0.16)	9.66(0.56)
Cover	0.93(0.01)	0.90(0.01)	0.91(0.01)	0.91(0.01)	0.93(0.01)	0.94(0.01)	0.90(0.01)	0.28(0.02)
Length	0.22(0.00)	0.32(0.00)	0.33(0.00)	0.32(0.00)	0.27(0.00)	0.28(0.00)	0.27(0.00)	0.20(0.00)
	β is sparse, $\Sigma_{ij} = 0.9^{\mathbf{1}(i\neq j)}$, $\widehat{s}_{\min} = 10$.							
$\sqrt{n}$Bias	0.04(0.14)	3.80(0.18)	0.42(0.20)	−0.06(0.20)	0.39(0.15)	0.50(0.15)	−0.75(0.15)	−3.62(0.16)
nMSE	9.14(0.57)	30.04(2.05)	19.74(1.27)	20.76(1.46)	11.69(0.76)	11.90(0.75)	11.87(0.74)	26.11(1.62)
Cover	0.94(0.01)	0.83(0.02)	0.93(0.01)	0.90(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)	0.56(0.02)
Length	0.57(0.00)	0.77(0.01)	0.81(0.01)	0.78(0.01)	0.66(0.01)	0.68(0.00)	0.61(0.00)	0.44(0.00)
	β is sparse, $\Sigma_{ij} = 0.9^{\lvert i-j\rvert}$, $\widehat{s}_{\min} = 6$.							
$\sqrt{n}$Bias	0.03(0.11)	−0.25(0.11)	−0.04(0.12)	−0.04(0.11)	0.42(0.10)	−0.02(0.11)	0.45(0.10)	−0.98(0.12)
nMSE	5.95(0.37)	6.09(0.40)	6.63(0.42)	6.07(0.41)	5.58(0.33)	6.35(0.44)	5.55(0.34)	7.60(0.54)
Cover	0.93(0.01)	0.92(0.01)	0.94(0.01)	0.93(0.01)	0.90(0.02)	0.95(0.01)	0.88(0.01)	0.78(0.02)
Length	0.46(0.00)	0.45(0.00)	0.46(0.00)	0.45(0.00)	0.35(0.00)	0.49(0.00)	0.37(0.00)	0.33(0.00)
	β is moderately sparse, Independent predictors, $\widehat{s}_{\min} = 10$.							
$\sqrt{n}$Bias	0.05(0.05)	−0.62(0.08)	0.02(0.12)	0.14(0.08)	0.19(0.07)	0.17(0.08)	−0.78(0.07)	−0.64(0.06)
nMSE	1.18(0.08)	3.71(0.24)	7.10(0.88)	2.98(0.19)	2.71(0.17)	3.48(0.23)	3.37(0.21)	2.24(0.16)
Cover	0.96(0.01)	0.89(0.01)	0.92(0.01)	0.87(0.02)	0.95(0.01)	0.95(0.01)	0.88(0.01)	0.82(0.02)
Length	0.22(0.00)	0.33(0.00)	0.45(0.00)	0.28(0.00)	0.33(0.00)	0.38(0.00)	0.29(0.00)	0.20(0.00)
	β is moderately sparse, $\Sigma_{ij} = 0.9^{\mathbf{1}(i\neq j)}$, $\widehat{s}_{\min} = 10$.							
$\sqrt{n}$Bias	−0.14(0.12)	−0.28(0.11)	−0.11(0.12)	−0.20(0.11)	1.10(0.10)	−0.00(0.11)	0.17(0.10)	0.13(0.10)
nMSE	6.99(0.45)	5.87(0.37)	6.76(0.41)	5.99(0.38)	6.20(0.38)	6.12(0.42)	5.06(0.29)	5.33(0.31)
Cover	0.94(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)	0.80(0.02)	0.95(0.01)	0.90(0.01)	0.80(0.02)
Length	0.50(0.00)	0.45(0.00)	0.49(0.00)	0.45(0.00)	0.35(0.00)	0.51(0.00)	0.37(0.00)	0.30(0.00)
	β is moderately sparse, $\Sigma_{ij} = 0.9^{\lvert i-j\rvert}$, $\widehat{s}_{\min} = 10$.							
$\sqrt{n}$Bias	0.09(0.12)	−0.12(0.11)	0.04(0.12)	0.01(0.11)	1.11(0.10)	0.02(0.12)	0.32(0.10)	0.20(0.10)
nMSE	7.23(0.46)	6.02(0.38)	7.12(0.44)	6.44(0.41)	6.43(0.39)	6.44(0.42)	5.46(0.32)	5.39(0.33)
Cover	0.94(0.01)	0.95(0.01)	0.95(0.01)	0.93(0.01)	0.80(0.02)	0.95(0.01)	0.89(0.01)	0.79(0.02)
Length	0.51(0.00)	0.45(0.00)	0.50(0.00)	0.45(0.00)	0.35(0.00)	0.51(0.00)	0.37(0.00)	0.29(0.00)
		β is dense, $\Sigma_{ij} = 0.9^{\lvert i-j\rvert}$, $\widehat{s}_{\min} = 10$.						
$\sqrt{n}$Bias	-	−1.32(0.14)	−0.13(0.21)	−0.16(0.18)	1.95(0.13)	−0.14(0.17)	−0.63(0.12)	0.37(0.12)
nMSE	-	12.24(0.88)	22.89(1.50)	13.12(1.05)	14.43(0.80)	11.95(0.91)	7.30(0.39)	6.89(0.49)
Cover	-	0.89(0.01)	0.93(0.01)	0.84(0.02)	0.74(0.02)	0.94(0.01)	0.77(0.02)	0.73(0.02)
Length	-	0.58(0.00)	0.85(0.01)	0.54(0.00)	0.43(0.00)	0.71(0.00)	0.35(0.00)	0.30(0.00)

When β is not sparse, we omit the results for "Oracle". $\widehat{s}_{\min}$ is the minimum model size used in R-Split. The numbers in the parenthesis are the standard errors of the estimated values. The nominal coverage probability is 0.95.

Table 2: Notations are the same as in Table 1. The results are based on Setting 2 with $(n, p) = (100, 500)$.

	Oracle	Double	Double-2CV	PODS	R-Split	PODS-Split	De-sparsified	Alasso+OLS
	β and γ are sparse, $R_y^2 = 0.8$, $R_d^2 = 0.5$, $\widehat{s}_{\min} = 6$							
$\sqrt{n}$Bias	0.05(0.03)	−0.02(0.05)	0.23(0.05)	−0.04(0.05)	0.14(0.05)	0.06(0.05)	0.38(0.05)	−0.94(0.06)
nMSE	0.46(0.03)	1.15(0.08)	1.49(0.10)	1.16(0.08)	1.13(0.08)	1.21(0.08)	1.27(0.09)	2.79(0.37)
Cover	0.95(0.01)	0.95(0.01)	0.91(0.01)	0.94(0.01)	0.95(0.01)	0.95(0.01)	0.94(0.01)	0.70(0.02)
Length	0.14(0.00)	0.21(0.00)	0.22(0.00)	0.22(0.00)	0.22(0.00)	0.23(0.00)	0.22(0.00)	0.17(0.00)
	β and γ are sparse, $R_y^2 = 0.8$, $R_d^2 = 0.9$, $\widehat{s}_{\min} = 6$							
$\sqrt{n}$Bias	0.03(0.01)	−0.03(0.05)	0.02(0.05)	−0.01(0.05)	0.09(0.03)	0.02(0.05)	0.18(0.03)	−1.00(0.05)
nMSE	0.10(0.01)	1.04(0.07)	1.25(0.08)	1.03(0.07)	0.34(0.03)	1.12(0.07)	0.41(0.03)	2.18(0.14)
Cover	0.95(0.01)	0.95(0.01)	0.93(0.01)	0.94(0.01)	0.94(0.01)	0.96(0.01)	0.94(0.01)	0.58(0.02)
Length	0.06(0.00)	0.20(0.00)	0.21(0.00)	0.20(0.00)	0.14(0.00)	0.23(0.00)	0.12(0.00)	0.13(0.00)
	β is moderately sparse and γ is dense, $R_y^2 = 0.8$, $R_d^2 = 0.3$, $\widehat{s}_{\min} = 10$							
$\sqrt{n}$Bias	0.02(0.03)	−0.27(0.05)	0.60(0.04)	0.08(0.04)	0.36(0.04)	0.23(0.04)	0.34(0.04)	−0.77(0.05)
nMSE	0.48(0.03)	1.19(0.09)	1.25(0.09)	0.92(0.07)	0.94(0.07)	0.91(0.06)	0.96(0.07)	2.03(0.13)
Cover	0.94(0.01)	0.92(0.01)	0.86(0.02)	0.93(0.01)	0.90(0.01)	0.93(0.01)	0.93(0.01)	0.67(0.02)
Length	0.14(0.00)	0.19(0.00)	0.17(0.00)	0.18(0.00)	0.17(0.00)	0.19(0.00)	0.19(0.00)	0.15(0.00)
	β is moderately sparse and γ is dense, $R_y^2 = 0.8$, $R_d^2 = 0.8$, $\widehat{s}_{\min} = 10$							
$\sqrt{n}$Bias	0.03(0.02)	−0.43(0.05)	0.26(0.04)	−0.03(0.05)	0.24(0.03)	0.10(0.03)	0.33(0.03)	−0.56(0.05)
nMSE	0.22(0.01)	1.55(0.10)	1.00(0.10)	1.02(0.07)	0.43(0.03)	0.59(0.03)	0.58(0.04)	1.33(0.10)
Cover	0.94(0.01)	0.90(0.01)	0.85(0.02)	0.95(0.01)	0.90(0.01)	0.94(0.01)	0.92(0.01)	0.68(0.02)
Length	0.09(0.00)	0.20(0.00)	0.14(0.00)	0.19(0.00)	0.11(0.00)	0.14(0.00)	0.13(0.00)	0.11(0.00)
		β is dense and γ is dense, $R_y^2 = 0.8$, $R_d^2 = 0.8$ $\widehat{s}_{\min} = 15$						
$\sqrt{n}$Bias	–	2.01(0.07)	4.28(0.06)	1.95(0.14)	4.06(0.04)	3.47(0.04)	4.01(0.04)	2.28(0.07)
nMSE	-	6.20(0.38)	20.04(1.01)	6.13(0.38)	17.18(0.83)	12.89(0.65)	16.91(0.82)	7.35(0.51)
Cover	-	0.51(0.02)	0.03(0.01)	0.39(0.02)	0.00(0.00)	0.16(0.02)	0.00(0.00)	0.20(0.02)
Length	-	0.21(0.00)	0.18(0.00)	0.16(0.00)	0.13(0.00)	0.20(0.00)	0.14(0.00)	0.11(0.00)

When β is not sparse, we omit the results for "Oracle". The reduced forms of R^2 are defined by $R_y^2 = 1 - \frac{\mathbb{E}(\varepsilon_i + \alpha\nu_i)^2}{\mathrm{Var}(Y_i)} = 1 - \frac{\sigma_\varepsilon^2 + \alpha^2\sigma_\nu^2}{\mathrm{Var}(Y_i)}$, and $R_d^2 = 1 - \frac{\sigma_\nu^2}{\mathrm{Var}(D_i)}$. The numbers in the parenthesis are the standard errors of the estimated values. The nominal coverage probability is 0.95.

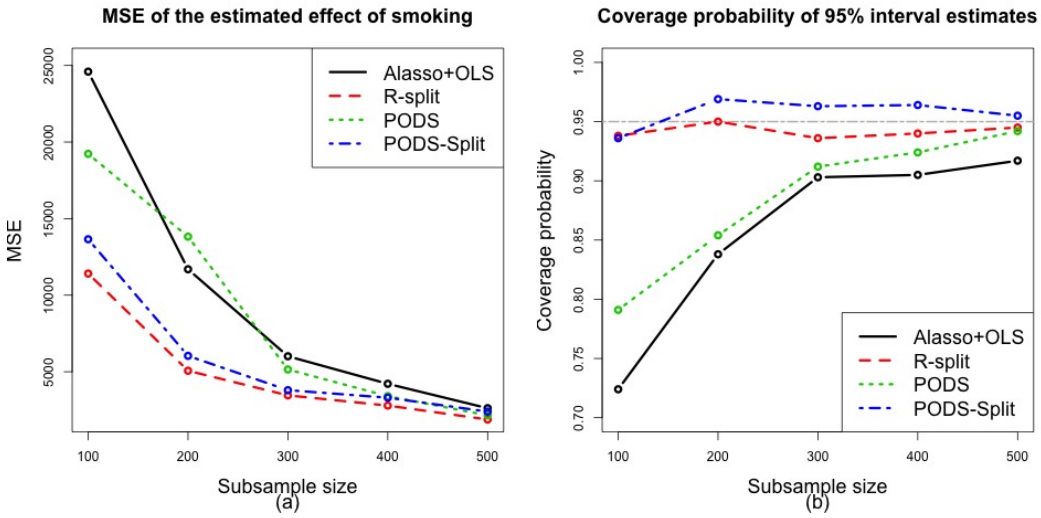

Figure 3: Finite sample performance for estimating effect of smoking on infants birth weights, aggregated over 1,000 replications.