Debiased Post Selection Inference

by

Jingshen Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2019

Doctoral Committee:

Professor Xuming He, Chair
Professor Matias D. Cattaneo
Assistant Professor Gongjun Xu
Professor Ji Zhu

©Jingshen Wang

jshwang@umich.edu

ORCID iD: 0000-0001-9498-1185

2019

To my parents

ACKNOWLEDGMENTS

I would like to first thank my advisor Xuming He for his support, patience, and guidance during the past years. Xuming's rigor and incisiveness on research have profoundly influenced my way of approaching statistical problems. I will greatly miss our conversations, meetings, and phone calls, which were not only fun but also influenced my perception on academia and personal life.

I would like to thank Gongjun Xu for the helpful discussions on my work. Gongjun's seemingly unlimited availability to meet and to discuss any questions were in fact indispensable to me. I am immensely grateful to Matias Cattaneo for many helpful advice and discussions during my job search process. Matias is not only an exemplary researcher, a motivating teacher and a caring mentor, but also an inspirational person who always has abundance of positive energy around. I would like to thank Rocio Titiunik for sharing her experience in personal and academic lives with me. The power in Rocio's words gives me the confidence to overcome unforeseen challenges in the future. I would like to express my deep gratitude to Ji Zhu for bringing me to Michigan and giving me this wonderful PhD journey in the department. I would also specially thank my dear friend Xinwei Ma for many interesting discussions on diverse research topics.

Finally, I would like to acknowledge my beloved parents and Alexander Giessing for their unwavering support in my life.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

This dissertation concerns the post-selection bias issue 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. We show that the over-fitting bias can be reduced or eliminated through data splitting, and more importantly, smoothing over random data splits or bootstrap-induced 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. Based on these insights, we thoroughly study the connections between our current framework and the estimates of the average treatment effects under the Neyman-Rubin causal model. A careful analysis shows that the post-selection bias issue can exist in a wider range of treatment effect estimation procedures. Under appropriate conditions we show that our 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.

CHAPTER 1

Introduction

In the modern era, we are often challenged by high dimensional data with many different characteristics per subject. For example, biomedical scientists may study the genomes of patients to choose a precise treatment and to learn the underlying cause of a disease; social scientists study individual behavior from multiple perspectives to decide the effectiveness of a training program. Thus, there is a crucial need to sort through this mass of information in high dimensional data, and provide valid statistical inference. In recent years, two lines of research appear to dominate the literature for high dimensional data analysis.

The first line of research provides statistical inference frameworks for scientists who start their research by running exploratory data analysis on high dimensional data, and form their research questions after model/variable selection. For example, one may assume that the observed data $\{Y_i, X_i\}_{i=1}^n$ are i.i.d and follow a high dimensional linear regression model

$$Y_i = X_i'\beta + \varepsilon_i, \quad i = 1, \cdots, n,$$

where Y_i is the outcome variable, X_i is the high dimensional covariate, β is a high dimensional sparse vector of coefficients, and ε is a noise variable. In this context, Lee et al. (2016) and Taylor and Tibshirani (2015) proposed a framework, called “selective inference”, which constructs exact confidence intervals for the selected regression coefficients conditional on the selected model. As long as the selective event can be rewritten as affine constraints on the response vector $Y = (Y_1, \cdots, Y_n)'$, selective inference forms valid confidence intervals for the selected coefficients. To be more specific, suppose that M is the set of all variables and $\widehat{M}$ is the selected set of variables, for $j \in \widehat{M}$, Lee et al. (2016) finds the confidence interval C_j^M for β_j^M with desired coverage probability $1-q$ that satisfies

$$\mathbb{P}\left(\beta_j^M \in C_j^M | \widehat{M} = M\right) = 1 - q.$$

Since the confidence interval is obtained after conditioning on the selection event, it follows that selective inference may produce a conservative inference procedure. Other than selective inference, Berk et al. (2013) and Kuchibhotla et al. (2018) carry out valid post-selection inference (PoSI) by considering all possible model selection procedures that could have produced the selected model. As the authors point out, the inferences are also generally conservative but have the advantage that they require neither perfect model selection nor affine constraint on the selection event.

The second line of research is developed for scientists who start with a pre-specified question (e.g., what is the treatment effect of a medical intervention, what is the effect of interest rates on housing price, etc.), and hope to construct confidence intervals to answer that question. In the presence of high dimensional covariates, standard point estimates in the classical theory of statistical inference are usually biased and methodological advances are required. In this thesis, we thoroughly study the bias issue after model selection when the parameter of interest is a fixed quantity, and our analysis is taken to be in the traditional sense of statistical inference.

As there is a growing literature on program evaluations, where estimation of the treatment effects is a valuable part of the statistical analysis in analyzing how treatments or social policies affect the outcome distributions of interest, we focus on the problem of statistical inference on treatment effects in the presence of high dimensional covariates. Suppose that we have n independent and identically distributed 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 is 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 thesis, 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 thesis is the sparsity in β. Formally, we require $M_0 = \text{supp}(\beta) = \{j \in \{1, \dots , p\} : \beta_j \neq 0\}$ 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. One is to perform inference after a sufficiently small model (that includes D) is selected, and the other is to perform debiased inference directly on a regularization method.

In the first chapter of this thesis, we focus on the method where inference is carried out on a selected model. Any reasonable model selection method can be used, for example Lasso (Tibshirani, 1996), the smoothly clipped absolute deviation penalized maximum likelihood (Fan and Li, 2001), the adaptive Lasso (Zou, 2006) and many others. When perfect model selection is attained, the resulting estimate of the treatment effect achieves the oracle property (Fan and Li, 2001), and post selection inference is asymptotically valid, e.g. 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 in Chapter 2.

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 Hong et al. (2018) and Chernozhukov et al. (2018) in a related context. An easy solution to avoid this over-fitting bias is the old idea of data splitting.

A main contribution of this thesis 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, which adopts random data splitting or bootstrap-induced data splitting, is similar in spirit to the bagging of Breiman (1996). For each split, model selection and OLS estimation are performed on two independent parts of the data, and the proposed estimator of α is the average of the estimates over many data splits. Data splitting has been used by other researchers for debiased inference. Wager and Athey (2017) used data splitting on random forests-based inference on the treatment effect and established the asymptotic normality for the estimator under the assumption that the subsample size with each split does not grow linearly with n, which is different from the splits that we consider for the regression approach. Additionally, Chernozhukov et al. (2018), Robins et al. (2017) and Wager et al. (2016) adopted the approach of data splitting and aggregation 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 this procedure as cross-estimation. As illustrated in our numerical studies, our proposed approach results in better efficiency by allowing repeated data splitting with overlapping sub-samples for estimation. We also note that under stronger parametric assumptions on the noise ε, one may follow Fithian et al. (2014) to apply the Rao-Blackwell theorem on the data splitting estimator to obtain an optimal estimator that utilizes the full data.

In the second chapter of this thesis, when the parameter of interest α is independent of the selected model, 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 variables D. Our analysis confirms a delicate bias-variance trade-off in the cases where the treatment variable is correlated with some of the covariates that are not active in the model conditional on the treatment.

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 those debiased inference methods. When the bias issue is not a main concern, we show that the two-stage selection procedure is not as efficient as the repeated data splitting approach in observational studies.

In the third chapter of this thesis, we consider a special case when $D \in \{0, 1\}$ is a binary random variable. Under the Neyman-Ruin causal model and the unconfoundedness assumption, see Neyman (1923) for detailed discussion, we provide an extension of the repeated data splitting approach to incorporate propensity score as part of the model, where larger approximation errors can be accommodated as long as the propensity score is well estimated. For the second part of Chapter 4, we provide a potentially interesting extension of the repeated data splitting approach for estimating heterogeneous treatment effect (HTE). This can be particularly useful to study in subgroup analysis, where the goal often includes reporting treatment effects within subgroups of subjects defined by a variable of interest. For instance, studies in biomedical science may evolve estimating treatment effects for a group of patients at a certain age; studies in marketing often try to estimate the treatment effect for the individuals for whom a job training program may be most beneficial.

The rest of this thesis is structured as follows. In Chapter 2, we use motivating examples to illustrate the bias issue for inference on α by refitting the OLS to a selected model, and we propose the repeated data splitting approach to eliminate the over-fitting bias. In Chapter 3, 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 the second part of

Chapter 3, we give theoretical and numerical comparisons for several methods of debiased inference. In the last part of Chapter 3, 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. In Chapter 4, we discuss an extension of our framework for estimating the average treatment effect and the heterogeneous treatment effect. Finally, we conclude our work in Chapter 5 with some future directions and discussions.

CHAPTER 2

Bias after Model Selection and Repeated Data Splitting

In this chapter, we first formalize the notations used in the thesis. Then we discuss the bias issue of the OLS estimator in a selected model, followed by a repeated data splitting approach to remove this bias.

2.1 Notations

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 $\boldsymbol{Z} = (Z_1^{\mathrm{T}}, \cdots, Z_n^{\mathrm{T}})$, $\boldsymbol{X} = (X_1^{\mathrm{T}}, \cdots, X_n^{\mathrm{T}})$, $D = (D_1, \cdots, D_n)^{\mathrm{T}}$, 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 a 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 $\Sigma = \mathbb{E}(Z_i^{\mathrm{T}} Z_i)$ 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

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

The performance of $\widehat{\alpha}_{\text{OLS}}$ is evaluated by Belloni et al. (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 $\widehat{\alpha}_{\text{OLS}}$ as

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

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 $\widehat{\alpha}_{\text{OLS}}$ as an estimator of α is in the same order of $1/\sqrt{n}$, which would result in biased inference.

If 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 two examples when β is sparse.

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 the Appendix. 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* $X_{i,j} = \widetilde{Z}_{i,j}$ *be the covariates, for* $i = 1, \ldots, n$, $j = 2, \ldots, p + 1$. *If a model selection procedure is the oracle, then*

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

where $\sigma^2_{\mathit{oracle}} = \sigma^2_{\varepsilon}(\Sigma^{-1}_{M_0})_{11}$, $\sigma^2_{\varepsilon} = \mathit{Var}(\varepsilon_i)$, *and* $(\Sigma^{-1}_{M_0})_{11}$ *denotes the first diagonal element of* $\Sigma^{-1}_{M_0}$. *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* $\widehat{\alpha}_{\mathit{OLS}}$ *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 2.1 are evaluated though 1000 Monte Carlo samples. From Figure 2.1(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-

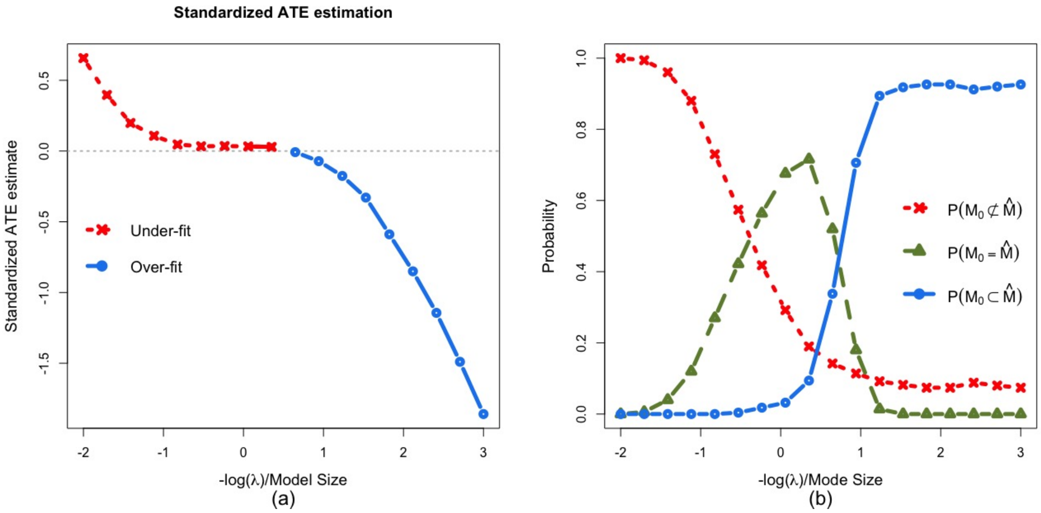


Figure 2.1: (a) The left panel shows standardized bias of Alasso+OLS estimator as the tuning parameter λ varies from $\exp(2)$ to $\exp(-3)$. The horizontal axis is $-\log(\lambda)$ as a measure of model size. (b) The right panel shows the probabilities of under-fitting $M_0 \not\subset \widehat{M}$, perfect selection $M_0 = \widehat{M}$, and no under-fitting $M_0 \subset \widehat{M}$ in Example 1.

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 2.1(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.

Example 2 (A simple model without covariates)**.** *To understand the over-fitting bias, we consider a simple model to illustrate the point, $Y = \alpha D + \varepsilon$, where ε is the white noise. For easy notation, suppose our covariates and the response are centered and thus the intercept is not considered. The treatment effect is the coefficient of D. When D is a binary random variable, in randomized experiments, this simple model suggests that the treatment assignment is not influenced by any potential confounding factors, both observed and un-*

observed. Due to model selection, as we discussed before, we have $\mathbb{E}(\varepsilon|\boldsymbol{X}_{\widehat{M}}) \neq 0$ *if* $\widehat{M}$ *contains any covariates from* $\boldsymbol{X}$*.*

In this example, the estimated coefficients from the working model with any endogenous variables is biased. To simplify the notation, consider the case where only one covariate (beyond the treatment variable) is included in the working model. In this case, the over-fitting bias can be further simplified into

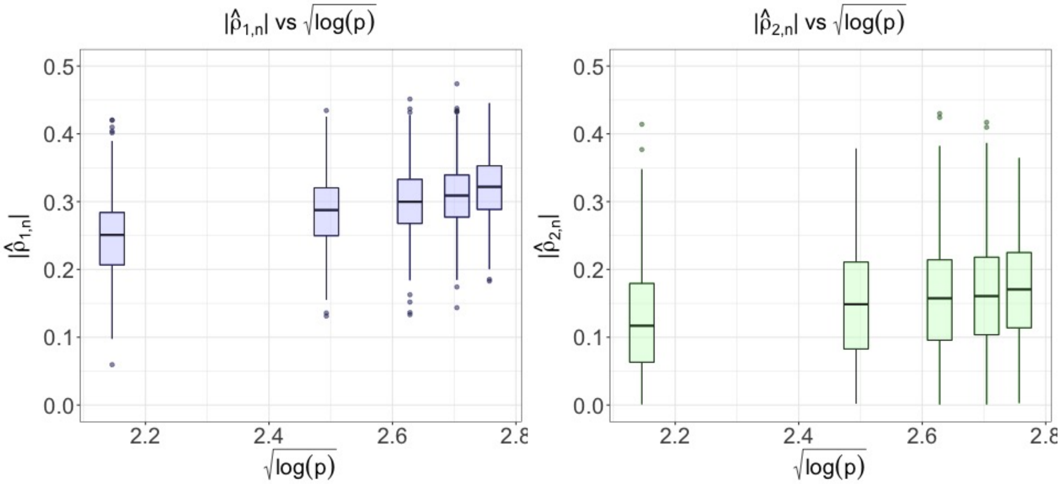
$$\mathbb{E}\widehat{\alpha}_{\text{OLS}} - \alpha = E\left\{\frac{\widehat{\rho}_{1,n}\widehat{\rho}_{2,n} - \widehat{\text{corr}}_n(\varepsilon, D)\|D\|_2^2/n}{\widehat{\rho}_{2,n}^2 - 1}\frac{\|\varepsilon\|_2}{\|D\|_2}\right\},$$

where $\widehat{\rho}_{1,n} = \widehat{\text{corr}}_n(\varepsilon, \boldsymbol{X}_{\widehat{M}})$ and $\widehat{\rho}_{2,n} = \widehat{\text{corr}}_n(D, \boldsymbol{X}_{\widehat{M}})$ are the correlations between the over-selected variable $\boldsymbol{X}_{\widehat{M}}$ and D and ε, respectively. These correlations are similar to spurious correlations, and may increase in magnitude with p, even when both $\boldsymbol{X}$ and D are generated completely independent of ε. Fan et al. (2018) addressed a related problem and derived the distribution of the maximum spurious correlation for high dimensional variables.

Next, we provide a simulation study to support the heuristic given above. Let $n = 100$, $p \in \{100, 500, 1000, 1500, 2000\}$, $\alpha = 1$, $\varepsilon_i \sim N(0, 1)$, and generate $\widetilde{Z}_i \sim N(0, \boldsymbol{I}_{p+1})$, then let $D_i = 1(Z_{i1} > 0)$ and $X_{ij} = \widetilde{Z}_{ij}$ be the covariates, for $i = 1, \cdots, n, j = 2, \cdots, p+1$. We proceed the model selection step with marginal screening. As the upper bound of over-fitting bias derived in (2.10) increases with $\sqrt{\log p}$, we plot $|\widehat{\rho}_{1,n}|$ and $|\widehat{\rho}_{2,n}|$ against $\sqrt{\log p}$. From the results shown in Figure 2.2, we observe that the sizes of $\widehat{\rho}_{1,n}$, and to a lesser extent $\widehat{\rho}_{2,n}$, increase with the dimension of the covariates.

Remark 1 (Over-fitting bias for predicting Y)**.** *The over-fitting bias issue we discussed in this section also applies when the goal is to predict the response* Y*. Consider the refitted OLS prediction* $\widehat{Y} = \widehat{\alpha}_{OLS}D + \boldsymbol{X}\widehat{\beta}_{OLS}$*, then even if* $M_0 \subseteq \widehat{M}$*, we have*

$$\mathbb{E}(\widehat{Y} - \alpha D - \boldsymbol{X}\beta) = \mathbb{E}\left(\boldsymbol{Z}_{\widehat{M}}(\boldsymbol{Z}_{\widehat{M}}'\boldsymbol{Z})^{-1}\boldsymbol{Z}_{\widehat{M}}'\varepsilon\right) \neq 0,$$



(a)

(b)

Figure 2.2: Based on 500 Monte Carlo samples. Panel (a)-(b) show the box-plots of $|\widehat{\rho}_{1,n}|$ and $|\widehat{\rho}_{2,n}|$ for different dimensions. The data generating process is given in the example in Chapter 2.2.

due to the correlation between $\boldsymbol{Z}_{\widehat{M}}$ and ε.

2.3 Repeated data splitting

Since the over-fitting bias is mainly caused by the spurious correlation between the over-selected variables and the noise, it can be easily 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. (2016) 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) and bootstrap-induced splitting (B-Split), in more detail.

2.3.1 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 n_2 as the 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 softwares for regularized regression, such as `glmnet`.

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
        $(\widehat{\alpha}_b, \widehat{\beta}_b^{\mathrm{T}}) = \arg\min \sum_{j \in T_2} (Y_j - \alpha D_j - X^{\mathrm{T}}_{j,\widehat{M}_b}\beta)^2$,
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 equals 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 Appendix 2.5.6

$$\widehat{\sigma}_n^2 = n \sum\nolimits_{j=1}^{n} \left(\frac{n-1}{n-n_2} \widehat{S}_j \right)^2 - \frac{n_2 n^2}{B^2(n-n_2)} \sum\nolimits_{b=1}^{B} (\widehat{\alpha}_b - \widetilde{\alpha})^2, \tag{2.3}$$

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 by $\widetilde{\alpha} \pm Z_{q/2} n^{-1/2} \widehat{\sigma}_n$, where $Z_{q/2}$ is the $1 - q/2$ quantile of the standard normal distribution.

2.3.2 B-Split

In Efron (2014), the author discussed a bootstrap smoothing method to account for the variability of model selection. In that setting, model selection and parameter estimation are performed on the same bootstrap samples, so the over-fitting bias would remain in high dimensional problems. We find that a simple modification to Efron's approach addresses the bias issue. The proposed method is to draw a bootstrap sample for model selection, and then estimate the treatment effect α using the observations that do not show up in the bootstrap sample. On average, a bootstrap sample takes $0.632n$ distinct observations from the original sample, even though the bootstrap sample size remains at n. In other words, we now use the bootstrap-induced splitting, by using the bootstrap sample (of size n) to perform model selection but choosing observations not used in the bootstrap sample, roughly 36.8% of the original sample, for parameter estimation.

Algorithm 2 B-Split

```
For $b \leftarrow 1$ to $B$ do
    Step 1 Draw a bootstrap sample $\boldsymbol{\mathcal{X}}_b^* := (\mathcal{X}_{b1}^*, \cdots, \mathcal{X}_{bn}^*)$ from $\mathcal{X}$. Let $w_{bi}^*$ be the number of
        times the $i$th observation $\mathcal{D}_i$ appears in the bootstrap sample, and let $v_{bi}^* = 1_{(w_{bi}^*=0)}$.
    Step 2. Select a model $\widehat{M}_b^*$ to predict $Y$ based on $\boldsymbol{\mathcal{X}}_b^*$.
    Step 3. Refit the selected model $\widehat{M}_b^*$ with the observations not in the bootstrap sample
        to get $(\widehat{\alpha}_b^*, \widehat{\beta}_b^{\mathrm{T}*})^{\mathrm{T}} = \arg\min \sum_{i=1}^n v_{bi}^*(Y_i - \alpha D_i - X_{i,\widehat{M}_b^*}^{\mathrm{T}}\beta)^2$.
The final smoothed estimate is $\widetilde{\alpha} = \frac{1}{B}\sum_{b=1}^B \widehat{\alpha}_b^*$.
```

We refer to this bootstrap-induced data splitting as B-Split. Clearly, there is similarity between B-Split and data carving as used in Fithian et al. (2014). Similar to R-Split, we can view the smoothed estimator obtained from Algorithm 2 as a conditional expectation, $\widetilde{\alpha} = \mathbb{E}(\widehat{\alpha}_b^* | \boldsymbol{\mathcal{X}})$. The weight $V^* = (V_1^*, \cdots, V_n^*)$ is from the set $\mathcal{V}^* = \{V^* \in R^n : \; V_i^* = 1_{(W_i^*=0)}, \; i = 1, \cdots, n, \; (W_1^*, \cdots, W_n^*) \sim \text{Mult}(n, 1/n)\}$, where $\text{Mult}(n, 1/n)$ denotes the multinomial distribution with n trails and each event has the success probability of $1/n$. Following Efron (2014) and Wager et al. (2014), we can construct a variance estimate for B-Split estimator as

$$\widehat{\sigma}_n^2 = n\sum_{j=1}^{n} \widehat{S}_j^{*2} - \frac{n^2}{B^2}\sum_{b=1}^{B} \left(\widehat{\alpha}_b^* - \widetilde{\alpha}\right)^2 , \tag{2.4}$$

where $\widehat{S}_j^* = \frac{1}{B}\sum_{b=1}^B (w_{bi}^* - \frac{1}{B}\sum_{k=1}^B w_{ki}^*)(\widehat{\alpha}_b^* - \widetilde{\alpha})$.

2.3.3 Theoretical investigation of R-Split

In this section, we study the theoretical properties of the smoothed estimators. Except for the space of weights V and V^*, B-Split and R-Split are intrinsically the same. To avoid redundancy, we focus on R-Split.

For a fixed model M and a weight $V \in \mathcal{V}_{n_2}$, define the covariance matrix in the given subsample as $\widehat{\Sigma}_{V,M} = n_2^{-1}\sum_{i=1}^{n} V_i Z_{i,M} Z_{i,M}^{\mathrm{T}}$, 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

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

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 definitions. For instance, in $\widehat{h}_{i,n}$, 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 (D_i, X_i) have zero mean and have bounded support with an upper bound C, i.e. $|D_i| \le C$, and $|X_{ij}| \le C$, for $i = 1, \cdots, n$, $j = 1, \cdots, p$. (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. *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*

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

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*

$$\sqrt{n}(\widetilde{\alpha} - \alpha) = \eta_n^{\mathrm{T}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i Z_i + o_p(1). \tag{2.5}$$

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

$$\widetilde{\sigma}_n^{-1} \sqrt{n}(\widetilde{\alpha} - \alpha) \rightsquigarrow N(0, 1). \tag{2.6}$$

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 to ensure that the under-fitting bias to be small. Since the selected model size allowed in Assumption 2 can be relatively large, we can choose a larger model than usual to control the under-fitting risk. The proof of the theorem is given in Appendix 2.5.2.

Since η_n plays a key role in the asymptotic expression of R-Split estimator, we consider a special case that $\widehat{M} = M$, where M is a fixed model. In this case, Assumption 3 is immediately satisfied since $\widehat{h}_{i,n} = 0$, for $i = 1, \cdots, n$. Then, η_n reduces to $e_1^{\mathrm{T}}\widehat{\Sigma}_M^{-1}\widetilde{I}_M$, and thus the linear representation in (2.5) simplifies into

$$\sqrt{n}(\widetilde{\alpha} - \alpha) = e_1^{\mathrm{T}}\widehat{\Sigma}_M^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i Z_{i,M} + o_p(1).$$

Therefore, the smoothed estimator $\widetilde{\alpha}$ shares the same asymptotic expression as the α estimate obtained from refitting model M with the full sample.

To consistently estimate the variance of the smoothed estimator, we adopt the nonparametric delta method proposed in Efron (2014). A cleaner version of the linear representation can be provided to build the foundation of the nonparametric delta method, however, stronger conditions are then required. We make the following assumptions.

Assumption 6. *In addition to Assumption 2, we have* $s \log p = o(n)$.

Assumption 7. *The quantities* $h_{i,n}$ *satisfy* $\sum_{i=1}^{n} h_{i,n}/\sqrt{n} = o_p(1)$.

Assumption 8. *There exists a constant vector* $\xi_n \in \mathbb{R}^{p+1}$ *that satisfies* $||\xi_n||_\infty \leq C_1$ *for a constant* C_1 *and*

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

Assumption 9. *The minimum s-sparse eigenvalues of the population covariance matrix is positive and bounded away from zero with $\lambda_{\min,s}(\Sigma) \geq \kappa_0 > 0$. There exists a positive constant $K < \infty$ such that, $\forall V \in \mathcal{V}_{n_2}$, $\mathbb{P}\left(\limsup_{n\to\infty} ||\widehat{\Sigma}^{-1}_{V,\widehat{M}} e_1||_2 \leq K\right) = 1$.*

Assumption 9 requires that uniformly over all possible models, the l_2-norm of the first column of inverse of the sample covariance matrix in a subsample of size n_2 is bounded above from infinity. Under Assumption 1 and Assumptions 6-9, we have

$$\widetilde{\alpha} = \alpha + \frac{1}{n}\sum_{i=1}^{n} U(\mathcal{X}_i) + o_p(1/\sqrt{n}), \tag{2.7}$$

where $U(\mathcal{X}_i) = \varepsilon_i \xi_n^{\mathrm{T}} Z_i$. Since $\mathbb{E}U(\mathcal{X}_i) = 0$ and $\mathbb{E}U(\mathcal{X}_i)^2 < \infty$, $\widetilde{\alpha}$ is asymptotically linear with the influence function $U(\mathcal{X}_i)$. The proof is provided in Appendix 2.5.3. As $n \to \infty$, $\widetilde{\alpha}$ is asymptotically normally distributed,

$$\sigma^{-1}\sqrt{n}\,(\widetilde{\alpha} - \alpha) \rightsquigarrow N(0,1), \tag{2.8}$$

where $\sigma^2 = \mathbb{E}U(\mathcal{X}_i)^2$, and σ^2 can be consistently estimated by the nonparametric delta method so that $\widehat{\sigma}_n - \sigma = o_p(1)$, where $\widehat{\sigma}_n$ is provided in (2.3) for R-Split. Similarly, we have (2.8) holds for B-Split with $\widehat{\sigma}_n$ provided in (2.4).

Remark 2 (Relationship between the oracle and the smoothed estimators)**.** *In R-Split, if perfect model selection is achieved in all splits, the influence function $U(\mathcal{X}_i)$ and the asymptotic variance reduces to $U(\mathcal{X}_i) = \varepsilon_i e_1^{\mathrm{T}} \Sigma_{M_0}^{-1} Z_{i,M_0}$ and $\sigma^2 = \mathbb{E}\{U(\mathcal{X}_i)^2\} = \sigma_\varepsilon^2 (\Sigma_{M_0}^{-1})_{11}$. In this case σ^2 equals the asymptotic variance of the oracle estimator, which implies that under model selection consistency, the smoothed estimator $\widetilde{\alpha}$ achieves oracle efficiency. However, if $\widehat{M_b}$ has a positive probability to be a larger model than M_0, $\widetilde{\alpha}$ is not expected to be oracle.*

Remark 3 (Comparison between R-Split and cross-estimation)**.** *As we mentioned in Section 1, cross-estimation can be used to removed the over-fitting bias. Take 2-fold cross-*

estimation for simplicity, suppose $V \in \mathcal{V}_{n_2}$ with $n_2 = n/2$, then α can be estimated by

$$\begin{aligned}\widetilde{\alpha}_{cv} =& \frac{1}{2}\Bigg\{ e_1^{\mathrm{T}} \left(\frac{1}{n}\sum_{i=1}^{n} V_i Z_{i,\widehat{M}_1} Z_{i,\widehat{M}_1} \right)^{-1} \frac{1}{n}\sum_{i=1}^{n} V_i Z_{i,\widehat{M}_1} Y_i \\ &+ e_1^{\mathrm{T}} \left(\frac{1}{n}\sum_{i=1}^{n} (1-V_i) Z_{i,\widehat{M}_2} Z_{i,\widehat{M}_2} \right)^{-1} \frac{1}{n}\sum_{i=1}^{n} (1-V_i) Z_{i,\widehat{M}_2} Y_i \Bigg\},\end{aligned}$$

where $\widehat{M_1}$ is the selected model from the subsample indexed by $\{i : V_i = 0,\ i = 1, \cdots, n\}$, and $\widehat{M_2}$ is selected from subsample indexed by $\{i : V_i = 1,\ i = 1, \cdots, n\}$. We show in Appendix 2.5.4 that the variance of $\widetilde{\alpha}_{cv}$ satisfies

$$\mathit{Var}(\sqrt{n}(\widetilde{\alpha}_{cv} - \alpha)) = \mathbb{E}\left\{\mathit{Var}(\sqrt{n}(\widetilde{\alpha}_{cv} - \alpha)|\boldsymbol{\mathcal{X}})\right\} + \mathit{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha)) \geq \mathit{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha)). \tag{2.9}$$

Thus, R-Split is more efficient than cross-estimation.

2.4 Finite-sample comparison between R-Split and B-Split

We have the flexibility to choose $r_v = n_2/n$ in R-Split, where n_2 is the number of the observations used in the estimation. Let ω be the variance of the estimated effect $\widehat{\alpha}_b$ from a single split, and ρ denotes the correlation between the estimates of two different random splits. Then the variance of R-Split estimator is of the same order of $\rho\omega$ as $B \to \infty$,

$$\begin{aligned}\mathrm{Var}(\widetilde{\alpha}) &= \frac{1}{B^2}\sum\nolimits_{b=1}^{B} \mathrm{Var}(\widehat{\alpha}_b) + \frac{1}{B^2}\sum\nolimits_{b_1 \neq b_2} \mathrm{Cov}(\widehat{\alpha}_{b_1}, \widehat{\alpha}_{b_2}), \\ &= \frac{1}{B}\mathrm{Var}(\widehat{\alpha}_1) + (1 - \frac{1}{B})\mathrm{Var}(\widehat{\alpha}_1)\mathrm{corr}(\widehat{\alpha}_1, \widehat{\alpha}_2), \\ &\to \mathrm{Var}(\widehat{\alpha}_1)\rho := \omega\rho, \quad \text{as } B \to \infty.\end{aligned}$$

From this point of view we see the choice of the ratio r_v can play a role in $\mathrm{Var}(\widetilde{\alpha})$: by making r_v smaller, we reduce the overlap between different subsamples for estimation,

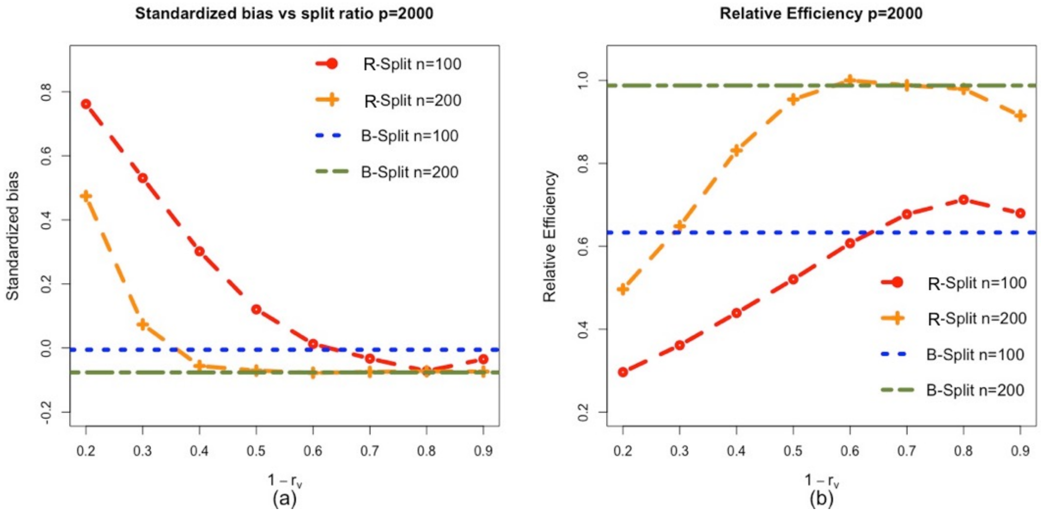


Figure 2.3: Summary for the equal correlation design with $\Sigma_{jk} = 0.3$ and $|\widehat{M}| = 10$ as the fraction of the data for model building $1 - r_v$ changes from 0.2 to 0.9. The horizontal lines capture the performance of B-Split estimator, which do not change with r_v. Panel (a) shows the standardized bias of the smoothed estimator. Panel (b) shows the relative efficiency of the smoothed estimators against the oracle estimator.

which leads to decreased correlation ρ. However, the smaller sample size reduces the accuracy of estimation in each split, which results in larger ω. The optimal choice of r_v at a given sample size is difficult to pin down and it depends on the underlying model. To illustrate this point, we choose $B = 2000$ in this subsection so that $\omega\rho$ provides a more accurate approximation of Var($\widetilde{\alpha}$).

Consider the same data generating process in Example 1 except for we set $\Sigma_{jk} = 0.3^{\mathbf{1}(j \neq k)}$, $(n, p) = (100, 2000)$ or $(200, 2000)$, and $\beta = (1, 1, 1, 1, 0, \cdots, 0)^{\mathrm{T}}$. We use Lasso for model selection, is implemented with R package `glmnet`. Furthermore, in each split, we select a model from the Lasso path whose model size is the closest to $s = 10$, a model size that keeps the under-fitting risk at a negligible level with strong signals. We report two quantities through simulation: (a) the ratio of the bias of R/B-Split estimator relative to the standard deviation of the oracle estimator, (b) the asymptotic relative efficiency of R/B-Split estimator to the oracle estimator.

The results are shown in Figure 2.3 as r_v varies from 0.1 to 0.8 for R-Split. We summarize the results in two points. From Figure 2.3(a), B-Split tends to have small bias and so does R-Split with r_v below 0.4. From Figure 2.3(b), the smoothed estimators from B-Split and R-Split with $r_v \approx 0.4$ are nearly as efficient as the oracle at $n = 200$ but the relative efficiency is never above 0.7 at $n = 100$. In this model, B-Split does well, and R-Split with at least 60% of the data in the model selection stage does almost as well in terms of both bias and variance. For smaller n (say $n = 100$), the under-fitting bias would be an issue if n_1, the subsample size for model selection is small. Then, R-Split benefits from the flexibility of choosing r_v to be small, that is, R-Split can outperform B-Split in such cases. For larger n, B-split is usually a solid choice in this case and many other cases that we have considered. The impact of the choice of r_v in R-Split is expected to diminish as n increases. Since R-Split and B-Split have similar performances whenever $r_v \in [0.6, 0.7]$, in the following subsections, we focus on the performance of R-Split with $r_v = 0.7$.

2.5 Proofs

2.5.1 Useful lemmas

In this section, we prove two useful lemmas that shall be used in the later proofs.

Lemma 1. *Under Assumption 1,* $\|\boldsymbol{Z}^T\varepsilon/\sqrt{n}\|_\infty = O_p(\sqrt{\log p})$.

Proof: Let $\delta_j = (\sum_{i=1}^n Z_{ij}^2)^{1/2}$ and $U_{ij} = \varepsilon_i Z_{ij}/\delta_j$. For $K > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\max_j \Big|\sum_{i=1}^n Z_{ij}\varepsilon_i/\sqrt{n}\Big| > \sqrt{\log p}K\right)\\
&\leq \mathbb{E}\left\{\mathbb{P}\left(\max_j \delta_j \cdot \max_j \Big|\sum_{i=1}^n U_{ij}/\sqrt{n}\Big| > \sqrt{\log p}K \;\Big|\; \boldsymbol{Z}\right)\right\}\\
&\leq p\mathbb{E}\left\{\mathbb{P}\left(\Big|\sum_{i=1}^n U_{ij}/\sqrt{n}\Big| > \sqrt{\log p}K/\max_j \delta_j \;\Big|\; \boldsymbol{Z}\right)\right\}\\
&\leq 2\exp\left(\log p - \frac{\log pK^2}{2\sigma_\varepsilon^2 C^2}\right),
\end{aligned}
$$

and the right hand side converges to zero when K is sufficiently large.

As an application of this Lemma, we can provide an upper bound of the over-fitting bias. Following the derivation of the bias decomposition in (2.2), and under the assumption that there exists a positive constant λ_0 such that

$$\mathbb{P}\left(\lim_{n\to\infty}\lambda_{s,\min}^{-1}\big(\frac{1}{n}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}\big)\geq\lambda_0\right)=1,$$

we have

$$\begin{aligned}
b_{n1}=&e_1^{\mathrm{T}}\left(\frac{1}{n}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{\widehat{M}}\right)^{-1}\frac{1}{\sqrt{n}}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\varepsilon\\
\leq&\|e_1\|_2\lambda_{\min}^{-1}\left(\frac{1}{n}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{\widehat{M}}\right)\|\frac{1}{\sqrt{n}}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\varepsilon\|_2\\
\leq&\lambda_{\min}^{-1}\left(\frac{1}{n}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{\widehat{M}}\right)|\widehat{M}|^{1/2}\cdot\|\boldsymbol{Z}^{\mathrm{T}}\varepsilon/\sqrt{n}\|_\infty\\
=&O_p\left(|\widehat{M}|^{1/2}\sqrt{\log p}\right). \qquad (2.10)
\end{aligned}$$

Lemma 2. *Under Assumptions 1 and 13 , we have*

$$\max_{|M|\leq s}\left|D^{\mathrm{T}}(\mathbf{I}-\boldsymbol{P}_M)D/n-(\Sigma_{11}-\Sigma_{D,M}^{\mathrm{T}}\Sigma_M^{-1}\Sigma_{D,M})\right|=o_p(1),$$

where $\boldsymbol{P}_M=\boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^{\mathrm{T}}$.

Proof: Denote $\|A\|=\{\mathrm{tr}(AA^{\mathrm{T}})\}^{1/2}$ for an arbitrary matrix A. To obtain the result, we prove the following two uniform convergence results hold:

$$\max_{|M|\leq s}\|n^{-1}D^{\mathrm{T}}\boldsymbol{X}_M-\Sigma_{D,M}\|=o_p(1), \qquad (2.11)$$

$$\max_{|M|\leq s}\|n^{-1}\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M-\Sigma_M\|=o_p(1). \qquad (2.12)$$

In (2.11), let $\sum_{i=1}^{n} D_i X_{ij}/n = \widehat{\sigma}_{D,j}$ and $\Sigma_{D,M} = (\sigma_{D,j}, j \in M) \in \mathbb{R}^{|M|}$, we have

$$\|n^{-1}D^{\mathrm{T}}\boldsymbol{X}_M - \Sigma_{D,M}\| =\left\{\sum_{j\in M}(\widehat{\sigma}_{D,j} - \sigma_{D,j})^2\right\}^{1/2} \leq s^{1/2}\max_{j\in M}|\widehat{\sigma}_{D,j} - \sigma_{D,j}|,$$

when $|M| \leq s$. Therefore, $\forall\epsilon > 0$ by adopting similar arguments used in Lemma 1

$$\begin{aligned}\mathbb{P}\left(\max_{|M|\leq s}\|n^{-1}D^{\mathrm{T}}\boldsymbol{X}_M - \Sigma_{D,M}\| > \epsilon\right) \leq & \sum_{|M|\leq s}\sum_{j\in M}\mathbb{P}\left(|\widehat{\sigma}_{D,j} - \sigma_{D,j}| > s^{-1/2}\epsilon\right)\\ \leq & sp^s 2\exp\left(-\frac{n\epsilon^2}{2C^2 s}\right)\\ = & \exp\left(s\log p + \log s - \frac{n\epsilon^2}{2C^2 s}\right). \qquad (2.13)\end{aligned}$$

By Assumption 13, the right hand of (2.13) converges toward 0 as $n \to \infty$. Applying similar techniques to those used in (2.11), we can also demonstrate (2.12). As a minor generalization of this lemma, we have

$$\mathbb{P}\left(\max_{j}\max_{|M|\leq s}\|Z_j^{\mathrm{T}}\boldsymbol{Z}_M/n - \Sigma_{j,M}\|_2 > \epsilon\right) \leq \exp\left(\log p + s\log p + \log s - \frac{n\epsilon^2}{2C^2 s}\right) = o(1).$$

Similarly, since the covariates are bounded by same constant C and $n_2/n = r_v$ is bounded away from 0 and 1, we have for a random subsample V,

$$\begin{aligned}&\mathbb{P}\left(\max_{j}\max_{|M|\leq s}\|Z_{j,V}^{\mathrm{T}}\boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}\|_2 > \epsilon\right)\\ &\leq \mathbb{E}\left\{\mathbb{P}\left(\max_{j}\max_{|M|\leq s}\|Z_{j,V}^{\mathrm{T}}\boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}\|_2 > \epsilon \mid V\right)\right\}\\ &\leq \exp\left(\log p + s\log p + \log s - \frac{r_v n\epsilon^2}{2C^2 s}\right) = o(1).\end{aligned}$$

2.5.2 Proof of Theorem 1 in Section 2.3.3

In this part, we provide the proof of Theorem 1.

Proof:

Step 1. The estimated treatment effect based on model M through ordinary least squares by using full sample can be written as

$$\begin{aligned}\widehat{\alpha}_M &= e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}Y \\ &= \alpha + e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}\varepsilon + e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}(\boldsymbol{X}\beta + R_n) \\ &= \alpha + e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\widetilde{\boldsymbol{I}}_M\boldsymbol{Z}^{\mathrm{T}}\varepsilon + (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)(\boldsymbol{X}\beta + R_n),\end{aligned} \tag{2.14}$$

where $e_1 = (1, 0, \cdots, 0)^{\mathrm{T}} \in \mathbb{R}^{p+1}$, and $\boldsymbol{P}_M = \boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^{\mathrm{T}}$. Since the decomposition given above is very important to understand the bias issue after model selection, we provide a detailed derivation for the last equality. By block matrix inversion, the first row of matrix $(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}$ equals

$$\left((D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}, -(D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}\boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\right),$$

and then multiply this quantity by $\boldsymbol{Z}_M^{\mathrm{T}} = (D^{\mathrm{T}}, \boldsymbol{X}_M^{\mathrm{T}})$, we get

$$\begin{aligned}e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}} &= \left((D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}, -(D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}\boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\right)(D^{\mathrm{T}}, \boldsymbol{X}_M^{\mathrm{T}}) \\ &= (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}} - (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}\boldsymbol{X}_M(\boldsymbol{X}_M^{\mathrm{T}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^{\mathrm{T}} \\ &= (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}} - (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}\boldsymbol{P}_M \\ &= (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M).\end{aligned} \tag{2.15}$$

Therefore, we obtain

$$e_1^{\mathrm{T}}(\boldsymbol{Z}_M^{\mathrm{T}}\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}_M^{\mathrm{T}}(\boldsymbol{X}\beta + R_n) = (D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)D)^{-1}D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_M)(\boldsymbol{X}\beta + R_n).$$

For a random model $\widehat{M}$, we may replace M with $\widehat{M}$ and get the decomposition provided in

(2.2) in Section 2.2:

$$\begin{aligned}\sqrt{n}(\widehat{\alpha}_{\text{OLS}} - \alpha) = e_1^{\mathrm{T}} \left(\frac{1}{n}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{\widehat{M}}\right)^{-1} \frac{1}{\sqrt{n}}\boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}}\varepsilon \\ + \left(\frac{1}{n}D^{\mathrm{T}}(I-\boldsymbol{P}_M)D\right)^{-1}\frac{1}{\sqrt{n}}D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}})(\boldsymbol{X}\beta+R_n).\end{aligned}$$

Step 2. Now suppose we take a subsample of size n_2 indexed by weight $V = (V_1, \cdots, V_n)$, let $\boldsymbol{Z}_V = (D_V, X_V)$ be the design matrix with rows $\{Z_i : \ V_i = 1, i = 1, \cdots, n\}$ and $g_V(\mathbf{W}) = \{g(W_i) : \ V_i = 1, i = 1, \cdots, n\}$. Denote the covariance matrix and the projection matrix in this subsample as $\widehat{\Sigma}_{V,M} = \boldsymbol{Z}_{V,M}^{\mathrm{T}}\boldsymbol{Z}_{V,M}/n$, and $\boldsymbol{P}_{V,M} = \boldsymbol{X}_{V,M}(\boldsymbol{X}_{V,M}^{\mathrm{T}}\boldsymbol{X}_{V,M})^{-1}\boldsymbol{X}_{V,M}^{\mathrm{T}}$ respectively. Let $\mathcal{X} = \{(Y_i, Z_i)\}_{i=1}^n$. Consider the smoothed estimator $\widetilde{\alpha}$:

$$\begin{aligned}\sqrt{n}(\widetilde{\alpha} - \alpha) =& \mathbb{E}\left(\sqrt{n}(\widehat{\alpha}_{\widehat{M}} - \alpha_0)|\boldsymbol{\mathcal{X}}\right) \\ =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}V_i\middle|\boldsymbol{\mathcal{X}}\right)Z_i\varepsilon_i \\ &+ \mathbb{E}\left(\sqrt{n}(D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V)^{-1}D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})g_V(\mathbf{W})|\boldsymbol{\mathcal{X}}\right) \\ =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_n^{\mathrm{T}}Z_i\varepsilon_i + r_{n1} + r_{n2},\end{aligned}$$

where

$$\begin{aligned}r_{n1} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}V_i\middle|\boldsymbol{\mathcal{X}}\right) - \eta_n\right\}Z_i\varepsilon_i, \\ r_{n2} &= \mathbb{E}\left(\sqrt{n}(D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V)^{-1}D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})X_V\beta|\boldsymbol{\mathcal{X}}\right).\end{aligned}$$

We next show that $r_{n1} = o_p(1)$ and $r_{n2} = o_p(1)$ in Steps 3 and 4 respectively.

Step 3. (Behavior of r_{n1}.) In r_{n1}, by conditioning on $V_i = 1$

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

The term $\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}\middle|\boldsymbol{\mathcal{X}}, V_i = 1\right)$ is the average of $e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}$ over all possible models but excluding the ith point. Let $\widetilde{V} = (\widetilde{V}_1, \cdots, \widetilde{V}_n)$ be another set of splitting weight that is independent with V, and denote $\widetilde{M}$ as the selected model indexed by $\widetilde{V}$. Following the definition in Assumption 3, the remainder term r_{n1} can be decomposed into two parts

$$\begin{aligned}r_{n1} = &\underbrace{\left\{r_v\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{\widetilde{V},\widetilde{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widetilde{M}}\middle|\boldsymbol{\mathcal{X}}\right) - \eta_n\right\}^{\mathrm{T}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\varepsilon_i}_{r_{n1}^a} \\ &+ \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}V_i\middle|\boldsymbol{\mathcal{X}}\right) - \mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{\widetilde{M},\widetilde{V}}^{-1}\widetilde{\boldsymbol{I}}_{\widetilde{M}}V_i\middle|\boldsymbol{\mathcal{X}}\right)\right\}^{\mathrm{T}} Z_i\varepsilon_i}_{r_{n1}^b}.\end{aligned}$$

By Assumption 3, $r_{n1}^b = \sum_{i=1}^n h_{i,n}/\sqrt{n} = o_p(1)$. Next, by Hölder's inequality, Assumption 4 and Lemma 1, we have

$$r_{n1}^a \le \left\|r_v\mathbb{E}\big(e_1^{\mathrm{T}}\widehat{\Sigma}_{\widetilde{V},\widetilde{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widetilde{M}}|\boldsymbol{\mathcal{X}}\big) - \eta_n\right\|_1 \cdot \|\boldsymbol{Z}^{\mathrm{T}}\varepsilon/\sqrt{n}\|_\infty = o_p(1).$$

Therefore, $r_{n1} = o_p(1)$.

Step 4. (Behavior of r_{n2}.) r_{n2} captures the under-fitting bias, and is small by Assumption 5:

$$r_{n2} = \mathbb{E}\Big((D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{V,\widehat{M}})D_V/n)^{-1} \cdot D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{V,\widehat{M}})g_V(\boldsymbol{W})/\sqrt{n}|\boldsymbol{\mathcal{X}}\Big) = o_p(1).$$

Finally, the results in Steps 3 and 4 imply

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

2.5.3 Derivation of (2.7) in Section 2.3.3

In this part, we provide the derivation of expression that includes the influence functions in (2.7) is provided in 2.5.3. Following similar idea to the proof of Theorem 1, by direct calculation

$$\begin{aligned}\sqrt{n}(\widetilde{\alpha}-\alpha)=&\mathbb{E}\left(\sqrt{n}(\widehat{\alpha}_{\widehat{M}}-\alpha_0)|\mathcal{X}\right)\\=&\frac{1}{\sqrt{n}}\xi_n^{\mathrm{T}}\sum_{i=1}^{n}Z_i\varepsilon_i+t_{n1}+t_{n2}+r_{n2},\end{aligned}$$

where

$$\begin{aligned}t_{n1}&=\frac{1}{\sqrt{n}}\left\{\mathbb{E}\left(e_1^{\mathrm{T}}\Sigma_{\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}V_i\middle|\mathcal{X}\right)-\xi_n\right\}^{\mathrm{T}}\sum_{i=1}^{n}Z_i\varepsilon_i,\\t_{n2}&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left\{\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}-e_1^{\mathrm{T}}\Sigma_{\widehat{M}}^{-1}\widetilde{\boldsymbol{I}}_{\widehat{M}}\right)^{\mathrm{T}}V_iZ_i\varepsilon_i\middle|\mathcal{X}\right\}.\end{aligned}$$

In this expression, t_{n1} can be bounded following similar steps in Section B.1 under Assumptions 7 and 8. In t_{n2}, let $B=\begin{pmatrix}n\\n_2\end{pmatrix}^{-1}$ and we have

$$\begin{aligned}t_{n2}=&\sum_{b=1}^{B}\mathbb{P}(V=v_b)\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{v_b,\widehat{M}_b}^{-1}-e_1^{\mathrm{T}}\Sigma_{\widehat{M}_b}^{-1}\right)^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n}v_{ib}\varepsilon_iZ_{i,\widehat{M}_b}\\=&\frac{1}{B}\sum_{b=1}^{B}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{v_b,\widehat{M}_b}^{-1}-e_1^{\mathrm{T}}\Sigma_{\widehat{M}_b}^{-1}\right)^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n}v_{ib}\varepsilon_iZ_{i,\widehat{M}_b}=:\frac{1}{B}\sum_{b=1}^{B}t_{n,v_b}.\end{aligned}$$

Denote by $T_{1,b}$ the subsample for model building, and $T_{2,b}$ the subsample for parameter estimation. Define $\mu_{i,b}=e_1^{\mathrm{T}}(\widehat{\Sigma}_{v_b,\widehat{M}_b}^{-1}-\Sigma_{\widehat{M}_b}^{-1})Z_{i,\widehat{M}_b}$ which is independent with $\{\varepsilon_i,\ i\in T_{2,b}\}$, then t_{n,v_b} satisfies

$$\begin{aligned}t_{n,v_b} &= \left(e_1^{\mathrm{T}}\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b} - e_1^{\mathrm{T}}\Sigma^{-1}_{\widehat{M}_b}\right)^{\mathrm{T}} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_{ib}\varepsilon_i Z_{i,\widehat{M}_b} \\ &= \frac{\sqrt{n}}{n_2}\sum_{i\in T_{2,b}} \varepsilon_i\mu_{i,b} := \sqrt{n}u_b\sigma_\varepsilon\frac{1}{n_2}\Big(\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big)^{1/2},\end{aligned}$$

where

$$u_b = \Big(\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big)^{-1/2}\sum_{i\in T_{2,b}}\varepsilon_i\mu_{i,b}.$$

We note that $\{u_1,\cdots,u_B\}$ are dependent but identically distributed random variables. The variance of t_{n,v_b} then equals

$$\mathrm{Var}(t_{n,v_b}) = \mathbb{E}\left(\mathrm{Var}(t_{n,v_b})|\mu_{i,b}\right) + \mathrm{Var}\left(\mathbb{E}(t_{n,v_b})|\mu_{i,b}\right) = \frac{n\sigma_\varepsilon^2}{n_2}\mathbb{E}\Big(\frac{1}{n_2}\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big).$$

Next, we provide an upper bound for $\sum_{i\in T_{2,b}}\mu_{i,b}^2/n_2$. Note that

$$\begin{aligned}\frac{1}{n_2}\sum_{i\in T_{2,b}}\mu_{i,b}^2 &= e_1^{\mathrm{T}}\Big(\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b} - \Sigma^{-1}_{\widehat{M}_b}\Big)\frac{1}{n_2}\sum_{i\in T_{2,b}} Z_{i,\widehat{M}_b}Z^{\mathrm{T}}_{i,\widehat{M}_b}\Big(\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b} - \Sigma^{-1}_{\widehat{M}_b}\Big)e_1 \\ &= e_1^{\mathrm{T}}\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b}\Big(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{\widehat{M}_b}\Big)\Sigma^{-1}_{\widehat{M}_b}\Big(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{\widehat{M}_b}\Big)\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b}e_1 \\ &\le \lambda_{\max}\left(\Sigma^{-1}_{\widehat{M}_b}\right)\left\|e_1^{\mathrm{T}}\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b}\left(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{\widehat{M}_b}\right)\right\|_2^2 \\ &\le \lambda^{-1}_{\min}\left(\Sigma_{\widehat{M}_b}\right)\left\|e_1^{\mathrm{T}}\widehat{\Sigma}^{-1}_{v_b,\widehat{M}_b}\right\|_2^2\lambda^2_{\max}\left(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{\widehat{M}_b}\right) \\ &\le K^2/\kappa_0\lambda^2_{\max}\left(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{\widehat{M}_b}\right),\end{aligned}$$

where the last step is obtained by Assumption 9. Since the covariates are bounded by C, by applying Corollary 5.50 in Vershynin (2016), we obtain that under Assumption 6 and the fact that $\widehat{M}_b$ is selected independent with $\widehat{\Sigma}_{v_b}$, $\mathbb{P}\{\lambda_{\max}(\widehat{\Sigma}_{v_b,\widehat{M}_b} - \Sigma_{M_b}) \ge \epsilon\} \le$

$2\exp(-n\varepsilon^2/C)$. Therefore for all possible models,

$$\mathbb{P}\left[\lambda_{\max}\left(\widehat{\Sigma}_{v_b,\widehat{M}_b}-\Sigma_{\widehat{M}_b}\right)\geq\epsilon \text{ for any } \widehat{M}_b\right]\leq 2\exp\left(s\log p-\frac{n\epsilon^2}{C}\right), \qquad (2.16)$$

by Assumption 6, and the above upper bound converges to 0 as $n\to\infty$. Therefore with probability tending to 1, for all $\epsilon>0$ and for all $v_b\in\mathcal{V}_{n_2}$, $n_2^{-1}\sum_{i\in T_{2,b}}\mu_{i,b}^2$ is bounded by ϵ. By letting $H_n^\epsilon=\{\text{for all } v_b\in\mathcal{V}_{n_2}, n_2^{-1}\sum_{i\in T_{2,b}}\mu_{i,b}^2\leq\epsilon\}$, we have $\mathbb{P}(H_n^\epsilon)\to 1$ as $n\to\infty$. For any $\epsilon_0>0$,

$$\begin{aligned}
&\mathbb{P}(t_{n2}>\epsilon_0)\\
=&\,\mathbb{P}\Big\{\frac{1}{B}\sum\nolimits_{b=1}^{B}u_b\sigma_\varepsilon n_2^{-1}\Big(\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big)^{1/2}>n^{-1/2}\epsilon_0\Big\}\\
\leq&\,\mathbb{P}\Big\{\sigma_\varepsilon\Big(\frac{1}{B}\sum\nolimits_{b=1}^{B}u_b^2\Big)^{1/2}\Big(\frac{1}{B}\sum\nolimits_{b=1}^{B}\frac{1}{n_2}\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big)^{1/2}>n^{-1/2}n_2^{1/2}\epsilon_0\Big\}\\
\leq&\,\mathbb{P}\Big\{\sigma_\varepsilon\Big(\frac{1}{B}\sum\nolimits_{b=1}^{B}u_b^2\Big)^{1/2}\Big(\frac{1}{B}\sum\nolimits_{b=1}^{B}\frac{1}{n_2}\sum\nolimits_{i\in T_{2,b}}\mu_{i,b}^2\Big)^{1/2}>n^{-1/2}n_2^{1/2}\epsilon_0\;\Big|\;H_n^\epsilon\Big\}\mathbb{P}(H_n^\epsilon)+\mathbb{P}(H_n^{\epsilon,c})\\
\leq&\,\mathbb{P}\Big\{\sigma_\varepsilon\epsilon\Big(\frac{1}{B}\sum\nolimits_{b=1}^{B}u_b^2\Big)^{1/2}>n^{-1/2}n_2^{1/2}\epsilon_0\Big\}\mathbb{P}(H_n^\epsilon)+\mathbb{P}(H_n^{\epsilon,c})\\
\leq&\,\frac{\epsilon^2\sigma_\varepsilon^2\text{var}\Big\{\left(\frac{1}{B}\sum_{b=1}^{B}u_b^2\right)^{1/2}\Big\}}{n^{-1}n_2\epsilon_0^2}\mathbb{P}(H_n^\epsilon)+\mathbb{P}(H_n^{\epsilon,c})\\
=&\,O(\epsilon^2\epsilon_0^{-2})\mathbb{P}(H_n^\epsilon)+\mathbb{P}(H_n^{\epsilon,c}).
\end{aligned}$$

Therefore, by letting ϵ go to zero, we have $t_{n2}=o_p(1)$. This completes the proof.

2.5.4 Derivation of (2.9) in Section 2.3.3

In this subsection, we provide the derivation of the comparison between cross-estimation and R-Split is provided in 2.5.4. Since V and $V^c=\{1-V_1,\cdots,1-V_n\}$ are identically distributed random vectors, then

$$\mathbb{E}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}}-\alpha)|\boldsymbol{\mathcal{X}})=\sqrt{n}\mathbb{E}\left\{e_1^{\mathrm{T}}\left(\frac{1}{n}\sum_{i=1}^{n}V_iZ_{i,\widehat{M}_1}Z_{i,\widehat{M}_1}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}V_iZ_{i,\widehat{M}_1}\varepsilon_i\middle|\boldsymbol{\mathcal{X}}\right\}=\sqrt{n}(\widetilde{\alpha}-\alpha).$$

Thus Var $\{\mathbb{E}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)|\boldsymbol{\mathcal{X}})\} = \text{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha))$, and

$$\begin{aligned}
\text{Var}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)) &= \mathbb{E}\left\{\text{Var}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)|\boldsymbol{\mathcal{X}})\right\} + \text{Var}\left\{\mathbb{E}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)|\boldsymbol{\mathcal{X}})\right\} \\
&= \mathbb{E}\left\{\text{Var}(\sqrt{n}(\widetilde{\alpha}_{\text{cv}} - \alpha)|\boldsymbol{\mathcal{X}})\right\} + \text{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha)) + \geq \text{Var}(\sqrt{n}(\widetilde{\alpha} - \alpha)).
\end{aligned}$$

2.5.5 Derivation of (3.12) in Section 3.2

In this part, we provide the proof of the alternative expression for the variance of R-Split that has been shown in Section 2.5.5. Recall the additional assumptions we made in Section 5.

Assumption 10. *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,*

$$\left\|\mathbb{E}\left\{\frac{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})X_V/n}{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})D_V/n}\middle|\mathcal{X}\right\}\right\|_\infty = O_p(\sqrt{\log p}). \tag{2.17}$$

Assumption 11. *The maximal $s-$sparse eigenvalue satisfies* $\mathbb{P}(\limsup_{n\to\infty} \lambda_{\max,s}(X^{\mathrm{T}}X/n) \leq K_0) = 1$. *The maximum eigenvalue of* $\widehat{\Sigma}$ *is bounded by* $\log p$ *in probability.*

We start with generalizing the result stated in (2.16). For all possible models,

$$\begin{aligned}
&\mathbb{P}\left\{\lambda_{\max}\left(\widehat{\Sigma}_{V,\widehat{M}_V} - \widehat{\Sigma}_{n,\widehat{M}_V}\right) \geq \epsilon\right\} \qquad (2.18)\\
\leq&\mathbb{P}\left\{\lambda_{\max}\left(\widehat{\Sigma}_{V,\widehat{M}_V} - \Sigma_{\widehat{M}_V}\right) \geq \epsilon/2\right\} + \mathbb{P}\left\{\lambda_{\max}\left(\widehat{\Sigma}_{n,\widehat{M}_V} - \Sigma_{\widehat{M}_V}\right) \geq \epsilon/2\right\} \\
\leq&\mathbb{E}\,\mathbb{P}\left\{\lambda_{\max}\left(\widehat{\Sigma}_{V,\widehat{M}_V} - \Sigma_{\widehat{M}_V}\right) \geq \epsilon/2 \mid V\right\} + \mathbb{E}\,\mathbb{P}\left\{\lambda_{\max}\left(\widehat{\Sigma}_{n,\widehat{M}_V} - \Sigma_{\widehat{M}_V}\right) \geq \epsilon/2 \mid V\right\} \\
\leq&4\exp\left(s\log p - \frac{n_2\epsilon^2}{C}\right) = 4\exp\left(s\log p - \frac{r_v n\epsilon^2}{C}\right) = o(1).
\end{aligned}$$

Let $\widehat{\eta}_n = r_v\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}^{-1}_{V,\widehat{M}}\mathbf{I}_{\widehat{M}}|\mathcal{X}\right)^{\mathrm{T}}$, by Assumption 4, we have $||\eta_n - \widehat{\eta}_n||_1 = o_p(1/\sqrt{\log p})$.

We next prove the followings

$$\eta_n^{\mathrm{T}}\widehat{\Sigma}_n\eta_n = \widehat{\eta}_n^{\mathrm{T}}\widehat{\Sigma}_n\widehat{\eta}_n + o_p(1), \tag{2.19}$$

$$\widehat{\eta}_n^{\mathrm{T}}\widehat{\Sigma}_n\widehat{\eta}_n \leq r_v^2\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}\widehat{\Sigma}_{\widehat{M}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1|\mathcal{X}\right) = \mathbb{E}\left(r_v e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1|\mathcal{X}\right) + o_p(1). \tag{2.20}$$

With (2.19) and (2.20), if we further assume the selected model size satisfies ultra-sparsity in the sense that $|\widehat{M}|\log p/\sqrt{n} = o(1)$, we obtain

$$\mathbb{E}\left(r_v e_1^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1|\mathcal{X}\right) = \mathbb{E}\left\{(\Sigma_{\widehat{M}}^{-1})_{11}\bigg|\mathcal{X}\right\} + o_p(1)$$

by Lemma 2 under Assumptions 1, 4 and 13. Therefore, we have $\widetilde{\sigma}_n^2 \leq \sigma_\varepsilon^2\mathbb{E}\left\{(\Sigma_{\widehat{M}}^{-1})_{11}\bigg|\mathcal{X}\right\}+ o_p(1)$. We next prove (2.19) in step 1 and prove (2.20) in step 2.

Step 1. In (2.19), by Assumptions 11 and 4,

$$\begin{aligned}\eta_n^{\mathrm{T}}\widehat{\Sigma}_n\eta_n - \widehat{\eta}_n^{\mathrm{T}}\widehat{\Sigma}_n\widehat{\eta}_n &= (\eta_n-\widehat{\eta}_n)^{\mathrm{T}}\widehat{\Sigma}_n(\eta_n-\widehat{\eta}_n) + 2(\eta_n-\widehat{\eta}_n)^{\mathrm{T}}\widehat{\Sigma}_n\widehat{\eta}_n\\ &\leq \lambda_{\max}(\widehat{\Sigma}_n)||\eta_n-\widehat{\eta}_n||_1^2 + ||\eta_n-\widehat{\eta}_n||_1||\widehat{\Sigma}_n\widehat{\eta}_n||_\infty\\ &= o_p(1) + ||\eta_n-\widehat{\eta}_n||_1||\widehat{\Sigma}_n\widehat{\eta}_n||_\infty,\end{aligned}$$

In the second part,

$$\begin{aligned}\widehat{\Sigma}_n\widehat{\eta}_n =& r_v\mathbb{E}\left(\widehat{\Sigma}_n I_{\widehat{M}}^{\mathrm{T}}\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1|\mathcal{X}\right)\\ =&\mathbb{E}\left\{\left(\frac{1}{n}\sum_{i=1}^n Z_iZ_{i,\widehat{M}}^{\mathrm{T}} - \frac{1}{n_2}\sum_{i=1}^n V_iZ_iZ_{i,\widehat{M}}^{\mathrm{T}}\right)\left(\frac{1}{n_2}\sum_{i=1}^n V_iZ_{i,\widehat{M}}Z_{i,\widehat{M}}^{\mathrm{T}}\right)^{-1}e_1\bigg|\mathcal{X}\right\}\\ &+\mathbb{E}\left\{\left(\frac{1}{n_2}\sum_{i=1}^n V_iZ_iZ_{i,\widehat{M}}^{\mathrm{T}}\right)\left(\frac{1}{n_2}\sum_{i=1}^n V_iZ_{i,\widehat{M}}Z_{i,\widehat{M}}^{\mathrm{T}}\right)^{-1}\bigg|\mathcal{X}\right\}\\ :=&q_{n1}^a + q_{n1}^b,\end{aligned}$$

where q^b_{n1} can be further simplified

$$q^b_{n1} = \mathbb{E}\left\{ \frac{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})X_V}{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})D_V} \middle| \mathcal{X} \right\},$$

and satisfies $\left|\left|q^b_{n1}\right|\right|_\infty = O_p(\sqrt{\log p})$ by Assumption 3.11. In q^a_{n1},

$$\begin{aligned}
||q^a_{n1}||_\infty = & \mathbb{E}\left\{ \left(\frac{1}{n}\sum_{i=1}^n Z_i Z^{\mathrm{T}}_{i,\widehat{M}} - \frac{1}{n_2}\sum_{i=1}^n V_i Z_i Z^{\mathrm{T}}_{i,\widehat{M}} \right) \left(\frac{1}{n_2}\sum_{i=1}^n V_i Z_{i,\widehat{M}} Z^{\mathrm{T}}_{i,\widehat{M}} \right)^{-1} e_1 \middle| \mathcal{X} \right\} \\
& \leq r_v \mathbb{E}\left\{ ||\widehat{\Sigma}^{-1}_{V,\widehat{M}} e_1||_2 \max_j ||Z_j^{\mathrm{T}} \boldsymbol{Z}_{\widehat{M}}/n - \Sigma_{j,\widehat{M}}||_2 \middle| \mathcal{X} \right\} \\
& \quad + r_v \mathbb{E}\left\{ ||\widehat{\Sigma}^{-1}_{V,\widehat{M}} e_1||_2 \max_j ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,\widehat{M}}/n_2 - \Sigma_{j,\widehat{M}}||_2 \middle| \mathcal{X} \right\} \\
& \lesssim_P r_v K \cdot \max_j \max_{|M|\leq s} ||Z_j^{\mathrm{T}} \boldsymbol{Z}_M/n - \Sigma_{j,M}||_2 \\
& \quad + r_v K \mathbb{E}\left\{ \max_j \max_{|M|\leq s} ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}||_2 \middle| \mathcal{X} \right\} \\
& \lesssim_P o_p(1) + r_v K \mathbb{E}\left\{ \max_j \max_{|M|\leq s} ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}||_2 \middle| \mathcal{X} \right\},
\end{aligned}$$

where the last inequality is obtained by Lemma 2. Define an event

$$\begin{aligned}
& \mathbb{E}\left\{ \max_j \max_{|M|\leq s} ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}||_2 \right\} \\
= & \mathbb{E}\left\{ \max_j \max_{|M|\leq s} ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}||_2 \,\middle|\, L_n^\epsilon \right\} \mathbb{P}(L_n^\epsilon) \\
& \quad + \mathbb{E}\left\{ \max_j \max_{|M|\leq s} ||Z^{\mathrm{T}}_{j,V} \boldsymbol{Z}_{V,M}/n_2 - \Sigma_{j,M}||_2 \,\middle|\, L_n^{\epsilon,c} \right\} \mathbb{P}(L_n^{\epsilon,c}) \\
\leq & s^{1/2}(C + ||\Sigma||_\infty) \exp\left(\log p + s\log p + \log s - \frac{r_v n \epsilon^2}{2C^2 s} \right) + \epsilon = o(1);
\end{aligned}$$

by letting ϵ go to zero, then we get $||q^a_{n1}||_\infty = o_p(1)$. This completes the proof of (2.19).

Step 2. For the first part of (2.20), we have

$$
\begin{aligned}
\widehat{\eta}_n^{\mathrm{T}} \widehat{\Sigma}_n \widehat{\eta}_n =& r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widetilde{\boldsymbol{I}}_{\widehat{M}} | \mathcal{X} \right) \frac{1}{n} \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} \mathbb{E}\left(\widetilde{\boldsymbol{I}}_{\widehat{M}}^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) \\
=& r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \boldsymbol{Z}_{\widehat{M}}^{\mathrm{T}} / \sqrt{n} | \mathcal{X} \right) \mathbb{E}\left(\boldsymbol{Z}_{\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 / \sqrt{n} | \mathcal{X} \right) \\
\leq& r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right).
\end{aligned}
$$

Next we prove the difference between $r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{\widehat{M}}^{-1} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right)$ and $\mathbb{E}\left(r_v e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right)$ is negligible. Consider

$$
\begin{aligned}
& r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) - \mathbb{E}\left(r_v e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) \\
=& r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) - \mathbb{E}\left(r_v e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{V,\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) \\
=& r_v^2 \mathbb{E}\left\{ e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \left(\widehat{\Sigma}_{\widehat{M}} - \frac{1}{r_v} \widehat{\Sigma}_{V,\widehat{M}} \right) \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right\} \\
=& r_v^2 \mathbb{E}\left\{ e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} - \frac{1}{n_2} \sum_{i=1}^{n} V_i Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} \right) \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right\} \\
\leq& r_v^2 \mathbb{E}\left\{ \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} - \frac{1}{n_2} \sum_{i=1}^{n} V_i Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} \right) ||\widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1||_2^2 | \mathcal{X} \right\} \\
\leq& r_v^2 \mathbb{E}\left\{ \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} - \frac{1}{n_2} \sum_{i=1}^{n} V_i Z_{i,\widehat{M}} Z_{i,\widehat{M}}^{\mathrm{T}} \right) K^2 | \mathcal{X} \right\}
\end{aligned}
$$

By letting $G_n^{\epsilon} = \left\{ \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^{n} Z_{i,\widehat{M}_V} Z_{i,\widehat{M}_V}^{\mathrm{T}} - \frac{1}{n_2} \sum_{i=1}^{n} V_i Z_{i,\widehat{M}_V} Z_{i,\widehat{M}_V}^{\mathrm{T}} \right) K^2 \leq \epsilon \right\}$, by (2.18), we have $\mathbb{P}(G_n^{\epsilon}) \to 1$ as $n \to \infty$. Therefore, since the maximum eigenvalues are bounded above, we obtain $r_v^2 \mathbb{E}\left(e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} \widehat{\Sigma}_{\widehat{M}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) - \mathbb{E}\left(r_v e_1^{\mathrm{T}} \widehat{\Sigma}_{V,\widehat{M}}^{-1} e_1 | \mathcal{X} \right) = o_p(1)$, which is (2.20).

2.5.6 Derivation of variance estimation in R-Split via the non-parametric delta method in Section 2.3.1

Recall $v_b = (v_{b1}, \cdots, v_{bn})$ is the data splitting weights in the bth split: if $v_{bi} = 1$, data point $\mathcal{X}_i$ is used in the refitting step. v_b takes value from sample space $\mathcal{V}_{n_2}$, where $n_2 = \sum_{i=1}^{n} v_{bi}$ denotes the total number of samples used for refitting, and

$$\mathcal{V}_{n_2} = \left\{ V = (V_1, \cdots, V_n) \in R^n : \; V_i \in \{0, 1\}, \; \sum\nolimits_{i=1}^{n} V_i = n_2 \right\}.$$

Let $B = \binom{n}{n_2}$ and there are B components in $\mathcal{V}_{n_2}$. Since all the weights are independently generated, we have $\mathrm{pr}(V = v_b) = 1/B$, for $b = 1, \ldots, B$. Following the definition in Efron (2014), the ideal smoothed estimation is $\widetilde{\alpha} = \sum_{b=1}^{B} \mathbb{P}(V = v_b)\widehat{\alpha}_b$, and it can be viewed as a functional of the empirical distribution $\widehat{F}_n$, denotes as $T(\widehat{F}_n)$. When adding a point mass $\delta_{\mathcal{X}_j}$ at direction j, $j \in \{1, \ldots, n\}$, the empirical distribution $\widehat{F}_n$ changes to $(1-\epsilon)\widehat{F}_n + \epsilon\delta_{\mathcal{X}_j}$, and the influence function can be written as

$$U(\mathcal{X}_j, \widehat{F}_n) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)\widehat{F}_n + \epsilon\delta_{\mathcal{X}_j}) - T(\widehat{F}_n)}{\epsilon}. \tag{2.21}$$

Data splitting takes n_2 subsamples without replacement and without regard to the order. The subsamples can be viewed as taken all at once from the entire population of n objects, while each sample shares the same probability being chosen. Thus, the average number of a given sample $\mathcal{X}_j$ in the subsample E (#of $\mathcal{X}_j$ in the subsamples of size n_2) $= n_2/n$.

If a point mass $\epsilon\delta_{\mathcal{X}_j}$ is added in $\widehat{F}_n$, this means the probability of $\mathcal{X}_j$ increases from $1/n$ to $(1-\epsilon)/n + \epsilon$, and the probability of the other objects decreases from $1/n$ to $(1-\epsilon)/n$. We denote the perturbed empirical distribution function as $\widehat{F}_n^j$. After the perturbation,

$$\begin{aligned} & E\left(\#\text{of } \mathcal{X}_j \text{ in the subsamples of size } n_2\right) \qquad (2.22) \\ =& n_2\left(\frac{1-\epsilon}{n} + \epsilon\right) = \mathbb{P}\left(\mathcal{X}_j \text{ being selected in the subsamples of size } n_2 \text{ under } \widehat{F}_n^j\right). \end{aligned}$$

Define a subset $\mathcal{B}_j = \{v_b : v_b \in \mathcal{V}_{n_2} \text{and } v_{bj} = 1\} \subset \mathcal{V}_{n_2}$, which indexes the entire possible combinations that include $\mathcal{X}_j$. The cardinality of the subset $\mathcal{B}_j$ equals to

$$\binom{n-1}{n_2-1} = \frac{Bn_2}{n}. \tag{2.23}$$

After the perturbation on the jth direction, in $\mathcal{V}_{n_2}$, only the elements with $V_j = 1$ share the same probability of being chosen. This gives $\mathbb{P}(V = v_{b_1}, \mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j) = \mathbb{P}(V = v_{b_2}, \mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j)$, for all $b_1, b_2 \in \mathcal{B}_j$, and

$$\sum\nolimits_{b \in \mathcal{B}_j} \mathbb{P}(V = v_b, \mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j) = \mathbb{P}(\mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j). \tag{2.24}$$

From (2.22)–(2.24), we have

$$\mathbb{P}(V = v_b, \mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j) = n_2 \left\{(1-\epsilon)/n + \epsilon\right\} / \{Bn_2/n\},$$

and similarly $\mathbb{P}(V = v_b, \mathcal{X}_j \text{ not being chosen under } \widehat{F}_n^j) = \{1 - n_2 \left\{(1-\epsilon)/n + \epsilon\right\}\} / \{B - Bn_2/n\}$. Hence, after adding a small perturbation on jth direction,

$$\begin{aligned}\mathbb{P}(V = v_b \text{ under } \widehat{F}_n^j) =& v_{bj} \mathbb{P}(V = v_b, \mathcal{X}_j \text{ being chosen under } \widehat{F}_n^j) \\ &+ (1 - v_{bj}) \mathbb{P}(V = v_b, \mathcal{X}_j \text{ not being chosen under } \widehat{F}_n^j) \\ =& v_{bj} \frac{n_2 \left(\frac{1-\epsilon}{n} + \epsilon\right)}{Bn_2/n} + (1 - v_{bj}) \frac{1 - n_2 \left(\frac{1-\epsilon}{n} + \epsilon\right)}{B - Bn_2/n} \\ =& B^{-1} \left\{1 + \epsilon \frac{n(n-1)}{n - n_2} (v_{bj} - \frac{n_2}{n})\right\}.\end{aligned}$$

Using (2.21),

$$\begin{aligned}
U(\mathcal{X}_j, \widehat{F}_n) =& \lim_{\epsilon\to 0} \epsilon^{-1}\{T((1-\epsilon)\widehat{F}_n + \epsilon\delta_{\mathcal{X}_j}) - T(\widehat{F}_n)\} \\
=& \lim_{\epsilon\to 0} \epsilon^{-1} \sum\nolimits_{b=1}^{B} \left\{\mathbb{P}(V = v_b \text{ under } \widehat{F}_n^j) - \mathbb{P}(V = v_b \text{ under } \widehat{F}_n)\right\} \widehat{\alpha}_b \\
=& \frac{n(n-1)}{n-n_2} \frac{1}{B} \sum_{b=1}^{B} \left(v_{bj} - \frac{n_2}{n}\right) \widehat{\alpha}_b \\
=& \frac{n(n-1)}{n-n_2} \text{cov}(\widehat{\alpha}, V_j).
\end{aligned}$$

The nonparametric delta method suggests the standard deviation of the smoothed estimator to be estimated by

$$n^{-1}\sum\nolimits_{j=1}^{n} U^2(\mathcal{X}_j, \widehat{F}_n) = n\sum\nolimits_{j=1}^{n} \left\{\frac{n-1}{n-n_2}\widehat{S}_j(V, \mathcal{X})\right\}^2 \tag{2.25}$$

where the covariance can be estimated by the data splitting samples

$$\widehat{S}_j(V, \mathcal{X}) = B^{-1}\sum\nolimits_{b=1}^{B}(v_{bj} - B^{-1}\sum\nolimits_{k=1}^{B} v_{kj})\widehat{\alpha}_b. \tag{2.26}$$

In practice, the smoothed estimator are computed using a finite number B of the data splitting, and working with a large B can be computationally expensive. Without sufficient number of splitting, the formula in (2.26) is biased upward argued in Wager et al. (2014). Following similar method as in Wager et al. (2014), the Monte Carlo bias in M-Split can be corrected and the variance can be estimated through equation (2.3).

CHAPTER 3

Debiased Inference

In this chapter, we revisit the debiased inference procedure and discuss the connection between some popular methods in the literature. Then we propose a projection assisted double-selection (PODS) approach to have a strong control over the over-fitting bias, followed by a comparison between PODS and R-Split.

3.1 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 estimate 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.

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

$$\widehat{\alpha}_{\text{full}} = (\widehat{D}^{\mathrm{T}}D)^{-1}\widehat{D}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\text{full}}), \tag{3.1}$$

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 possible 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 can be written as

$$\widehat{\alpha}_{\text{double}} = (\widehat{D}_{\widehat{M}}^{\mathrm{T}}D)^{-1}\widehat{D}_{\widehat{M}}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\widehat{M}}), \tag{3.2}$$

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 $\widehat{\gamma}_{-\widehat{M}} = 0$.

Furthermore, $\widehat{\alpha}_{\text{double}}$ satisfies

$$\breve{\sigma}_n^{-1}\sqrt{n}(\widehat{\alpha}_{\text{double}} - \alpha) \rightsquigarrow N(0,1), \quad \breve{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{1}{||D - \boldsymbol{X}\widehat{\gamma}||_2^2/n} + o_p(1). \tag{3.3}$$

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

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

for some tuning constant λ_d. Then we have $e_1^{\mathrm{T}}\widehat{\boldsymbol{\Theta}} = \widehat{\nu}^{-2}(1, -\widehat{\gamma}_{\text{lasso}}^{\mathrm{T}})$, where $\widehat{\nu}^2 = \widehat{D}_{\text{lasso}}^{\mathrm{T}}D/n$, and the de-sparsified Lasso estimator for α can be written as

$$\begin{aligned}\widehat{\alpha}_{\text{desparse}} &= \widehat{\alpha}_{\text{lasso}} + e_1^{\mathrm{T}}\widehat{\boldsymbol{\Theta}}(D,\boldsymbol{X})^{\mathrm{T}}(D,\boldsymbol{X})(Y - \widehat{\alpha}_{\text{lasso}}D - \boldsymbol{X}\widehat{\beta}_{\text{lasso}})/n \\ &= (\widehat{D}_{\text{lasso}}^{\mathrm{T}}D)^{-1}\widehat{D}_{\text{lasso}}^{\mathrm{T}}(Y - \boldsymbol{X}\widehat{\beta}_{\text{lasso}}). \end{aligned} \tag{3.4}$$

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

$$\ddot{\sigma}_n^{-1}\sqrt{n}(\widehat{\alpha}_{\text{desparse}} - \alpha) \rightsquigarrow N(0,1), \quad \ddot{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{||D - \boldsymbol{X}\widehat{\gamma}_{\text{lasso}}||_2^2/n}{(||D - \boldsymbol{X}\widehat{\gamma}_{\text{lasso}}||_2^2/n + \lambda_d||\widehat{\gamma}_{\text{lasso}}||_1)^2}. \tag{3.5}$$

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 (3.3).

It follows from (3.2) and (3.4) 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

$$D = \boldsymbol{X}\gamma + \nu, \quad \mathbb{E}(\nu|\boldsymbol{X}) = 0, \quad \mathrm{Cov}(\nu) = \sigma_\nu^2 \boldsymbol{I}, \tag{3.6}$$

for some constant σ_ν^2. A good estimation of γ helps reduce the under-fitting term 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 equivalently to (3.1), 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 and van de Geer (2017) 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 (3.5) 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 assisted double-selection (PODS) as an enhancement of the post-double-selection of Belloni et al. (2014).

3.1.2 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, 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 with spurious correlation with D are less likely to enter $\widehat{M}$ and the risk of over-fiting is reduced.

3.1.2.1 Proposed Method

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_{n2} can be decomposed

$$\begin{aligned} b_{n2} = & \frac{1}{k_{n1}} \underbrace{\frac{1}{\sqrt{n}} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D}) \varepsilon}_{(I)} \\ & + \frac{1}{k_{n1}} \underbrace{\frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D}) \boldsymbol{X}_{\widehat{M}_Y}}_{(II)} \cdot (\boldsymbol{X}_{\widehat{M}_Y}^{\mathrm{T}} \boldsymbol{X}_{\widehat{M}_Y}/n)^{-1} \cdot \underbrace{\frac{1}{\sqrt{n}} \boldsymbol{X}_{\widehat{M}_Y} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D}) \varepsilon}_{(III)}, \end{aligned} \tag{3.7}$$

where $k_{n1} = D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}})D/n$ is a scaler. In (3.7), (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 selected variables in $\widehat{M}_Y$, the over-fitting bias can be reduced to a smaller scale. This motivates our proposed method of projection assisted double-selection (PODS).

A formal algorithm of PODS is given in Algorithm 3, 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, Lasso, or iterated Lasso, which is a tuning free method discussed in Belloni et al. (2014). 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 is expected to have low correlation with D, and then the over-fitting bias can be controlled. 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 3 PODS

```
    Step 1. Select a set of variables $\widehat{M}_D$ for the 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 $\boldsymbol{X}_{\widehat{M}^*}$ to get $\widehat{\alpha}$, which is the estimated coefficient of $D$.
```

The asymptotic variance of $\widehat{\alpha}$ from PODS can be estimated by

$$\widetilde{\sigma}_n^2 = \frac{\widehat{\sigma}_\varepsilon^2}{||D - \boldsymbol{X}\widehat{\gamma}^*||_2^2/n},$$

where $\widehat{\sigma}_\varepsilon^2 = Y^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}^*})Y \cdot n/(n - |\widehat{M}^*| - 1)$, and $\widehat{\gamma}^* \in \mathbb{R}^p$ is a sparse vector with $\widehat{\gamma}^*_{\widehat{M}^*} = (\boldsymbol{X}^{\mathrm{T}}_{\widehat{M}^*}\boldsymbol{X}_{\widehat{M}^*})^{-1}\boldsymbol{X}_{\widehat{M}^*}D$ and $\widehat{\gamma}^*_{-\widehat{M}^*} = 0$. To better understand the difference between PODS and the post-double-selection, we shall take a look at the difference between the distributions of $\widehat{M}_Y^*$ and $\widehat{M}_Y$.

Example 3 (Difference between $\widehat{M}_Y$ and $\widehat{M}_Y^*$)**.** *Consider the following model*

$$\begin{aligned} Y_i &= \alpha D_i + \varepsilon_i, \\ D_i &= \gamma_1 X_{i1} + \nu_i, \end{aligned} \tag{3.8}$$

where $(\varepsilon_i, \nu_i) \sim N(0, \boldsymbol{I}_2)$*, for* $i = 1, \cdots, n$*, where* $\boldsymbol{I}_2$ *is the 2 by 2 identity matrix, and* X_{i1} *is just the first component of* X_i*. Suppose that we perform model selection in Step 1 by marginal screening. Let* $\widehat{r}_{D,j} = |\widehat{corr}_n(D, X_{\cdot j})|$*, for* $j = 1, \cdots, p$*, and* $\widehat{r}_D = (\widehat{r}_{D,1}, \cdots, \widehat{r}_{D,p})$*. Now consider selecting two covariates, and*

$$\widehat{M}_D = \{1 \leq j \leq p : \ \widehat{r}_{D,j} \text{ is among the two largest elements of } \widehat{r}_D\}.$$

Then, the post-projection variables are $\boldsymbol{X}^*$ *and* Y^* *through* $\boldsymbol{X}^* = (\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})X$, $Y^* = (\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})Y$*. In the second step of model selection, for simplicity, we select one covariate from* $\boldsymbol{X}^*$ *in addition to* D*,that is,*

$$\widehat{M}^*_Y = \{\arg\max_{1\leq j\leq p}|\widehat{corr}_n(Y^*, X^*_j)|\}.$$

As for the post-double-selection, we have

$$\widehat{M}_Y = \{\arg\max_{1\leq j\leq p}|\widehat{corr}_n(Y, X_j)|\},$$

From (3.7) we note that (II)*, which is the partial sample covariance between* D *and the selected variable in the second stage, plays a key role in reducing the over-fitting bias. Therefore, for PODS and the post-double-selection, we need to compare* $\widehat{\rho}_{pods} = \frac{1}{n}D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})X_{\widehat{M}^*_Y}$ *and* $\widehat{\rho}_{double} = \frac{1}{n}D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})X_{\widehat{M}_Y}$.

(1). Simple heuristics. From a theoretical point of view, since the post-double-selection picks the variable that maximizes the absolute correlation with Y, and thus the selected variable in $\widehat{M}_Y$ is correlated with both D and the noise ε. In contrast, PODS selects the model

$$\widehat{M}^*_Y = \left\{\arg\max_{1\leq j\leq p}\left|\frac{1}{nc^*_{nj}}X^{\mathrm{T}}_j(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})Y\right|\right\},$$

where c^*_{nj} is the product of the variances of $(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})X_j$ and $(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}_D})Y$. The projection matrix $\boldsymbol{P}^*_{\widehat{M}_D}$ can be decomposed into

$$\boldsymbol{P}^*_{\widehat{M}_D} = -\boldsymbol{P}_{\widehat{M}_D} + \frac{1}{nk_{n2}}(D - \boldsymbol{P}_{\widehat{M}_D}D)(D - \boldsymbol{P}_{\widehat{M}_D}D)^{\mathrm{T}},$$

where $k_{n2} = D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})D/n$. That is, PODS selects

$$\widehat{M}_Y^* = \left\{ \arg\max{}_{1\le j\le p} \left| \frac{1}{nc_{nj}^*} X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon - \frac{1}{nc_{nj}^*} X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D}) D \cdot \frac{1}{k_{n2}} \underbrace{\frac{1}{n} D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon}_{O_p(1/\sqrt{n})} \right| \right\}. \tag{3.9}$$

Since the second term in the above expression is of smaller order than the first term (see the arguments in Appendix 3.5.3), it means that the selected variable from the second stage is mostly determined by $X_j^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}_D})\varepsilon$, which is little correlated with D. Therefore, PODS reduces the over-fitting bias.

(2). Numerical evidence. Next, we provide a simulation study to support the heuristics given above. In model (3.8), let $(n, p) = (100, 1000)$, $\alpha = 1$, $\gamma_1 = 1$ and $(D_i, X_i) \sim N(0, \boldsymbol{I}_{p+1})$ independent of (ε_i, ν_i), for $i = 1, \cdots, n$. A numerical comparison between $\widehat{\rho}_{\text{double}}$ and $\widehat{\rho}_{\text{pods}}$ based on 1000 Monte Carlo samples is presented in Figure 3.1.

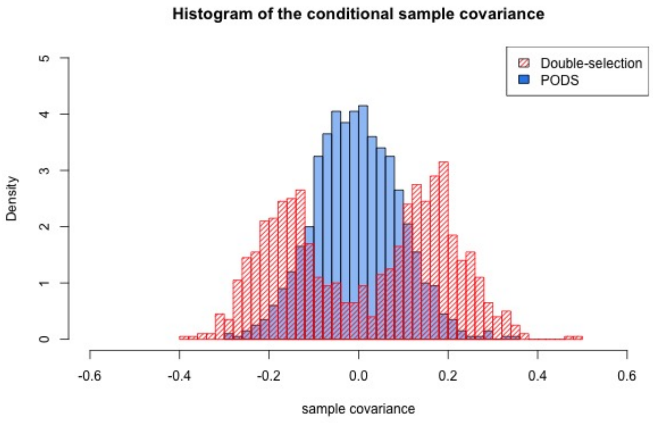


Figure 3.1: Based on 1000 Monte Carlo samples, the area of shading lines is the histogram of $\widehat{\rho}_{\text{double}}$, and the area with solid blue filling is the histogram of $\widehat{\rho}_{\text{pods}}$. The data generating process is given in Example 3.

The result in Figure 3.1 says that the distribution of $\widehat{\rho}_{\text{pods}}$ is centered around 0, while the distribution of $\widehat{\rho}_{\text{double}}$ clearly has two modes, neither of which centers around 0. This suggests that, the variable selected by PODS in the second stage tends to have a smaller correlation with D than the variable selected by the post-double-selection in general.

It is worth noting that the linear projection approach is also adopted in the correlated projection screening method (CPS) 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.

3.1.2.2 Theoretical investigation of PODS

We first introduce some additional notations 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 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 let the minimal $s-$sparse eigenvalue of a semi-positive definite matrix A as

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

We make the following assumptions to study the theoretical property of PODS.

Assumption 12. *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 13. *The cardinality of $\widehat{M}_Y^*$ is of the same order as s_0, which satisfies $s_0 \log p = o(\sqrt{n})$.*

Assumption 14. *There exists a positive constant* κ_2 *such that* $\lim_{n\to}\mathbb{P}(\lambda_{\min,s_d+s_0}(X^{\mathrm{T}}X/n) \geq \kappa_2) = 1$*, where* s_d *is the cardinality of* $\widehat{M}_D$*.*

Assumption 15. *The under-fitting bias is small in the sense:* $D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\mathbf{W}) = o_p(\sqrt{n})$*.*

Assumption 12 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 13 restricts the sparsity level of β and the selected model size. Although in this assumption we require β to be ultra-sparse, if the maximum correction between D and the over-selected variable in Assumption 12 is of order $O_p(1/\sqrt{n})$, the ultra-sparsity condition can be relaxed to $s_0\sqrt{\log p} = o(\sqrt{n})$. Assumption 14 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 Assumptions says that, unlike the treatment in Belloni et al. (2014), PODS no longer requires (3.6) to be true or γ to be ultra-sparse. Assumption 15 assumes a negligible under-fitting bias. In a boarder context, Chernozhukov et al. (2018) assumed a similar condition. Sufficient conditions for Assumption 15 are provided in Belloni et al. (2014).

Theorem 2 (Asymptotic normality of PODS)**.** *Under Assumption 1 and Assumption 12-15, 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 $\breve{\sigma}_n^2 = \sigma_\varepsilon^2/(D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D/n)$.

Remark 4 (Variance estimation)**.** *Under additional assumptions that* $||n^{-1/4}(I-\boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})||_2 = o_p(1)$ *and* $(\frac{1}{n}\sum_{i=1}^n R_{ni}^2)^{1/2} = O(\sqrt{s_0/n})$ *and the cardinality of* $\widehat{M}^*$ *is of the same order as* s_0*, we have* $\breve{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{1}{||D-\boldsymbol{X}\widehat{\gamma}^*||_2^2/n} + o_p(1)$*, and it can be consistently estimated by* $\widehat{\sigma}_\varepsilon^2/(||D-\boldsymbol{X}\widehat{\gamma}^*||_2/n)$ *where* $\widehat{\sigma}_\varepsilon^2 = Y^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}^*})Y \cdot n/(n-|\widehat{M}^*|-1)$.

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 in Remark 3 also apply to the post-double-selection estimator.

3.1.3 Data splitting in removing the over-fitting bias of the de-sparsified Lasso

In this subsection, we discuss the bias issue of the de-sparsified Lasso estimator of van de Geer et al. (2014). To simplify the discussion, we may work under the additionally model (3.6). The de-sparsified Lasso estimator can be decomposed

$$\begin{aligned}\sqrt{n}(\widehat{\alpha}_{\text{desparse}} - \alpha) =& \widehat{\nu}^{-2}\frac{1}{\sqrt{n}}\nu^{\mathrm{T}}\varepsilon + \underbrace{\widehat{\nu}^{-2}\frac{1}{\sqrt{n}}\nu^{\mathrm{T}}(\boldsymbol{X}\beta - \boldsymbol{X}\widehat{\beta}_{\text{Lasso}})}_{:=c_{n1}} \\ &+ \underbrace{\widehat{\nu}^{-2}\sqrt{n}(\widehat{\gamma}-\gamma)^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n(\widehat{\beta}_{\text{Lasso}} - \beta)}_{:=c_{n2}}. \qquad (3.10)\end{aligned}$$

The first term is centered since $\widehat{\nu}$ is obtained independent of ε. The term c_{n2} is similar to the under-fitting bias b_{n2}, and is small as long as either γ or β is well estimated. The term c_{n1} measures the correlation between ν and $\widehat{\beta}_{\text{Lasso}}$ and is generally not centered. Since $\widehat{\beta}_{\text{Lasso}}$ is obtained via using the information in D, and thereby is correlated with ξ. We note that c_{n1} can be reduced via data splitting.

Suppose that we split the data into two halves T_1 and T_2, and estimate γ and β on T_1, later the residuals of D and Y are obtained on T_2. Note that $\forall i \in T_2$, ν_i is independent with $\widehat{\beta}_b$, thereby c_{n1} is controlled. To mitigate the efficiency loss induced by data splitting, similar procedure as R-Split might be adopted. The confidence interval of the resulting estimate can be constructed by the non-parametric delta method via the normal approximation. To reduce c_{n1}, we may use data splitting in the de-sparsified Lasso estimator. We summarize the idea with repeated data splitting in Algorithm 4.

Algorithm 4 R-Split with the de-sparsified Lasso

```
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 $n_2 = n - n_1$. Let $v_{bi} = 1_{(i \in T_2)}$, for $i = 1, \cdots, n$.
    Step 2. Obtain $\widehat{\gamma}_b$ and $\widehat{\beta}_b$ on $T_1$.
    Step 3. "Predict" on $T_2$, for $i \in T_2$: $\widehat{D}_i = D_i - X_i^{\mathrm{T}}\widehat{\gamma}_b$, $\widehat{Y}_i = Y_i - Y_i^{\mathrm{T}}\widehat{\beta}_b$
    Step 4. Estimate $\alpha$ through: $\widehat{\alpha}_b = (\sum_{i \in T_2} \widehat{D}_i D_i)^{-1}(\sum_{i \in T_2} \widehat{D}_i \widehat{Y}_i)$.
The final "smoothed" estimate is $\widetilde{\alpha} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\alpha}_b$.
```

We note that the repeated data splitting is not the only strategy to reduce the term c_{n1}. When $p = o(\sqrt{n})$, the similar over-fitting bias problem in the context of the two-step estimator has also been identified in Cattaneo et al. (2017). It has been show in Cattaneo et al. (2017) that the jackknife can be used to removes the over-fitting bias and then delivers consistent point estimates. Except for jackknife, one may also adopt the cross-estimation to remove the over-fitting bias. In the high dimensional regime, unlike refitting on the random model, it remains unclear which is the most efficient strategy to combine data splitting with the de-sparsified Lasso. Therefore, we leave the theoretical and the numerical investigations of the de-sparsified Lasso with data splitting to future work.

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

In this thesis, we have considered two classes of methods for debiased inference. The first is based on one-stage selection, which includes R-Split. The second is built upon two-stage selection procedures, e.g. the de-sparsified Lasso, the post-double-selection, and PODS. Since the various two-stage selection methods have similar asymptotic representations, we use PODS as a representative in this section. The purpose of this section is to compare the statistical efficiencies between one- and two-stage selection methods for making inference on α.

To compare the asymptotic behavior of R-Split with PODS more explicitly, we provide alternative asymptotic variances expressions of R-Split and PODS estimators under additional assumptions. The proofs of the alternative expressions are provided in Appendix 2.5.5 and 3.5.4. We make two additional assumptions to simplify the asymptotic variance expression of R-Split. First, 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,

$$\left\| \mathbb{E}\left\{ \frac{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})\boldsymbol{X}_V/n}{D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M},V})D_V/n} \middle| \boldsymbol{\mathcal{X}} \right\} \right\|_\infty = O_p(\sqrt{\log p}). \tag{3.11}$$

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 constant $K_0 > 0$ such that $\mathbb{P}(\limsup_{n\to\infty} \lambda_{\max,s_d+s}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n) \le K_0) = 1$, $\lambda_{\max,s_d+s}(\Sigma) \le K_0$, and the maximum eigenvalue of $\widehat{\Sigma}$ is bounded by $\log p$ in probability. Under Assumptions 1, 4, and 13, and the two additional assumptions stated above, we show in the Appendix that the asymptotic variance of R-Split estimator satisfies

$$\widetilde{\sigma}_n^2 \leq \sigma_\varepsilon^2 \mathbb{E}\left\{ (\Sigma_{\widehat{M}}^{-1})_{11} \middle| \boldsymbol{\mathcal{X}} \right\} + o_p(1), \tag{3.12}$$

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

As for PODS, under Model (3.6), 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

$$\breve{\sigma}_n^2 = \sigma_\varepsilon^2 (\Sigma^{-1})_{11} + o_p(1) = \sigma_\varepsilon^2 / \sigma_\nu^2 + o_p(1). \tag{3.13}$$

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, (3.13) indicates that the two-stage selection method is not very efficient. From the comparison between (3.12) and (3.13), 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, it is convenient to use the same data generating process adopted in Example 3, except that we set $\sigma_\nu^2 = \text{Var}(\nu)$ as an increasing sequence from 0 to 1. The implementation details of various methods under comparison are provided in Section 6.2. 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 3.2. The variance of the oracle estimator is provided as a benchmark.

The results in Figure 3.2 indicate that R-Split is not as efficient as the oracle estimator,

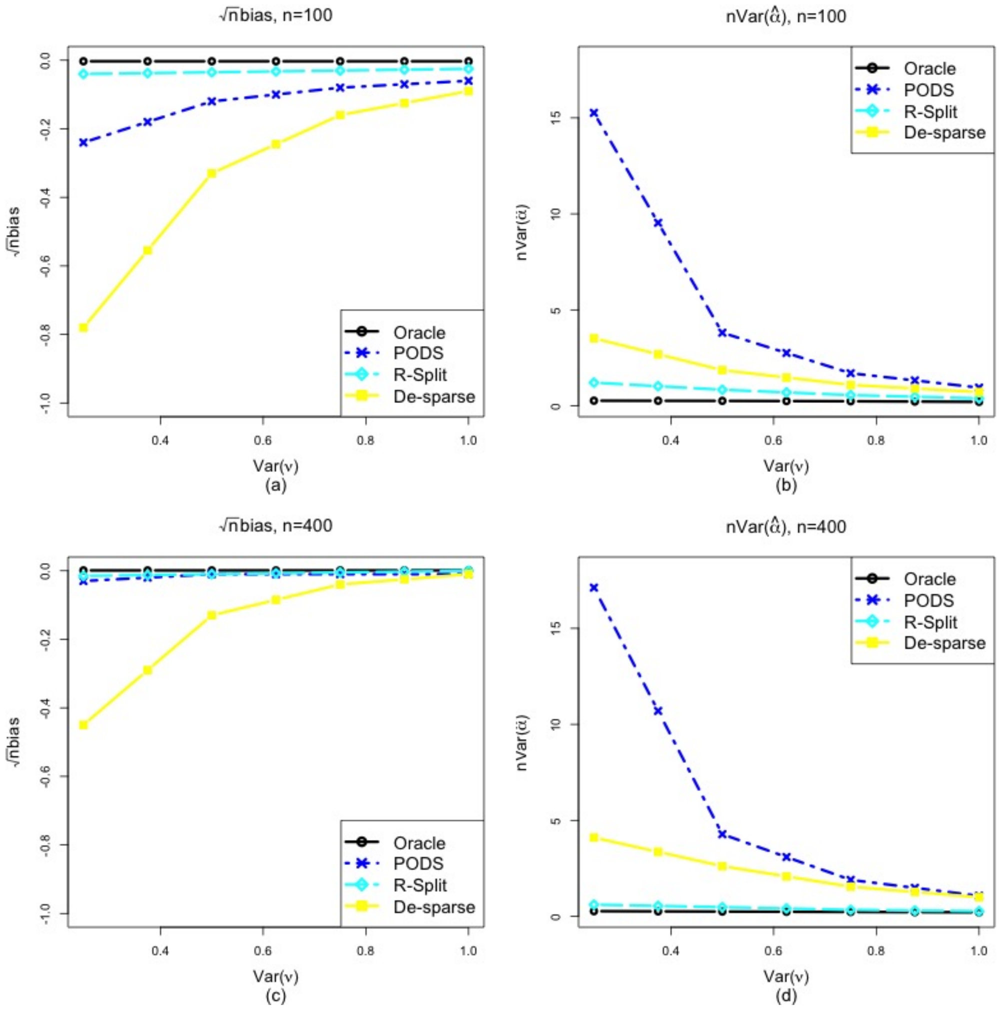


Figure 3.2: Finite sample comparison between R-Split and the two-stage selection methods based on Model (3.8). The data generating process is the same as Example 3, except for σ_ν^2 is a sequence from 0 to 1, and $\Sigma_{jk} = 0.9^{|j-k|}$ is the (j,k)-th element of Σ, for $j,k = 1,\cdots,p+1$. Panels (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.

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 4.

Algorithm 5 R-Split with PODS (or the post-double-selection)

```
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 $n_2 = n - n_1$. Let $v_{bi} = 1_{(i \in T_2)}$, for $i = 1, \cdots, n$.
    Step 2. Select a model $\widehat{M}_b$ by using PODS (or the post-double-selection) based on $T_1$.
    Step 3. Refit the model with the data in $T_2$ to get
        $(\widehat{\alpha}_b, \widehat{\beta}_b^{\mathrm{T}}) = \arg\min \sum_{i \in T_2} (Y_i - \alpha D_i - X_{i,\widehat{M}_b}^{\mathrm{T}} \beta)^2$,
The final "smoothed" estimate is $\widetilde{\alpha} = \frac{1}{B} \sum_{b=1}^{B} \widehat{\alpha}_b$.
```

The estimator derived in Algorithm 4 eliminates the over-fitting bias by data splitting in Step 1, while in each split, the risk of under-fitting is mitigated by PODS (or the post-double-selection). The estimator of the combined approach converges to a normal distribution under the same assumptions for Theorem 1. With the assistance of the two-stage selection in Step 2, Assumption 5 tends to hold more easily. The technical treatment for R-Split and PODS combination is similar to that of Theorem 1 and is omitted in the present thesis. 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.

3.3 Simulation study

This section reports finite sample performances of the proposed methods in comparison with several others through Monte Carlo simulations.

3.3.1 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 3.2.

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 $\widehat{s}_{\text{min}}$ given in Table 3.1-3.2. The implementation details of the adaptive Lasso is provided in Section 3.6 of the Supplementary Materials.

- "PODS-Split" is the combined approach we discussed at the end of Section 3.2. Its implementation is similar to R-Split, except that the minimum and maximum model sizes equal $\widehat{s}_{\text{min}}/2$ 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.

3.3.2 Results

In this subsection, we provide the finite sample comparisons in our simulation studies through Tables 3.1 and 3.2, one for each setting.

Table 3.1 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 3.1.1.

From the results in Table 3.2 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. estimator.

Table 3.1: 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 3.2: Notations are the same as in Table 3.1. 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, $R_y^2 = 0.8$, $R_d^2 = 0.5$, $\widehat{s}_{\text{min}} = 6$							
$\sqrt{n}$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, $R_y^2 = 0.8$, $R_d^2 = 0.9$, $\widehat{s}_{\text{min}} = 6$							
$\sqrt{n}$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, $R_y^2 = 0.8$, $R_d^2 = 0.3$, $\widehat{s}_{\text{min}} = 10$							
$\sqrt{n}$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, $R_y^2 = 0.8$, $R_d^2 = 0.8$, $\widehat{s}_{\text{min}} = 10$							
$\sqrt{n}$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, $R_y^2 = 0.8$, $R_d^2 = 0.8$ $\widehat{s}_{\text{min}} = 15$							
$\sqrt{n}$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 $R_y^2 = 1 - \frac{\mathbb{E}(\varepsilon_i + \alpha\nu_i)^2}{\text{Var}(Y_i)} = 1 - \frac{\sigma_\varepsilon^2 + \alpha^2\sigma_\nu^2}{\text{Var}(Y_i)}$, and $R_d^2 = 1 - \frac{\sigma_\nu^2}{\text{Var}(D_i)}$. The numbers in the parenthesis are the standard errors of the estimated values. The nominal coverage probability is 0.95.

3.4 Real data analysis

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 trails. However, randomized studies are not always feasible due to ethical and practical limitations, and most of the empirical evidence regarding the effect of smoking on birth weight is based on observational studies. A method often used is the regression analysis to adjust for the potential confounders, as done in Nijiati et al. (2008) and Zheng et al. (2016).

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 that are not older than 45 or younger than 18, with less 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 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 a few mother's 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 on average. 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, the post-double-selection and the approximate residual balancing are similar to PODS in this particular study, we include the results only for PODS, R-Split, PODS-Split and Alsso+OLS in Figure 3.3. 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.

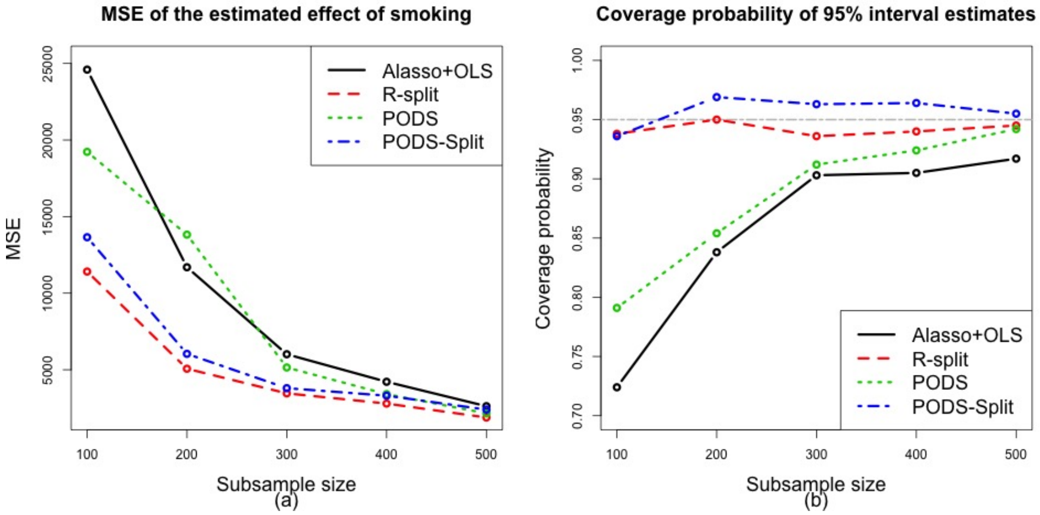


Figure 3.3: Finite sample performance for estimating effect of smoking on infants birth weights, aggregated over 1,000 replications.

In this specific dataset, to show that our comparison in this section is robust to different transformations of the covariates, we provide another set of comparison based on polynomial basis function expansion. Instead of replacing several continuous variables with their spline basis functions, we replace them with their orthogonal polynomial basis functions (up to 5th order). As an implementation detail, the orthogonal polynomial basis functions are generated via R function `poly` and the spline basis functions are generated via R function `ns`.

After Lasso screening, we control for father's age and race, infant's sex, plurality, infant delivery month, infant's Apgar score, the obstetric estimate of gestation, induction of labor, admission to NICU, anesthesia usage during delivery, mother's pre-pregnancy weight, mother's weight gain during pregnancy, mother's height, mother's status of gestational diabetes and a few mother's complications during pregnancy, and some interaction terms between these selected features. In total, we keep $p = 747$ variables. Based on the selected variables, we apply OLS on the year 2016 data. We conclude that the mothers who were the self-reported smokers delivered instant weighting 81.56g less than the other, and 48.25% of the variance of infant birth weight can be explained by the selected 747 variables.This is similar to the conclusion reached via spline basis function expansion. The results of PODS, R-Split, PODS-Split and Alsso+OLS are provided in Figure 3.4. We observe that similar conclusions can be reached compared to the results based on splines basis function expansion.

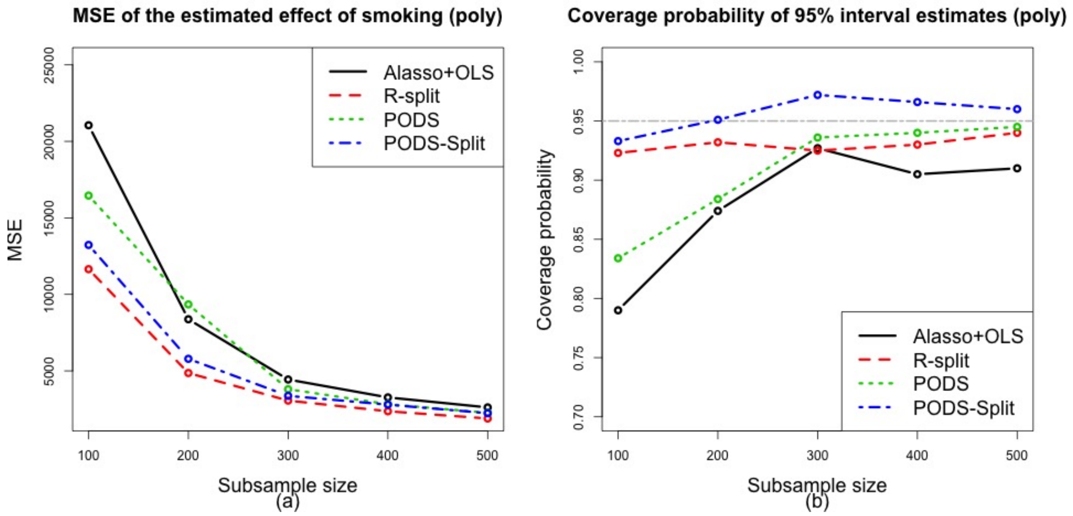


Figure 3.4: Finite sample performance for estimating effect of smoking on infants birth weights, aggregated over 1,000 replications. The continuous variables are replaced by their orthogonal polynomial expansions.

3.5 Proofs

3.5.1 Proof of Theorem 2

We prove Theorem 2 in Section 3.1.2 for PODS. Define $\widetilde{M}_D = \widehat{M}_D \cup (M_0 \cap \widehat{M}_Y^*)$, and $\widetilde{M}_Y = \widehat{M}_Y^* \backslash M_0$, therefore $\widehat{M}^* = \widetilde{M}_D \cup \widetilde{M}_Y$ and $\widetilde{M}_D \cap \widetilde{M}_Y = \emptyset$. The estimated α from PODS satisfies

$$\begin{aligned}
&\sqrt{n}(\widehat{\alpha} - \alpha) \\
=&\left\{\frac{1}{n} D_P^{\mathrm{T}} (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y}) D_P\right\}^{-1} \cdot \left\{\frac{1}{\sqrt{n}} D_P^{\mathrm{T}} (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y})(I - \boldsymbol{P}_{\widetilde{M}_D}) g(\boldsymbol{W}). + \frac{1}{\sqrt{n}} D_P^{\mathrm{T}} (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y}) \varepsilon_P\right\} \\
=&\left(\frac{1}{n} D_P^{\mathrm{T}} D_P - \frac{1}{n} D_P^{\mathrm{T}} \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} D_P\right)^{-1} \\
&\cdot \left(\frac{1}{\sqrt{n}} D_P^{\mathrm{T}} \varepsilon_P - \frac{1}{\sqrt{n}} D_P^{\mathrm{T}} \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} \varepsilon_P + \frac{1}{\sqrt{n}} D_P^{\mathrm{T}} (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y})(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) g(\boldsymbol{W})\right) \\
=&\left(\frac{1}{n} D_P^{\mathrm{T}} D_P - q_{n1}\right)^{-1} \left(\frac{1}{\sqrt{n}} D_P^{\mathrm{T}} \varepsilon_P - q_{n2} + q_{n3}\right),
\end{aligned}$$

where $D_P = (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) D$, $\varepsilon_P = (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\varepsilon$, $\boldsymbol{P}_{\widetilde{M}_D} = \boldsymbol{X}_{\widetilde{M}_D} \left(\boldsymbol{X}_{\widetilde{M}_D}^{\mathrm{T}} \boldsymbol{X}_{\widetilde{M}_D}\right) \boldsymbol{X}_{\widetilde{M}_D}^{\mathrm{T}}$, and

$$\widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} = (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \left(\boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y}\right)^{-1} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}),$$

which is a projection matrix that projects vectors onto space spanned by $(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y}$, and we define

$$\begin{aligned}
q_{n1} &= \frac{1}{n} D_P^{\mathrm{T}} \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} D_P, \\
q_{n2} &= \frac{1}{\sqrt{n}} D_P^{\mathrm{T}} \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} \varepsilon_P, \\
q_{n3} &= \frac{1}{\sqrt{n}} D_P^{\mathrm{T}} (\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y})(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) g(\boldsymbol{W}).
\end{aligned}$$

Our goal is to prove $q_{ni} = o_p(1)$, for $i = 1, 2, 3$, so that

$$\sqrt{n}(\widehat{\alpha} - \alpha) = (D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D/n)^{-1}(D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\varepsilon/\sqrt{n}) + o_p(1).$$

Since ε is uncorrelated with $X_{\widetilde{M}_D}$ and D, therefore given $\boldsymbol{Z}$

$$\widetilde{\sigma}_n^2 = \mathrm{Var}(\sqrt{n}(\widehat{\alpha} - \alpha)|\mathbf{Z}) = \sigma_\varepsilon^2 D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D/n + o_p(1).$$

The formal proofs of these remainder terms are given in Steps 1-4.

Step 1. In this step, we prove

$$\lambda_{\min}^{-1}(\boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\boldsymbol{X}_{\widetilde{M}_Y}/n) \le \lambda_{\min}^{-1}(\boldsymbol{X}_{\widehat{M}^*}^{\mathrm{T}}\boldsymbol{X}_{\widehat{M}^*}/n). \tag{3.14}$$

With (3.14), by Assumption 14, the right hand side is bounded above by $1/\kappa$ with probability going to 1. Therefore, the minimum eigenvalue of matrix $\boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\boldsymbol{X}_{\widetilde{M}_Y}/n$ is bounded in probability.

To prove (3.14), consider any $\mathbf{U} \in \mathbb{R}^{n\times s}$, and $\boldsymbol{A} = \mathbf{U}^{\mathrm{T}}\mathbf{U}$ be a symmetric matrix with minimum eigenvalue $\lambda_{\min}$, we have $\lambda_{\min}(A) = \min_{\nu^{\mathrm{T}}\nu=1} \nu^{\mathrm{T}}\boldsymbol{A}\nu$. Let $\delta = (\delta_1, \cdots, \delta_m)$ be a set of indices of the columns of $\mathbf{U}$ we are interested in. Therefore we observe $\mathbf{U}_\delta^{\mathrm{T}}\mathbf{U}_\delta$ be a sub-matrix of $\mathbf{U}^{\mathrm{T}}\mathbf{U}$, and there exists an index matrix $\boldsymbol{I}_\delta \in \mathbb{R}^{p\times m}$ such that $\mathbf{U}\boldsymbol{I}_\delta = \mathbf{U}_\delta$. By the definition of minimum eigenvalue, we have

$$\lambda_{\min}(\mathbf{U}_\delta^{\mathrm{T}}\mathbf{U}_\delta) = \min_{y\in\mathbb{R}^m|y^{\mathrm{T}}y=1} y^{\mathrm{T}}\mathbf{U}_\delta^{\mathrm{T}}\mathbf{U}_\delta y = \min = \min_{y\in\mathbb{R}^m|y^{\mathrm{T}}y=1} y^{\mathrm{T}}\boldsymbol{I}_\delta^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\boldsymbol{I}_\delta y \ge \min_{\nu^{\mathrm{T}}\nu} \nu^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\nu = \lambda_{\min}($$

Similarly we can prove $\lambda_{\max}(\mathbf{U}_\gamma^{\mathrm{T}}\mathbf{U}_\gamma) \le \lambda_{\max}(\mathbf{U}^{\mathrm{T}}\mathbf{U})$. Therefore (3.14) holds, since matrix $(\boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\boldsymbol{X}_{\widetilde{M}_Y}/n)^{-1}$ is a sub-matrix of $(\boldsymbol{X}_{\widehat{M}^*}^{\mathrm{T}}\boldsymbol{X}_{\widehat{M}^*}/n)^{-1}$.

Step 2. Consider q_{n1}. By Assumption 14 and the result in Step 1, we have

$$
\begin{aligned}
q_{n1} &\leq \left| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) D \right| \\
&= \left| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right)^{-1} \frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) D \right| \\
&\leq \lambda_{\min}^{-1} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right) \left\| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right\|_2^2 \\
&\lesssim_P s_0 \log p / n = o_p(1),
\end{aligned}
$$

where the last equality follows from Assumption 12.

Step 3. Consider q_{n2}. Again by Assumptions 12, 14 and the result in Step 1, we have

$$
\begin{aligned}
q_{n2} &= \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right)^{-1} \frac{1}{\sqrt{n}} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \varepsilon \\
&\leq \left\| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) X_{\widetilde{M}_Y} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right)^{-1} \right\|_1 \cdot \left\| \frac{1}{\sqrt{n}} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \varepsilon \right\|_\infty \\
&\lesssim_P \sqrt{s_0} \left\| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right)^{-1} \right\|_2 \cdot \sqrt{\log p} \\
&\lesssim_P \sqrt{s_0 \log p} \cdot \lambda_{\min}^{-1} \left(\frac{1}{n} \boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right) \left\| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right\|_2 \\
&\lesssim_P s_0 \sqrt{\log p} \cdot \left\| \frac{1}{n} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D}) \boldsymbol{X}_{\widetilde{M}_Y} \right\|_\infty = o_p(1).
\end{aligned}
$$

Step 4. Consider q_{n3}. By the property of the projection matrix, we have $P_{\widehat{M}^*} = P_{\widetilde{M}_D} + \widetilde{P}_{\widetilde{M}_Y}$, and then $\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*} = (\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})(\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y})(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})$, since $\widehat{M} = \widetilde{M}_D \cup \widetilde{M}_Y$ and $\widetilde{M}_D \cap \widetilde{M}_Y = \emptyset$. Under Assumption 15, $q_{n3} = \frac{1}{\sqrt{n}} D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*}) g(\mathbf{W}) = o_p(1)$.

3.5.2 Proof of Remark 3 in Section 3.1.2

In this subsection, we provide the proof of Remark 3 in Section 3.1.2 that the variance $\widetilde{\sigma}_n^2$ can be consistently estimated by $\breve{\sigma}_n^2 = \widehat{\sigma}_\varepsilon^2 / (D^{\mathrm{T}} (\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*}) D / n)$, where $\widehat{\sigma}_\varepsilon^2 = Y^{\mathrm{T}} (\boldsymbol{I} -$

$\mathbf{P}_{\widehat{M}^*})Y \cdot n/(n - |\widehat{M}^*| - 1)$. We start with proving three ancillary bounds. Let $\beta_{\widehat{M}^*} = \widetilde{I}^{\mathrm{T}}_{\widehat{M}^*}\boldsymbol{X}_{\widehat{M}^*}(\boldsymbol{X}^{\mathrm{T}}_{\widehat{M}^*}\boldsymbol{X}_{\widehat{M}^*})^{-1}\boldsymbol{X}^{\mathrm{T}}_{\widehat{M}^*}g(\boldsymbol{W}) \in \mathbb{R}^{p\times 1}$.

- The first bound. Under Assumptions 13, 14 and 15,

$$\begin{aligned}
||\beta - \beta_{\widehat{M}^*}||_1 &\lesssim_P \sqrt{s_0}||\beta - \beta_{\widehat{M}^*}||_2 \\
&\lesssim_P \sqrt{s_0}\lambda_{\min,s}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n)||\beta - \beta_{\widehat{M}^*}||_2 \\
&\leq \sqrt{s_0}||(\boldsymbol{X}_{M_0}\beta_{M_0} - \boldsymbol{X}\beta_{\widehat{M}^*})/\sqrt{n}||_2 \\
&\leq \sqrt{s_0}(||R/\sqrt{n}||_2 + ||(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(W)/\sqrt{n}||_2) \\
&= o(\sqrt{s_0^2/n}) + o_p(s_0^{1/2}n^{-1/4}).
\end{aligned}$$

- The second bound. Followed by the first bound above and Lemma 1,

$$\begin{aligned}
\varepsilon^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/n &= \varepsilon^{\mathrm{T}}\boldsymbol{X}(\beta - \beta_{\widehat{M}^*})/n + \varepsilon^{\mathrm{T}}R/n \\
&\lesssim_P ||\varepsilon^{\mathrm{T}}\boldsymbol{X}/n||_\infty||\beta - \beta_{\widehat{M}^*}||_1 + \sqrt{s_0}/n = o_p(1).
\end{aligned}$$

- The third bound. Under Assumptions 13 and 14,

$$\varepsilon^{\mathrm{T}}\boldsymbol{P}_{\widehat{M}^*}\varepsilon/n = \varepsilon^{\mathrm{T}}\boldsymbol{Z}_{\widehat{M}^*}/n(\boldsymbol{Z}^{\mathrm{T}}_{\widehat{M}^*}\boldsymbol{Z}_{\widehat{M}^*}/n)^{-1}\boldsymbol{Z}^{\mathrm{T}}_{\widehat{M}^*}\varepsilon/n \leq \frac{1}{\kappa_2}||\boldsymbol{Z}^{\mathrm{T}}_{\widehat{M}^*}\varepsilon/n||_2^2 = \frac{1}{\kappa_2}s_0\log p/n = o_p(1). \tag{3.15}$$

As shown in Step 1, $D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})D/n = D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})D/n + o_p(1)$. Since $s = o(n)$, it suffices to prove $Y^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}^*_{\widehat{M}^*})Y/n = \sigma^2_\varepsilon + o_p(1)$. Assisted by these three auxiliary bounded, recall for a model M, $\boldsymbol{P}^*_M = \boldsymbol{Z}_M(\boldsymbol{Z}^{\mathrm{T}}_M\boldsymbol{Z}_M)^{-1}\boldsymbol{Z}^{\mathrm{T}}_M$ is a projection matrix that includes D, we have

$$\begin{aligned}
Y^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}^*_{\widehat{M}^*})Y/n &= (g(\boldsymbol{W})+\varepsilon)^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}^*_{\widehat{M}^*})(g(\boldsymbol{W})+\varepsilon)/n \\
&= \big|\big|(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/\sqrt{n}\big|\big|_2 + 2\varepsilon^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/n + \varepsilon^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}^*})\varepsilon/n \\
&= o_p(n^{-1/4}) + 2\varepsilon^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/n + \sigma_\varepsilon^2 + o_p(1) + \varepsilon^{\mathrm{T}}\boldsymbol{P}_{\widehat{M}^*}\varepsilon/n \\
&= \sigma_\varepsilon^2 + o_p(1).
\end{aligned}$$

As a minor generalization of this results, as shown in Step 1 $D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}^*})D/n = D^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widetilde{M}_D})D/n + o_p(1)$, then we have $\breve{\sigma}_n^2 = \sigma_\varepsilon^2 \frac{1}{||D-\boldsymbol{X}\widehat{\gamma}^*||_2^2/n} + o_p(1)$, which echoes the result in (3.3).

3.5.3 Illustration of (3.9) in Section 3.1.2

In this part, we provide the arguments used in Example 3 in Section 3.1.2. Under the model 3.8, given $\{1\} \in \widehat{M}_D$, $(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})D = (\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})(X_1\gamma_1+\nu) = (\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\nu$ by the property of the projection matrix, and $k_{n2} = \nu^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\nu/n = \sigma_\nu^2 + o_p(1)$ if $s_d \log p = o(n)$ (the proof is similar to (3.15)). Therefore $\widehat{M}_Y^*$ is selected via

$$\widehat{M}_Y^* = \left\{ \arg\max_{1\le j\le p} \left| \frac{1}{c_{nj}^*} \left(\frac{1}{n} X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon - \frac{1}{n} X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\nu \cdot \frac{1}{k_{n2}} \frac{1}{n} \nu^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon \right) \right| \right\},$$

Since $\frac{1}{n}X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon = O_p(1/\sqrt{n})$ and $\frac{1}{n}X_j^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\nu = O_p(\sqrt{\log p/n})$, k_{n2} is lowered bounded by a constant, and $\frac{1}{n}\nu^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{\widehat{M}_D})\varepsilon = O_p(1/\sqrt{n})$, therefore the second term is a smaller order term as long as $\log p = o(n)$.

3.5.4 Derivation of (3.13) in Section 3.2

In this part, we provide the alternative variance expression of PODS used in Section 3.2. Given the two-stage model $D = \boldsymbol{X}\gamma + \nu$ which satisfies $\mathbb{E}(\nu|\boldsymbol{X}) = 0$, we then have

$\gamma = \Sigma_{DX}\Sigma_X^{-1}$, and

$$\begin{aligned}\sigma_\nu^2 &= \text{Var}(\nu_i) = \mathbb{E}\left\{\text{Var}(D_i|X_i)\right\} \\ &= \mathbb{E}\left\{\mathbb{E}(D_i - \mathbb{E}(D_i|X_i))^2|X_i)\right\} \\ &= \mathbb{E}\left\{\mathbb{E}(D_i^2 - 2D_iX_i^{\mathrm{T}}\gamma + \gamma^{\mathrm{T}}X_iX_i^{\mathrm{T}}\gamma|X_i)\right\} \\ &= \mathbb{E}(D_i^2) - \gamma^{\mathrm{T}}\Sigma\gamma = \Sigma_{11} - \gamma^{\mathrm{T}}\Sigma\gamma = \Sigma_{11} - \Sigma_{DX}\Sigma_X^{-1}\Sigma_{XD} = (\Sigma^{-1})_{11}.\end{aligned}$$

Therefore to prove

$$\breve{\sigma}_n^2 = \sigma_\varepsilon^2(\Sigma^{-1})_{11} + o_p(1) = \sigma_\varepsilon^2/\sigma_\nu^2 + o_p(1),$$

it remains to show that $||D - \boldsymbol{X}\widehat{\gamma}^*||_2^2/n = \sigma_\nu^2 + o_p(1)$. Except for the approximate error equals zero, the proof is the same as C.2 and thus is omitted.

3.5.5 Sufficient conditions to control the under-fitting bias

In this section, we discuss sufficient conditions to ensure the validity of Assumptions 5 and 15. Since the magnitude of under-fitting bias depends on the approximation errors R_n, the conditions discussed here require the approximation errors to be small.

3.5.5.1 Sufficient conditions for Assumption 5

Under Condition 1-3 below, the under-fitting bias of R-Split vanishes with a large n and Assumption 5 holds. We follow the notations used in Section 2.5.2, and work under Assumptions 1-4.

Condition 1. *The selected models satisfy*

$$\mathbb{E}\big((D_V^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{V,\widehat{M}})\boldsymbol{X}_V\beta/\sqrt{n})^2|\boldsymbol{\mathcal{X}}\big) = o_p(1). \tag{3.16}$$

Condition 2. *The approximation errors satisfy* $\mathbb{E}\left(\|\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}\|_2|\boldsymbol{\mathcal{X}}\right)=o_p(1)$.

Condition 3. $\mathbb{E}\left(e_1^{\mathrm{T}}\widehat{\Sigma}_{V.\widehat{M}}^{-2}e_1|\boldsymbol{\mathcal{X}}\right)\leq K_1$, *where* K_1 *is a positive constant.*

The expectations here are taken with respect to the distribution of V given the sample. These condition indicates that the under-fitting bias needs to be small on average, but it allows some $\widehat{M}$ to miss active variables in M_0. Condition 2 here contains an explicit requirement on the approximation errors.

We apply the transformation provided in (2.15) and then the under-fitting bias reduces to

$$
\begin{aligned}
&\mathbb{E}\left((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V/n)^{-1}\cdot D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})g_V(\boldsymbol{W})/\sqrt{n}|\boldsymbol{\mathcal{X}}\right)\\
=&\mathbb{E}\left((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V/n)^{-1}\cdot D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})\boldsymbol{X}_V\beta/\sqrt{n}|\boldsymbol{\mathcal{X}}\right)\\
&+\mathbb{E}\left(e_1^{\mathrm{T}}(\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{V,\widehat{M}}/n)^{-1}\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}|\boldsymbol{\mathcal{X}}\right)\\
\leq&\left(\mathbb{E}\big((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V/n)^{-2}|\boldsymbol{\mathcal{X}}\right)^{1/2}\cdot\left(\mathbb{E}\big((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})\boldsymbol{X}_V\beta/\sqrt{n})^2|\boldsymbol{\mathcal{X}}\big)\big)\right)^{1/2}\\
&+\mathbb{E}\left((e_1^{\mathrm{T}}(\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{V,\widehat{M}}/n)^{-1}\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}|\boldsymbol{\mathcal{X}}\right),\\
=&K_1\left(\mathbb{E}\big((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})\boldsymbol{X}_V\beta/\sqrt{n})^2|\boldsymbol{\mathcal{X}}\big)\right)^{1/2}+\mathbb{E}\left(e_1^{\mathrm{T}}(\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{V,\widehat{M}}/n)^{-1}\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}|\boldsymbol{\mathcal{X}}\right),\\
=&\mathbb{E}\left(e_1^{\mathrm{T}}(\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{V,\widehat{M}}/n)^{-1}\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}|\boldsymbol{\mathcal{X}}\right)+o_p(1),
\end{aligned}
$$

where the last step is obtained via (3.16). The quantity inside the conditional expectation of the above line is upper bounded by

$$
e_1^{\mathrm{T}}(\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}\boldsymbol{Z}_{V,\widehat{M}}/n)^{-1}\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}\leq\|\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1\|_2\cdot\|\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}\|_2. \tag{3.17}
$$

Thus under Assumption 9, the under-fitting bias of R-Split is bounded by

$$
\begin{aligned}
&\mathbb{E}\left((D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})D_V/n)^{-1}\cdot D_V^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{P}_{V,\widehat{M}})g_V(\boldsymbol{W})/\sqrt{n}|\boldsymbol{\mathcal{X}}\right)\\
\leq&\,\mathbb{E}\left(\|\widehat{\Sigma}_{V,\widehat{M}}^{-1}e_1\|_2\cdot\|\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}\|_2|\boldsymbol{\mathcal{X}}\right)\leq K\,\mathbb{E}\left(\|\boldsymbol{Z}_{V,\widehat{M}}^{\mathrm{T}}R_{n,V}/\sqrt{n}\|_2|\boldsymbol{\mathcal{X}}\right)=o_p(1).
\end{aligned}
$$

3.5.5.2 Sufficient conditions for Assumption 15

Sufficient conditions for controlling the under-fitting bias without smoothing can be developed similarly. Note that we follow the notations used in Section 3.5.1, and we work under Assumption 1 and Assumptions 12-14.

Condition 4. *Suppose that the sure screening property holds for $\widehat{M}^*$, i.e. $M_0 \subseteq \widehat{M}^*$, and we assume that $D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})R_n = o_p(1/\sqrt{n})$.*

Given that $M_0 \subseteq \widehat{M}^*$ and $q_{n1} = o_p(1)$, the bias term in Assumption 15 satisfies

$$\begin{aligned} D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\mathbf{W})/\sqrt{n} &= D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})\boldsymbol{X}\beta/\sqrt{n} + D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})R_n/\sqrt{n} \\ &= D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})R_n/\sqrt{n} = o_p(1). \end{aligned}$$

Condition 5. *We assume that the selected model $\widehat{M}^*$ satisfies $\min_{\beta:\beta_j=0,j\notin\widehat{M}^*} \|g(\boldsymbol{W}) - \boldsymbol{X}\beta\|_2 = o_p(n^{1/4})$.*

Notice that under the given condition we have $\|n^{-1/4}(I - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})\|_2 = o_p(1)$. Following the proof in Section 3.5.1, under Assumption 12 and 14 , we have

$$\begin{aligned} \frac{1}{\sqrt{n}}D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W}) &= \frac{1}{\sqrt{n}}D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})^2(\boldsymbol{I} - \widetilde{\boldsymbol{P}}_{\widetilde{M}_Y})(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})g(\boldsymbol{W}) \\ &= \frac{1}{\sqrt{n}}D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W}) \\ &\le \big|\big|D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\big|\big|_2 \cdot \big|\big|\frac{1}{\sqrt{n}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})\big|\big|_2 \\ &\lesssim_p \big|\big|D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\big|\big|_2 \cdot \lambda_{\min}(\boldsymbol{X}_{\widetilde{M}_Y}^{\mathrm{T}}\boldsymbol{X}_{\widetilde{M}_Y}/n) \cdot \big|\big|\frac{1}{\sqrt{n}}(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})\big|\big|_2 \\ &\le \big|\big|D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\boldsymbol{X}_{\widetilde{M}_Y}/\sqrt{n}\big|\big|_2 \cdot \big|\big|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/\sqrt{n}\big|\big|_2 \\ &\le \sqrt{s}\big|\big|D^{\mathrm{T}}(\boldsymbol{I} - \boldsymbol{P}_{\widetilde{M}_D})\boldsymbol{X}_{\widetilde{M}_Y}/\sqrt{n}\big|\big|_\infty \cdot \big|\big|(\boldsymbol{I} - \boldsymbol{P}_{\widehat{M}^*})g(\boldsymbol{W})/\sqrt{n}\big|\big|_2 \\ &= O_p(\sqrt{s\log p}) \cdot o_p(n^{-1/4}) = o_p(1). \end{aligned}$$

3.6 Implementation details

In the simulation studies and the examples provided in Section 2.2 and 3.3, we used the adaptive Lasso Zou (2006) for model selection. In this part, we provide the implementation details for the adaptive Lasso.

The adaptive Lasso weights are chosen via the high dimensional ordinary least squares projection for screening variables (HOLP) in Wang and Leng (2016). Following Wang and Leng (2016), HOLP is used to select a screening set $\widehat{\mathcal{H}}$ of size d,

$$\begin{aligned} &\widetilde{\theta}_{\mathrm{H}} = \boldsymbol{Z}(\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}})^{-1}\boldsymbol{Z}Y, \qquad (3.18)\\ &\widehat{\mathcal{H}} = \{j : |\widetilde{\theta}_{\mathrm{H}}|_j \text{ are among the largest } d \text{ of all } |\widetilde{\theta}'_{\mathrm{H}}|_j\text{'s}\}. \end{aligned}$$

The authors suggest selecting at least n variables to preserve the true model with an overwhelming probability. In the cases when $p \approx n$ or $p \leq n$, $\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}$ becomes singular but screening is not necessary and can be omitted. In our simulation studies, we choose $d = 300$. Then let $\widehat{\mathcal{H}}_k$ be the kth component of set $\widehat{\mathcal{H}}$, and we denote the covariates after screening as $X_{i,\widehat{\mathcal{H}}}$,and the observations after screening as $\{Y_i, D_i, X_{i,\widehat{\mathcal{H}}}\}_{i=1}^n$, for $i = 1, \cdots, n$. By allowing a relatively higher penalty for the zero coefficients and a lower penalty for nonzero coefficients, we take $\widetilde{\theta}_{\mathcal{H}}$ from the screening step to weight the L_1 penalties. The adaptive Lasso is then implemented as follows

$$\begin{aligned} &(\widehat{\alpha}_L, \widehat{\beta}'_L)' = \underset{\alpha\in R,\beta\in R^{(d-1)\times 1}}{\arg\min}\ n^{-1}\sum_{i=1}^{n}(Y_i - \alpha D_i - X'_{i,\widehat{\mathcal{H}}}\beta)^2 + \lambda\sum_{k\in\widehat{\mathcal{H}}}\frac{|\beta_k|}{w_k}, \qquad (3.19)\\ &\widehat{M} = \text{supp}(\widehat{\beta}_L) = \left\{k:\ \beta_k \neq 0, k \in \widehat{\mathcal{H}}\right\} \end{aligned}$$

where w_k is the $\widehat{\mathcal{H}}_k$th component of $|\widetilde{\theta}_{\mathrm{H}}|$ defined in (3.18), β_k is the coefficient for $X_{i,\widehat{\mathcal{H}_k}}$ and $\widehat{M}$ is the selected model with cardinality $|\widehat{M}| = s$.

Additionally, from package `glmnet`, the selected model size s can be controlled by tuning parameter λ. The Lasso coefficient path is built over a grid of tuning parameter

$\lambda_1 > \cdots > \lambda_Q$ and each λ_q indexes a different set of sparse solutions $\widehat{\beta}_{\lambda_q}$. We use $\widehat{M}_s$ as the model on the path whose model size is the closest to s. To be more specific, let the support set $\widehat{M}_{\lambda_q} = \text{supp}(\widehat{\beta}_{\lambda_q})$, clearly, the support sets satisfy

$$\widehat{M}_{\lambda_1} \subset \widehat{M}_{\lambda_2} \subset \cdots \subset \widehat{M}_{\lambda_Q}, \text{ and } |\widehat{M}_{\lambda_1}| < |\widehat{M}_{\lambda_2}| < \cdots < |\widehat{M}_{\lambda_Q}|,$$

where $|\widehat{M}_{\lambda_q}|$ refers the cardinality of each set. To control the working model size s, we simply pick the model $\widehat{M}_s$ through:

$$|\widehat{M}_s| = \underset{\widehat{M}_{\lambda_q},\ q\in\{1,\cdots,Q\}}{\arg\min} \left| s - |\widehat{M}_{\lambda_q}| \right|.$$

CHAPTER 4

Average Treatment Effects Estimation

In this chapter, we consider a special case where $D \in \{0, 1\}$ is a binary random variable and study the average treatment effect (ATE) estimation following the Neyman-Rubin causal model; see Neyman (1923) and Rubin (1974). In the first part of the chapter, we discuss the ATE estimation with regression adjustment and show that the procedures and theoretical properties discussed in Chapter 2 and 3 also apply to this context. As the validity of the procedures discussed in the previous chapter generally depends on a correctly specified regression model (1.2), in the second part of the chapter, we provide an extension of R-Split that combines the doubly robust estimator studied in Robins and Rotnitzky (1995) and Hahn (1998). This extension is proven to provide a consistent estimate of ATE if either the propensity score model or the model (1.2) is correctly specified. In the last part of the chapter, we provide a potentially interesting extension of repeated data splitting approach for estimating heterogeneous treatment effect (HTE).

4.1 ATE estimation with regression adjustment

4.1.1 Notation and setup

In a special case where $D_i \in \{0, 1\}$ represents the treatment indicator, suppose that there are two potential outcomes for each unit, denoted by $Y_i(0)$ for the outcome under the control and $Y_i(1)$ for the outcome under an active treatment, and then Y_i stands for the realized (or observed) outcome. We work under the framework that the potential confounders W_i are not affected by the treatment, and should be observed before any treatment. The parameter of interest is the average treatment effect (ATE) which is defined through $\alpha = \mathbb{E}(Y_i(1) - Y_i(0))$.

When the treatment assignment is not completely random, we may build a model for the response in the treated and the control group separately, and this yields the model:

$$\begin{aligned} Y_i &= D_i(\mu_{D_i}(W_i) + \varepsilon_i^{D_i}) + (1 - D_i)(\mu_{1-D_i}(W_i) + \varepsilon_i^{1-D_i}), \qquad (4.1) \\ \mu_{D_i}(W_i) &= X_i^{\mathrm{T}}\beta^{D_i} + R_{ni}^{D_i}, \quad i = 1, \ldots, n, \end{aligned}$$

for $D_i \in \{0, 1\}$, where $\mu_{D_i}(W_i)$ and $\mu_{1-D_i}(W_i)$ are unknown functions that can be approximated by a linear combination of the covariates X_i. Following the similar setup in Chapter 1, we assume that $\beta^{D_i} \in \mathbb{R}^p$ is an unknown sparse parameter vector, $R_{ni}^{D_i}$ is the approximation error, and $\varepsilon_i^{D_i}$ is the random error that satisfies the assumption $E(\varepsilon_i^{D_i}|W_i) = 0$. In this context, if the approximation error is negligible, the average treatment effect equals $\alpha = (\mathbb{E}X_1)^{\mathrm{T}}(\beta^1 - \beta^0)$, where $\mathbb{E}X_1$ denotes the mean of the covariate vector. Then we note that Model (4.1) is equivalent to the following linear model with interactions

$$Y_i = \alpha D_i + X_i^{\mathrm{T}}\beta^0 + D_i(X_i - \mathbb{E}X_1)^{\mathrm{T}}\delta + R_{ni} + \varepsilon_i, \quad i = 1, \ldots, n, \qquad (4.2)$$

where $\delta = \beta^1 - \beta^0$, $R_{ni} = D_i R_{ni}^1 + (1 - D_i)R_{ni}^0$ and $\varepsilon_i = D_i\varepsilon_i^1 + (1 - D_i)\varepsilon_i^0$. Therefore to estimate α, we can consider the regression model

$$Y_i = \alpha D_i + \widetilde{X}_i^{\mathrm{T}}\beta + R_{ni} + \varepsilon_i, \quad i = 1, \ldots, n, \qquad (4.3)$$

where $\widetilde{X}_i = (X_i^{\mathrm{T}}, D_i(X_i - \bar{X})^{\mathrm{T}})^{\mathrm{T}}$ and $\beta = (\beta^{0\ \mathrm{T}}, \delta^{\mathrm{T}})^{\mathrm{T}}$. Models (4.2) and (4.3) are asymptotically equivalent as long as $s_0 = o(\sqrt{n})$, where s_0 is the size of the support set of β. Except that the dimension of the covariates $\widetilde{X}_i$ has doubled from that of X_i, what we discuss in the previous section still applies to the ATE estimation in this context. We note that the procedures and the theoretical properties to be discussed in the paper for model (1.1) apply to the estimation of the ATE through Model (4.3). To simplify presentation, we shall focus on Model (1.1) in the rest of the thesis except in the simulation study.

In the literature of average treatment effect estimation, it is known that regression adjustment in randomized trials provides consistent estimates when p is fixed, and is as least as efficient as the naive estimate which is the mean difference between the responses in the treated group and the control group -even when the linear model is misspecified; see Lin (2013) and Imbens and Rubin (2015). However, in the presence of high dimensional covariates, such a statement needs to be examined with caution. It has been shown in Bloniarz et al. (2016), given an ultra-sparsity assumption in the sense that $s_0 \log p = o(\sqrt{n})$, regression adjustments using Lasso are more efficient than the naive estimate. Once the ultra-sparsity assumption is violated, it is difficult to consistently estimate the average treatment effect from Lasso without further restrictions on Σ. Wager et al. (2016) proposed an estimate which is built on cross-estimation, and then the ultra-sparsity assumption can be relaxed to $s_0 = o(n)$.

When the goal is to estimate the average treatment effect in non-randomized trials, regression adjustment may not work well when the regression model is misspecified. Another type of estimators has been constructed by estimating the propensity scores, and doubly robust estimators may be used to remove the bias; see Farrell (2015), Robins et al. (2017) and Chernozhukov et al. (2018). To achieve the $\sqrt{n}$-rate of convergence, these procedures would need either a consistent estimator of the propensity score or a correctly specified linear model for the potential outcome. More recently Athey et al. (2018) proposed the approximate residual balancing method without estimating the propensity score, but they required the potential outcome to follow a ultra-sparse linear model to achieve asymptotic normality of the estimator.

Finally, we note that the procedures and the theoretical properties discussed in the previous chapters for model (1.1) apply to the estimation of ATE through Model (4.3).

4.1.2 Simulation study

When D is binary, we consider a case similar to the many-cluster model used in Section 5.2 of Athey et al. (2018). Consider for $i = 1, \cdots, n$,

$$Y_i = (C_i + X_i)^{\mathrm{T}} \beta + D_i + \varepsilon_i, \tag{4.4}$$

where $X_i \sim N(0, \boldsymbol{I}_p)$, ε_i are white noise, $\beta = c_y \beta_0$ with the constants c_y chosen for desired values of R^2 given in Table 4.1, and $C_i \in \mathbb{R}^p$ are the cluster centers defined as follows. We first choose 10 cluster centers $\{c_1, \cdots, c_{10}\}$ as a random sample from $N(0, \boldsymbol{I}_p)$, and then draw C_i uniformly at random from the 10 cluster centers. The variable D_i is drawn from the Bernoulli distribution with probability η for the first 5 clusters and with probability $1 - \eta$ for the remaining clusters. For comparison, we include the methods discussed in Section 3.3 and "BalanceHD" , which is the residual balancing method of Athey et al. (2018) for estimating average treatment effect when D is binary only, and is implemented by R package `balanceHD`.

Encouragingly, similar conclusions can be reached from the results in Table 4.1 for Setting 3 in comparison to Settings 1 and 2 in Chapter 3.3. Meanwhile, we observe that the residual balancing method can successfully reduce the bias in most of the cases. The balancing weights used in this method involve a bias-variance trade-off. When n is small but p is large, the balancing weights can successfully reduce the bias but at a cost of higher variance. For the sample size up to $n = 200$, the variance of the estimator can be much higher than R-Split and PODS for sparse models.

Table 4.1: The results are based on the many-cluster setting in Model (4.4) with $p = 500$.

	Oracle	Double	Double-2CV	PODS	R-Split	PODS-Split	De-sparsified	Alasso+OLS	BalanceHD
β is sparse, $R^2 = 0.5$, $\widehat{s}_{\min} = 6$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	−0.08(0.09)	−0.81(0.12)	−0.37(0.19)	−0.11(0.12)	−0.14(0.11)	−0.11(0.12)	−0.69(0.11)	−2.29(0.11)	−0.07(0.12)
nMSE	3.91(0.23)	8.08(0.52)	18.22(2.42)	7.52(0.51)	5.96(0.37)	6.78(0.40)	6.33(0.38)	10.98(0.72)	7.58(0.46)
Cover	0.97(0.01)	0.94(0.01)	0.90(0.01)	0.92(0.01)	0.94(0.01)	0.95(0.01)	0.92(0.01)	0.61(0.02)	0.94(0.01)
Length	0.40(0.00)	0.52(0.00)	0.60(0.01)	0.46(0.00)	0.47(0.00)	0.50(0.00)	0.48(0.00)	0.31(0.00)	0.53(0.00)
β is sparse, $R^2 = 0.8$, $\widehat{s}_{\min} = 6$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	0.07(0.09)	−0.64(0.11)	−3.80(3.86)	0.20(0.11)	0.10(0.09)	0.09(0.10)	−0.86(0.11)	−1.21(0.10)	0.10(0.15)
nMSE	3.70(0.24)	6.27(0.52)	14.85(0.88)	6.09(0.70)	4.27(0.27)	4.81(0.33)	7.20(0.47)	6.79(0.47)	12.01(0.80)
Cover	0.96(0.01)	0.93(0.01)	0.91(0.01)	0.91(0.01)	0.96(0.01)	0.96(0.01)	0.92(0.01)	0.78(0.02)	0.93(0.01)
Length	0.41(0.00)	0.45(0.00)	0.68(0.01)	0.40(0.00)	0.44(0.00)	0.47(0.00)	0.47(0.00)	0.34(0.00)	0.63(0.00)
β is moderately sparse, $R^2 = 0.5$, $\widehat{s}_{\min} = 10$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	−0.06(0.10)	−0.46(0.14)	−0.19(0.18)	−0.08(0.13)	−0.22(0.13)	−0.13(0.13)	−0.46(0.12)	−2.41(0.12)	−0.05(0.13)
nMSE	5.43(0.36)	9.34(0.73)	15.72(1.53)	8.88(0.66)	7.87(0.62)	8.57(0.67)	7.31(0.56)	12.80(0.97)	8.29(0.61)
Cover	0.93(0.01)	0.92(0.01)	0.90(0.01)	0.92(0.01)	0.93(0.01)	0.94(0.01)	0.93(0.01)	0.65(0.02)	0.92(0.01)
Length	0.46(0.00)	0.56(0.01)	0.61(0.01)	0.52(0.00)	0.53(0.00)	0.55(0.00)	0.52(0.00)	0.34(0.00)	0.55(0.00)
β is moderately sparse, $R^2 = 0.8$, $\widehat{s}_{\min} = 10$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	−0.04(0.10)	−2.09(0.19)	−0.37(0.32)	−0.28(0.19)	−0.35(0.15)	−0.37(0.18)	−2.45(0.15)	−1.71(0.13)	−0.35(0.21)
nMSE	5.35(0.32)	21.62(1.45)	50.86(5.97)	18.98(1.44)	11.76(0.79)	16.59(1.09)	17.75(1.18)	11.83(0.85)	21.61(1.46)
Cover	0.95(0.01)	0.87(0.01)	0.88(0.01)	0.83(0.02)	0.93(0.01)	0.92(0.01)	0.84(0.02)	0.77(0.02)	0.93(0.01)
Length	0.46(0.00)	0.73(0.01)	0.97(0.01)	0.58(0.01)	0.64(0.00)	0.75(0.00)	0.59(0.00)	0.41(0.00)	0.83(0.00)
β is dense, $R^2 = 0.5$, $\widehat{s}_{\min} = 10$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	-	−0.12(0.14)	−0.25(0.18)	0.02(0.13)	0.01(0.13)	0.04(0.13)	−0.18(0.12)	−2.47(0.12)	0.05(0.12)
nMSE	-	9.42(0.85)	16.41(1.68)	8.95(0.84)	7.77(0.53)	7.94(0.52)	7.11(0.49)	12.91(0.90)	7.58(0.51)
Cover	-	0.94(0.01)	0.90(0.01)	0.92(0.01)	0.95(0.01)	0.95(0.01)	0.93(0.01)	0.65(0.02)	0.95(0.01)
Length	-	0.57(0.01)	0.62(0.01)	0.52(0.00)	0.55(0.00)	0.56(0.00)	0.53(0.00)	0.35(0.00)	0.55(0.00)
β is dense, $R^2 = 0.8$, $\widehat{s}_{\min} = 10$, $\eta = 0.25$, $n = 100$									
$\sqrt{n}$Bias	-	−0.80(0.23)	0.63(0.83)	−0.09(0.23)	−0.19(0.22)	0.01(0.22)	−0.94(0.21)	−2.54(0.19)	−0.03(0.22)
nMSE	-	27.42(2.23)	35.65(3.14)	27.57(2.54)	23.67(1.57)	24.62(1.59)	24.21(1.47)	24.57(1.71)	24.99(1.58)
Cover	-	0.92(0.01)	0.89(0.01)	0.90(0.01)	0.93(0.01)	0.94(0.01)	0.89(0.01)	0.71(0.02)	0.91(0.01)
Length	-	0.91(0.01)	1.02(0.01)	0.81(0.01)	0.85(0.01)	0.88(0.01)	0.79(0.00)	0.53(0.00)	0.89(0.00)
β is moderately sparse, $\widehat{s}_{\min} = 10$, $R^2 = 0.5$, $\eta = 0.1$, $n = 200$									
$\sqrt{n}$Bias	0.03(0.07)	−0.12(0.08)	0.14(0.08)	0.12(0.08)	0.03(0.08)	0.06(0.08)	−0.21(0.08)	−1.56(0.08)	0.10(0.09)
nMSE	2.36(0.15)	3.22(0.19)	3.31(0.20)	3.09(0.18)	3.07(0.20)	3.31(0.20)	2.97(0.18)	5.39(0.35)	3.64(0.21)
Cover	0.93(0.01)	0.95(0.01)	0.96(0.01)	0.93(0.01)	0.92(0.01)	0.94(0.01)	0.92(0.01)	0.65(0.02)	0.97(0.01)
Length	0.29(0.00)	0.35(0.00)	0.36(0.00)	0.33(0.00)	0.33(0.00)	0.35(0.00)	0.33(0.00)	0.23(0.00)	0.38(0.00)
β is moderately sparse, $\widehat{s}_{\min} = 10$, $R^2 = 0.8$, $\eta = 0.1$, $n = 200$									
$\sqrt{n}$Bias	0.03(0.07)	−0.59(0.08)	0.12(0.12)	0.01(0.08)	−0.00(0.08)	−0.03(0.08)	−0.50(0.08)	−1.15(0.08)	0.12(0.11)
nMSE	2.36(0.15)	3.30(0.22)	6.88(0.39)	2.83(0.18)	2.89(0.18)	3.21(0.19)	3.58(0.21)	4.55(0.30)	5.59(0.30)
Cover	0.93(0.01)	0.91(0.01)	0.96(0.01)	0.91(0.01)	0.94(0.01)	0.95(0.01)	0.92(0.01)	0.73(0.02)	0.95(0.01)
Length	0.29(0.00)	0.32(0.00)	0.52(0.00)	0.29(0.00)	0.33(0.00)	0.34(0.00)	0.33(0.00)	0.24(0.00)	0.44(0.00)

When β is not sparse, we omit the results for “Oracle”. The numbers in the parenthesis are the standard errors of the estimated values. The nominal coverage probability is 0.95.

4.2 ATE estimation with doubly robust estimator

As the validity of regression adjustment estimator replies on a correctly specified linear model in (1.2) for the response, in this section, we discuss the doubly-robust estimator that combines regression imputation and inverse probability weighting. The doubly-robust estimator proposed by Robins and Rotnitzky (1995) and Hahn (1998) remains consistent if either the response model or propensity score model is misspecified. We will first discuss a generalization of the doubly-roust estimator that incorporates high dimensional covariates. Our analysis shows that as long as model selection is adopted in the estimation procedure, doubly robust estimator can not avoid the over-fitting bias issue. As doubly robust estimator was originally proposed in the semi-parametric inference literature, later in the section, we shall discuss the connection between the post-selection bias and the "over-fitting" bias in the semi-parametric literature.

4.2.1 Doubly robust estimator with repeated data splitting

We write the propensity score, i.e. the conditional probability of receiving the treatment given the covariates, as

$$e(X_i) = \mathbb{P}(D_i = 1|X_i). \tag{4.5}$$

To estimate the propensity score in high dimensions, we assume that the propensity score satisfies

$$\log\left(\frac{e(X_i)}{1-e(X_i)}\right) = X_i^{\mathrm{T}}\gamma + R_{ni}^{e}, \quad i = 1, \cdots, n, \tag{4.6}$$

where R_{ni}^{e} is an approximation error, and γ is a sparse vector of coefficients with $||\gamma||_0 \leq s_\gamma$. To estimate ATE, we work under the following assumption.

Assumption 16. *We assume that (a) mean independence:* $\mathbb{E}[Y(d)|D, X] = \mathbb{E}[Y(d)|X]$*, and (b) overlap:* $e(X)$ *is bounded away from zero for all* $d \in \{0, 1\}$*, i.e. there exist a positive constant* η *such that* $\eta \leq e(z) \leq 1 - \eta$*.*

The mean independence condition in (a) is a relaxation of the unconfoundedness assumption discussed in Rubin (1991). Intuitively, this condition requires that the treatment assignment mechanism is similar to randomization within the group of units that share similar features. The second overlap condition requires that for each unit in the treated group, there exists a good proxy for this unit in the control group.

Given an i.i.d sample $\{Y_i, D_i, X_i\}_{i=1}^n$, a doubly robust estimator proposed in Robins and Rotnitzky (1995) and Hahn (1998) is defined as $\widehat{\alpha}_{\text{DR}} = \widehat{\mu}_1 - \widehat{\mu}_0$ with

$$\begin{aligned}
\widehat{\mu}_1 &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_iY_i}{\widehat{e}(X_i)} + \frac{(\widehat{e}(X_i) - D_i)\widehat{\mu}_1(X_i)}{\widehat{e}(X_i)}\right], \\
\widehat{\mu}_0 &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-D_i)Y_i}{1-\widehat{e}(X_i)} + \frac{(D_i - \widehat{e}(X_i))\widehat{\mu}_0(X_i)}{1-\widehat{e}(X_i)}\right],
\end{aligned} \tag{4.7}$$

where $\widehat{\mu}_d(\cdot)$ and $\widehat{e}(\cdot)$ are some generic model-based estimators of $\mu_d(\cdot)$ and $e(\cdot)$ that may require additional specification. $\widehat{\alpha}_{\text{DR}}$ is a robust estimation of α in the sense that it remains consistent if the model for either $\mu_d(\cdot)$ or $e(\cdot)$ is misspecified.

In high dimensions, when model selection is adopted for estimating the propensity score $e(\cdot)$ and conditional mean $\mu_d(\cdot)$, a similar post-selection bias issue discussed in Chapter 2 also exists in $\widehat{\alpha}_{\text{DR}}$. To illustrate this bias issue, suppose the propensity score $e(X)$ is known for now, then the doubly robust estimator in (4.7) reduces to $\widehat{\alpha}^* = \widehat{\mu}_1^* - \widehat{\mu}_0^*$ where

$$\begin{aligned}
\widehat{\mu}_1^* &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_iY_i}{e(X_i)} + \frac{(\widehat{e}(X_i) - D_i)\widehat{\mu}_1(X_i)}{e(X_i)}\right], \\
\widehat{\mu}_0^* &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1-D_i)Y_i}{1-e(X_i)} + \frac{(D_i - e(X_i))\widehat{\mu}_0(X_i)}{1-e(X_i)}\right].
\end{aligned}$$

If we estimate $\mu_d(X)$ with the refitted OLS estimator $\widehat{\mu}_{d,\text{OLS}}(X) = X'\widehat{\beta}_{\text{OLS}}$ discussed in Remark 1, we have in general $\mathbb{E}[Y - \widehat{\mu}_{d,\text{OLS}}(X)] \neq 0$ unless perfect model selection is achieved. Therefore, even if the propensity score is known, the doubly robust estimator is still biased due to either over-fitting or under-fitting. To avoid the post selection bias, we propose an extension of R-Split that incorporates doubly-robust estimator. We formalize this idea in Algorithm 6.

Similar to R-Split in Chapter 2, any reasonable model selection method can be adopted in Step 2. In Step 3, because the estimated functionals $\widehat{e}(\cdot)$ and $\widehat{\mu}_d(\cdot)$ are independent with the data in S_2, the over-fitting bias can be removed. In theory, we require that B equals the number of all possible subsample of size n_2, and the doubly-robust smoothed estimator can

Algorithm 6 ATE: R-Split with Doubly-robust estimator

```
For $b \leftarrow 1$ to $B$ do
    Step 1. Randomly split the data $\{(Y_i, D_i, X_i)\}_{i=1}^{n}$ into group $S_1$ of size $n_1$
        and group $S_2$ of size $n_2 = n - n_1$. Let $v_{bi} = 1_{(i \in S_2)}$, for $i = 1, \cdots, n$.
    Step 2. Obtain $\widehat{e}(\cdot)$ and $\widehat{\mu}_d(\cdot)$ on $S_1$ for $d \in \{0, 1\}$.
    Step 3. "Predict" on $S_2$:
        $\widehat{\mu}_{1,b} = \frac{1}{n_2} \sum_{i \in S_2} \left[ \frac{D_i Y_i}{\widehat{e}(X_i)} + \frac{(\widehat{e}(X_i) - D_i)\widehat{\mu}_1(X_i)}{\widehat{e}(X_i)} \right]$
        $\widehat{\mu}_{0,b} = \frac{1}{n_2} \sum_{i \in S_2} \left[ \frac{(1-D_i)Y_i}{1-\widehat{e}(X_i)} + \frac{(D_i - \widehat{e}(X_i))\widehat{\mu}_0(X_i)}{1-\widehat{e}(X_i)} \right].$
The doubly-robust "smoothed" estimator of $\alpha$ is $\widetilde{\alpha}_{\text{DR}} = \frac{1}{B} \sum_{b=1}^{B} (\widehat{\mu}_{1,b} - \widehat{\mu}_{0,b})$.
```

be viewed $\widetilde{\alpha}_{\text{DR}} = \mathbb{E}\left(\widehat{\mu}_{1,b} - \widehat{\mu}_{0,b} | \mathcal{X}\right)$. To study the theoretical property of $\widetilde{\alpha}_{\text{DR}}$, we assume that the observed data follow the data generating process listed in Assumption 1, and the model selection procedure satisfies Assumption 17.

Assumption 17. *For all $d \in \{0, 1\}$, the estimators of $\mu_d(\cdot)$ and $e(\cdot)$ satisfy that*

(a) $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)} \right)^2 \middle| \mathcal{X} \right] = o_p(1).$

(b) $\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\mu}_V^d(X_i) - \mu_d(X_i) \right)^2 \middle| \mathcal{X} \right] = o_p(1).$

(c) $\mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} (\widehat{\mu}_{1,v_b}(X_i) - \mu_1(X_i))^2 \right)^{1/2} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)} \right)^2 \right)^{1/2} \middle| \mathcal{X} \right] = o_p(n^{-1/2}).$

Assumption 17 (a) requires estimation bias in $\widehat{e}_V(\cdot)$ to be negligible after aggregating over all splits. If we use logistic Lasso proposed in Meier et al. (2008) to estimate the propensity score, this assumption places a limit on the sparsity level of s_γ in model (4.6). As the asymptotic results in Corollary 1 of Farrell (2015) indicate that $\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)} \right)^2 = O_p(s_\gamma \log p / n)$ under some mild conditions, we expect Assumption 17 (a) to be satisfied as long as $s_\gamma \log p / n = o(1)$ but further investigation is necessary. Assumption 17 (b) is similar to Assumption 5 in Chapter 2, see Chapter 3.5.5 for more discussion. Assumption 17 (c) requires a rate on the product of errors and is thus easier to satisfy as long as one function is moderately sparse; see Chernozhukov et al. (2018) and Dukes et al. (2018) for more discussion.

Theorem 3. *Under Assumption 1, 16, and 17, by letting $\widetilde{\mu}_d = \mathbb{E}(\widehat{\mu}_{d,V}|\mathcal{X})$, we have for $d \in \{0,1\}$,*

$$\sqrt{n}(\widetilde{\mu}_d - \mu_d) = \frac{1}{n}\sum_{i=1}^{n}\phi_d(Z_i) + o_p(1),$$

where

$$\phi_1(Z_i) = \frac{D_iY_i}{e(X_i)} - \mu_1 + \frac{(e(X_i) - D_i)\mu_1(X_i)}{e(X_i)}$$
$$\phi_0(Z_i) = \frac{(1 - D_i)Y_i}{1 - e(X_i)} - \mu_0 + \frac{(D_i - e(X_i))\mu_0(X_i)}{1 - e(X_i)}$$

Given the linear expression above, we have $\sqrt{n}\sigma_d^{-1}(\widetilde{\mu}_d - \mu_d) \rightsquigarrow N(0,1)$, where $\sigma_d^2 = \mathbb{E}\left[\frac{Var(\varepsilon_i^d)}{e(X)\mathbf{1}_{(d=1)}+(1-e(X))\mathbf{1}_{(d=0)}}\right] + \mathbb{E}\left(\mu_d(X_i) - \mu_d\right)^2$.

4.2.2 Comparison between R-Split estimator and $\widetilde{\alpha}_{\text{DR}}$

In this section, we provide a heuristic comparison on the statistical asymptotic efficiencies between R-Split estimator and $\widetilde{\alpha}_{\text{DR}}$ discussed in Algorithm 6. As R-Split works with model (1.2), a simplified asymptotic variance expression of $\widetilde{\alpha}_{\text{DR}}$ under this model shall be provided. Given model (1.2), we have $\mathbb{E}\left(\mu_d(X_i) - \mu_d\right)^2 = 0$, therefore under the assumptions for Theorem 3

$$\sigma_{\text{DR}}^2 = \mathbb{E}\left(\frac{\sigma_\varepsilon^2}{e(X_i)(1 - e(X_i))}\right). \tag{4.8}$$

where σ_{DR}^2 is the asymptotic variance of $\sqrt{n}(\widetilde{\alpha}_{\text{DR}} - \alpha)$. Given a fixed model M, following a similar derivation in Section 3.5.4, we have $(\Sigma_M^{-1})_{11} = \text{Var}(D) - \text{Var}(\mathbb{E}(D|X_M)) = \mathbb{E}(\text{Var}(D|X_M))$. Therefore for R-Split, given the derivation in (3.12) and the fact that

$D \in \{0,1\}$ is a binary random variable, we have

$$\widetilde{\sigma}_n^2 \leq \mathbb{E}\left(\frac{\sigma_\varepsilon^2}{e(X_{i,\widehat{M}})(1-e(X_{i,\widehat{M}}))}\bigg|\boldsymbol{\mathcal{X}}\right) + o_p(1). \tag{4.9}$$

From the comparison between (4.8) and (4.9), unless $e(X_{\widehat{M}}) \approx e(X)$ for possible models $\widehat{M}$, R-Split estimator has smaller asymptotic variance than $\widetilde{\alpha}_{\text{DR}}$. Recall that the propensity score $e(X_{\widehat{M}})$ is essentially the proportion of the treated units in the strata with features $X_{\widehat{M}}$. In observational studies, $e(X)$ is likely to be much smaller than $e(X_{\widehat{M}})$ since $e(X)$ calculates the proportion of the treated units in a much finer strata. On the other hand, in randomized trials, because treatments are randomly assigned and the probability of being assigned to the treatment group equals $P(D=1)$ in all possible strata, therefore we normally have $e(X_{\widehat{M}}) = e(X)$. As a conclusion, we expect that R-Split is more efficient than $\widehat{\alpha}_{\text{DR}}$ in observational studies.

4.3 Heterogeneous treatment effects

In this section, we propose a potential extension of R-Split for estimating heterogeneous treatment effect, which can be particularly interesting in subgroup analysis ; see Wang et al. (2007) and Pocock et al. (2002). For example, studies in marketing often try to identify a subgroup of individuals for whom a job training program is most beneficial; studies in biomedical science evolve identifying subgroups of patients for whom the treatment may be most beneficial or harmful. Therefore, in this section, we consider the problem of estimating

$$\mu_d(w) = \mathbb{E}[Y(d)|W=w], \quad \mu_d(\mathcal{W}) = \mathbb{E}[Y(t)|W \in \mathcal{W}], \text{ for } d \in \{0,1\},$$

where W is a continuous treatment or exposure variable as part of the covariates X in the sense that $X = (W, Z^{\mathrm{T}})^{\mathrm{T}}$. Then, the heterogeneous treatment effect can be calculated by the difference between $\mu_1(w)$ and $\mu_0(w)$, i.e.,

$$\tau(w) = \mathbb{E}[Y(1) - Y(0)|W = w] = \mu_1(w) - \mu_0(w).$$

As the quantities of interest are defined in terms of the potential outcomes that are not observed, we must consider assumptions under which these quantities can be expressed based on observed data. Thus, in this section, we again work under Assumption 16. Under Assumption 16 and given $W = w$, $\mu_1(w)$ can be identified with the observed data through

$$\begin{aligned}\mu_1(w_0) =& \mathbb{E}[Y(1)|W = w] \\ =& \mathbb{E}\left\{\mathbb{E}\left(Y(1)|Z, W = w\right)|W = w\right\} \\ =& \mathbb{E}\left\{\mathbb{E}\left(Y(1)|D = 1, Z, W = w\right)|W = w\right\} \\ =& \mathbb{E}\left\{\mathbb{E}\left(Y|D = 1, Z, W = w\right)|W = w\right\},\end{aligned}$$

then the conditional treatment effect equals

$$\tau(w) = \mathbb{E}\left\{\mathbb{E}\left(Y|D = 0, Z, W = w\right)|W = w\right\} - \mathbb{E}\left\{\mathbb{E}\left(Y|D = d, X\right)|W = w\right\}.$$

Following the derivations above, it would be interesting to derive the semi-parametric efficiency bound and the efficient influence function for estimating $\tau(w)$, but further studies are needed.

4.4 Proofs

4.4.1 Proof of Theorem 3

As the proofs for $\widetilde{\mu}_1$ and $\widetilde{\mu}_0$ are very similar, to avoid redundancy, we focus on $\widetilde{\mu}_1$. In a single split, $\widehat{\mu}_{1,b}$ can be decomposed

$$\widehat{\mu}_{1,v_b} - \mu_1 = \frac{1}{n_2} \sum_{i=1}^{n} v_{bi} \phi_1(Z_i) + (r_{n1,b} + r_{n2,b} + r_{n3,b})/\sqrt{n_2},$$

where

$$\begin{aligned}
\phi_1(Z_i) &= \frac{D_i Y_i}{e(X_i)} - \mu_1 + \frac{(e(X_i) - D_i)\mu_1(X_i)}{e(X_i)}, \\
r_{n1,b} &= \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n} v_{bi} D_i (Y_i - \mu_1(X_i)) \left(\frac{1}{\widehat{e}_{v_b}(X_i)} - \frac{1}{e(X_i)} \right), \\
r_{n3,b} &= \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n} v_{bi} (\widehat{\mu}_{1,v_b}(X_i) - \mu_1(X_i)) \cdot \frac{e(X_i) - D_i}{e(X_i)}, \\
r_{n3,b} &= \frac{1}{\sqrt{n_2}} \sum_{i=1}^{n} v_{bi} D_i (\widehat{\mu}_{1,v_b}(X_i) - \mu_1(X_i)) \frac{\widehat{e}_{v_b}(X_i) - p_t(X_i)}{\widehat{e}_{v_b}(X_i) p_t(X_i)}.
\end{aligned}$$

As the weights are independently generated with data, the smoothed estimator is of the form $\widetilde{\mu}_t = \mathbb{E}(\widehat{\mu}_{1,v_b} | \boldsymbol{\mathcal{X}})$ satisfies

$$\begin{aligned}
\sqrt{n}(\widetilde{\mu}_t - \mu_1) &= \sqrt{n} \left(\mathbb{E}(\widehat{\mu}_{1,V} | \boldsymbol{\mathcal{X}}) - \mu_1 \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \phi(Z_i) + \mathbb{E}(r_{n1,b} + r_{n2,b} + r_{n3,b} | \boldsymbol{\mathcal{X}}).
\end{aligned}$$

The proof proceed by showing that $\mathbb{E}(r_{ni,b}|\boldsymbol{\mathcal{X}}) = o_p(1)$ for $i = 1, 2, 3$. Applying the mean independence condition in (16), we may write $r_{1n,b}$ as

$$\begin{aligned}
\mathbb{E}(r_{1n,b}|\boldsymbol{\mathcal{X}}) &= \mathbb{E}\left[\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n} V_i D_i (Y_i - \mu_1(X_i))\left(\frac{1}{\widehat{e}_{v_b}(X_i)} - \frac{1}{e(X_i)}\right)\bigg|\boldsymbol{\mathcal{X}}\right] \\
&= \mathbb{E}\left[\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n} V_i D_i \varepsilon_i^1 \left(\frac{1}{\widehat{e}_{v_b}(X_i)} - \frac{1}{e(X_i)}\right)\bigg|\boldsymbol{\mathcal{X}}\right] \\
&= \frac{1}{B}\sum_{b=1}^{B}\frac{1}{\sqrt{n_2}}\sum_{i=1}^{n} v_{ib} D_i \varepsilon_i^1 \left(\frac{1}{\widehat{e}_{v_b}(X_i)} - \frac{1}{e(X_i)}\right),
\end{aligned}$$

To this end, it is enough to control the variance of $\mathbb{E}(R_{11,b}|\boldsymbol{\mathcal{X}})$. By data splitting, the noise term ε_i^1 is independent of $\widehat{e}_V(\cdot)$, we have

$$\begin{aligned}
\text{Var}(\mathbb{E}(R_{11,b}|\boldsymbol{\mathcal{X}})|\{Z_i\}_{i=1}^n) &\le \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_2}\text{Var}\left[\sum_{i=1}^{n} v_{ib} D_{it} \varepsilon_i^1 \left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)}\right)\bigg|\{Z_i\}_{i=1}^n\right] \\
&= \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_2}\sum_{i=1}^{n}\sigma_{\varepsilon^1}^2 \mathbb{E}\left[D_{it} v_{ib}\left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)}\right)^2\bigg|\{Z_i\}_{i=1}^n\right] \\
&\le \sigma_{\varepsilon^1}^2 \frac{1}{B}\sum_{b=1}^{B}\frac{1}{n_2}\sum_{i=1}^{n} v_{ib}\mathbb{E}\left[\left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)}\right)^2\bigg|\{Z_i\}_{i=1}^n\right] \\
&= \sigma_{\varepsilon^1}^2 \sum_{i=1}^{n}\left(\frac{1}{n_2}\frac{1}{B}\sum_{b=1}^{B} v_{ib}\right)\mathbb{E}\left[\left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)}\right)^2\bigg|\{Z_i\}_{i=1}^n\right] \\
&= \sigma_{\varepsilon^1}^2 \mathbb{E}\left[\frac{1}{n}\left(\frac{1}{\widehat{e}_V(X_i)} - \frac{1}{e(X_i)}\right)^2\bigg|\{Z_i\}_{i=1}^n\right] = o_p(1),
\end{aligned}$$

where the last step is obtained by Assumption 17 (a). Variance of $\mathbb{E}(R_{12,b}|\boldsymbol{\mathcal{X}})$ can be controlled similarly by Assumption 17 (b).

Applying Assumption 17 (c) and by Cauchy's inequality, we have

$$
\begin{aligned}
\mathbb{E}(R_{n2,b}|\boldsymbol{\mathcal{X}}) &= \mathbb{E}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i(\widehat{\mu}_{1,V}(X_i)-\mu_1(X_i))\frac{\widehat{e}_V(X_i)-e(X_i)}{\widehat{e}_V(X_i)e(X_i)}\middle|\boldsymbol{\mathcal{X}}\right] \\
&\leq \sqrt{n}\,\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{1,v_b}(X_i)-\mu_1(X_i))^2\right)^{1/2}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\widehat{e}_V(X_i)}-\frac{1}{e(X_i)}\right)^2\right)^{1/2}\middle|\boldsymbol{\mathcal{X}}\right] \\
&= o_p(1).
\end{aligned}
$$

CHAPTER 5

Conclusion and Future Work

In the previous chapters, we address 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 also 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 thesis, 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.

Our current work on post selection inference discuss the bias issues under the framework of high dimensional linear or partially linear sparse models. In the future, built on these insights, it would be interesting to generalize our work to other problems. For example, the inferential problems in high dimensional dense models or in mis-specified regression models could be potentially interesting. The motivation for this line of research is that any statistical model is only an approximation to the real world. Therefore, it would be meaningful to push the limits of standard statistical technique and propose more robust alternatives whenever possible.

Post model selection inference also has broad applications in other statistical areas such as subgroup analysis briefly mentioned in Chapter 4. It is known in Thomas and Bornkamp (2017) that a naive estimation of the treatment effect, which ignores a subgroup selection has taken place, leads to biased estimates that give overly optimistic conclusions. As this bias in subgroup analysis is in a spirit similar to the post-selection over-fitting bias, the procedures proposed in the current thesis have the potential to provide valid inference after subgroup selection.

Another future direction is to develop "precision model selection" methods for estimating heterogeneous treatment effects. There is a vast literature on model selection for prediction, but relatively little work on model selection for causal inference. The purpose of what we call "precision model selection is to assign models tailored towards individuals on the basis of their unique characteristics that distinguish a given set of individuals from other individuals.

BIBLIOGRAPHY

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2018. doi: 10.1111/rssb.12268.

Alexandre Belloni, Victor Chernozhukov, et al. 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.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S Sekhon, and Bin Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

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.

Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Two-step estimation and inference with possibly many included covariates. Technical report, working paper, Michigan, 2017.

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.

Oliver Dukes, Vahe Avagyan, and Stijn Vansteelandt. High-dimensional doubly robust tests for regression parameters. *arXiv preprint arXiv:1805.06714*, 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, Qi-Man Shao, and Wen-Xin Zhou. Are discoveries spurious? distributions of maximum spurious correlations and their applications. *Ann. Statist.*, 46(3):989–1017, 06 2018. doi: 10.1214/17-AOS1575.

Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

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.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Jana Jankova and Sara van de Geer. Semi-parametric efficiency bounds for high-dimensional models. *arXiv preprint arXiv:1601.00815*, 2017.

Arun Kumar Kuchibhotla, Lawrence D Brown, Andreas Buja, Edward I George, and Linda Zhao. A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. *arXiv preprint arXiv:1802.05801*, 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.

Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedmans critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

J Lumley, S Oliver, and E Waters. Interventions for promoting smoking cessation during pregnancy. *Cochrane Database Syst Rev*, 2:CD001055, 2000.

Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 53–71, 2008.

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.

JS 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.

Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

Alessandro Rinaldo, Larry Wasserman, Max G'Sell, Jing Lei, and Ryan Tibshirani. Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*, 2016.

James M Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429): 122–129, 1995.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, Aad van der Vaart, et al. 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 nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Donald B Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, pages 1213–1234, 1991.

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

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Marius Thomas and Björn Bornkamp. Comparing approaches to treatment effect estimation for subgroups in clinical trials. *Statistics in Biopharmaceutical Research*, 9(2): 160–171, 2017.

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, Ruben Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Roman Vershynin. High-dimensional probability. *An Introduction with Applications*, 2016.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2017. doi: doi.org/10.1080/01621459.2017.1319839.

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.

Rui Wang, Stephen W Lagakos, James H Ware, David J Hunter, and Jeffrey M Drazen. Statistics in medicinereporting of subgroup analyses in clinical trials. *New England Journal of Medicine*, 357(21):2189–2194, 2007.

Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 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.

Wei Zheng, Kohta Suzuki, Taichiro Tanaka, Moriyasu Kohama, Zentaro Yamagata, Okinawa Child Health Study Group, et al. Association between maternal smoking during pregnancy and low birthweight: Effects by maternal age. *PloS one*, 11(1):e0146241, 2016.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.