Lasso adjustments of treatment effect estimates in randomized experiments

Adam Bloniarz* Hanzhong Liu† Cun-Hui Zhang‡ Jasjeet Sekhon§ Bin Yu¶

July 13, 2015

Abstract

In randomized experiments, linear regression is often used to adjust for imbalances in covariates between treatment groups, yielding an estimate of the average treatment effect with lower asymptotic variance than the unadjusted estimator. If there are a large number of covariates, many of which are irrelevant to the potential outcomes, the Lasso can be used to both select relevant covariates and perform the adjustment. We study the resulting estimator under the Neyman-Rubin model for randomization. We present conditions on the covariates and potential outcomes which guarantee that the Lasso is more efficient than the unadjusted estimator and provide a conservative estimate of the asymptotic variance. Simulation and data examples show that Lasso-based adjustment can be advantageous even when $p < n$ and that a variant of Lasso, cv(Lasso+OLS), is similar to cv(Lasso) in terms of confidence length and coverage, but outperforms cv(Lasso) with much fewer covariates in the adjustment.

Keywords

Randomized experiment, Neyman-Rubin model, average treatment effect, high-dimensional statistics, Lasso, concentration inequality

1. Introduction

Randomized experiments are widely used to measure causal effects of treatments. The power of a randomized design derives from the blunt mechanism of the treatment assignment: randomization ensures that treatment assignment is not influenced by any potential confounding factors, both observed and unobserved. Thus, they are particularly useful when there is no rigorous theory of a system's dynamics, and full identification of confounders would be impossible. This advantage was cast elegantly in a mathematical terms in the early 20th century by Jerzy Neyman, who introduced a simple model for a randomized experiment, and showed that the difference of average outcomes in the treatment and control groups is statistically unbiased for the Average Treatment Effect (ATE) over the experimental sample [1].

However, no experiment occurs in a vacuum of scientific knowledge. Usually, many details are recorded about the units in a study besides their outcome and assigned treatment status, such as pre-treatment covariates. While the treatment assignment mechanism may not be a function of these covariates, analyses of experimental outcomes often take them into account with the goal of improving accuracy. In modern randomized experiments, the number of covariates can be very large—sometimes even larger than the number of units in the study. In clinical trials overseen by regulatory bodies like the FDA and MHRA, demographic and genetic information may be recorded about each patient. In applications in the tech industry, where randomization is often called A/B testing, or digital experimentation, there is often a huge amount of behavioral data collected for each user. However, in this 'big data' setting, much of this data may be irrelevant to the outcome being studied, and selection of important covariates is necessary.

To ground our discussion, we examine a randomized trial of the Pulmonary Artery Catheter (PAC) that was carried out in 65 intensive care units in the UK between 2001 and 2004, called PAC-man [2]. The PAC is a monitoring device commonly inserted into critically ill patients after admission to intensive care, and it provides a continuous measurement of several indicators of cardiac activity. However, insertion of PAC is an invasive procedure that carries some risk of complications (including death), and it involves significant expenditure both in equipment costs and personnel [3]. Controversy over its use came to a head when an observational study found that PAC had an adverse effect on patient survival and led to increased cost of care [4]. This led to several large-scale randomized

*Department of Statistics, University of California, Berkeley, CA.
†Department of Statistics, University of California, Berkeley, CA.
‡Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ
§Department of Political Science, Department of Statistics, University of California, Berkeley, CA
¶Department of Statistics, Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA

trials, including PAC-man.

In the PAC-man trial, randomization of treatment was largely successful, and a number of covariates were measured about each patient in the study. If covariate interactions are included, the number of covariates exceeds the number of units in the study; however, few of them are predictive of the patient's outcome. As it turned out, the (pre-treatment) estimated probability of death, was imbalanced between the treatment and control groups (p = 0.005, Wilcoxon rank sum test). Because the control group had, on average, a slightly higher risk of death, the unadjusted difference-in-means estimator may overestimate the benefits of receiving a PAC. Adjustment for this imbalance seems advantageous in this case, since the pre-treatment probability of death is clearly predictive of health outcomes post-treatment.

In this paper, we study regression-based adjustment with Lasso selecting covariates. Regression adjustment regresses the outcome against treatment, covariates, and, following the method introduced in [5], treatment-covariate interactions. However, standard linear regression suffers from over-fitting if a large number of covariates and interaction terms are included in the model. In such cases, researchers sometimes perform model selection based on observing which covariates are unbalanced given the realized randomization. This generally leads to misleading inferences because of incorrect test levels [6]. The Lasso [7] provides researchers with an alternative that can mitigate these problems and still perform model selection. We define an estimator, $\widehat{ATE}_{\text{Lasso}}$, which is the estimated coefficient of the treatment variable after running an l_1-penalized linear regression of the outcome on treatment, covariates and treatment $\times$ covariate interactions. Because of the geometry of the l_1 penalty, the Lasso will usually set many regression coefficients to 0, and is well defined even if the number of covariates is larger than the number of observations. The Lasso's theoretical properties under the standard linear model have been widely studied in the last decade; consistency properties for coefficient estimation, model selection, and out-of-sample prediction are well understood (see [8] for an overview).

In this paper, instead of assuming that the standard linear model is the true data-generating mechanism, we work under the aforementioned non-parametric model of randomization introduced by Neyman [1] and popularized by Donald Rubin [9]. In this model, the outcomes and covariates are fixed quantities, and the treatment group is assumed to be sampled without replacement from a finite population. The treatment indicator, rather than an error term, is the source of randomness, and it determines which of two potential outcomes is revealed to the experimenter. Unlike the standard linear model, the Neyman-Rubin model makes few assumptions not guaranteed by the randomization itself. The setup of the model does rely on the stable unit treatment value assumption (SUTVA), which states that there is only one version of treatment, and that the potential outcome of one unit should be unaffected by the particular assignment of treatments to the other units; however it makes no assumptions of linearity or exogeneity of error terms. Ordinary Least Squares (OLS) [10][33][5], logistic regression [12], and post-stratification [13] are among the adjustment methods that have been studied under this model.

To be useful to practitioners, the Lasso-based estimator must be consistent and yield a method to construct valid confidence intervals. We outline conditions on the covariates and potential outcomes that will guarantee these properties. Moreover, an upper bound for the asymptotic variance can be estimated from the model residuals, yielding asymptotically conservative confidence intervals for the average treatment effect which can be substantially narrower than the unadjusted confidence intervals. Simulation studies are provided to show the advantage of the Lasso adjusted estimator and to show situations where it breaks down. We also discuss an application of the estimator to the PAC-man data, and compare the estimates and confidence intervals derived from the unadjusted, OLS-adjusted, and Lasso-adjusted methods. We also compare different methods of selecting the Lasso tuning parameter on this data.

2. Framework and definitions

We give a brief outline of the Neyman-Rubin model for a randomized experiment; the reader is urged to consult [1], [9], and [14] for more details. We follow the notation introduced in [10] and [5]. For concreteness, we illustrate the model in the context of a randomized controlled trial such as PAC-man.

For each individual in the trial, indexed by i, $i = 1, ..., n$, the model assumes the existence of a set of potential outcomes, one for each possible level or variant of the drug. In this work, we restrict ourselves to a binary treatment (drug vs. no drug); we denote the potential outcomes as a_i and b_i, referred to as the outcomes under treatment and control, respectively. For each individual, only one of these potential outcomes will be revealed to the experimenter, since an individual cannot be both treated and untreated. The causal effect of the treatment on individual i, defined to be $a_i - b_i$, is unobservable. In this sense, the Neyman-Rubin framework recasts causal inference as a missing data problem. Instead of trying to infer individual-level effects, we will focus our attention on estimation of the average causal effect over the whole population, as outlined in the next section.

The only randomness in the model comes from the as-

signment of treatment, which is under the control of the experimenter. We define random treatment indicators T_i, which take on a value 1 for a treated individual, or 0 for an untreated individual. We will assume that the set of treated units is drawn without replacement from the full population, where the size of the treatment group is fixed; thus the T_i are identically distributed but not independent. The model for the observed outcome for individual i, say Y_i, is

$$Y_i = T_i a_i + (1 - T_i) b_i.$$

Note that the model does not incorporate any information about covariates. However, we will assume we have measured a vector of baseline covariates for each individual i. These might include, for example, age, gender, and genetic makeup. We denote the covariates for individual i as the column vector $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T \in \mathbb{R}^p$ and the full design matrix of the experiment as $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$.

Define the set of treated units as $A = \{i \in \{1, ..., n\} : T_i = 1\}$, and similarly define the set of control units B. The number of treated and control units are $n_A = |A|$ and $n_B = |B|$, respectively, where $n_A + n_B = n$. To indicate averages of quantities over these subsets, we introduce the notation $\bar{\cdot}_A$ and $\bar{\cdot}_B$. Thus, for example, the average value of the potential outcomes and the covariates in the treatment group are

$$\bar{a}_A = n_A^{-1} \textstyle\sum_{i \in A} a_i, \;\; \bar{\mathbf{x}}_A = n_A^{-1} \textstyle\sum_{i \in A} \mathbf{x}_i,$$

respectively. Note that these are random quantities, since the set A is determined by the random treatment assignment. When we want to take the average over the whole population, we will use the notation $\bar{\cdot}$, such as

$$\bar{a} = n^{-1} \textstyle\sum_{i=1}^n a_i, \;\; \bar{b} = n^{-1} \textstyle\sum_{i=1}^n b_i, \;\; \bar{\mathbf{x}} = n^{-1} \textstyle\sum_{i=1}^n \mathbf{x}_i.$$

3. Treatment effect estimation

Our main inferential goal will be estimation of the average effect of the treatment over the whole population. In our notation, this parameter is defined as

$$ATE = \bar{a} - \bar{b}.$$

The most natural estimator arises by replacing the population averages with the sample averages. This is defined as

$$\widehat{ATE}_{\text{unadj}} \stackrel{\text{def}}{=} \bar{a}_A - \bar{b}_B,$$

The subscript "unadj" indicates an estimate without regression adjustment. The foundational work in [1] points out that, under a randomized assignment of treatment, $\widehat{ATE}_{\text{unadj}}$ is unbiased for ATE, and derives a conservative procedure for estimating its variance.

While $\widehat{ATE}_{\text{unadj}}$ is an attractive estimator, covariate information can be used to make adjustments in the hope of reducing variance. A commonly used estimator is

$$\widehat{ATE}_{\text{adj}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(a)}\right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(b)}\right]$$

where $\hat{\boldsymbol{\beta}}^{(a)}, \hat{\boldsymbol{\beta}}^{(b)} \in \mathbb{R}^p$ are adjustment vectors derived from some procedure. The terms $\bar{\mathbf{x}}_A - \bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_B - \bar{\mathbf{x}}$ represent the fluctuation of the covariates in the subsample relative to the full sample, and the adjustment vectors fit the linear relationships between the covariates and potential outcomes under treatment and control. This is equivalent to imputing the unobserved potential outcomes; if we define

$$\hat{\bar{a}}_B = \bar{a}_A + (\bar{\mathbf{x}}_B - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}, \;\; \hat{\bar{b}}_A = \bar{b}_B + (\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)},$$

we can form the equivalent estimator

$$\widehat{ATE}_{\text{adj}} = n^{-1} \left(n_A \bar{a}_A + n_B \hat{\bar{a}}_B\right) - n^{-1} \left(n_B \bar{b}_B + n_A \hat{\bar{b}}_A\right).$$

If the adjustment vectors are fixed (non-random), then this estimator is still unbiased, but may have substantially smaller asymptotic and finite-sample variance than the unadjusted estimator.

In practice, the "ideal" linear adjustment vectors are unobservable. However, they can be estimated from the data, possibly at the expense of modest finite-sample bias. In the classical setup, when p is fixed and $p < n$, OLS can be used. The asymptotic properties of this kind of estimator are explored under the Neyman model in [33], [12], and [5]. We will follow a particular scheme which is studied in [5] and shown to have favorable properties. In this scheme, adjustment vectors are formed by regressing the outcome on treatment indicators, covariates, and treatment × covariate interactions. This is equivalent to running separate regressions in the treatment and control groups of outcome against an intercept and covariates. If we define $\hat{\boldsymbol{\beta}}^{(a)}_{\text{OLS}}$ and $\hat{\boldsymbol{\beta}}^{(b)}_{\text{OLS}}$ as the coefficients from the separate regressions, then the estimator is

$$\widehat{ATE}_{\text{OLS}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(a)}_{\text{OLS}}\right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(b)}_{\text{OLS}}\right].$$

This has some finite-sample bias, but [5] shows that it vanishes quickly at the rate of $1/n$. Moreover, for fixed p, under regularity conditions, the inclusion of interaction terms guarantees that it never has higher asymptotic variance than the unadjusted estimator, and asymptotically conservative confidence intervals for the true parameter can be constructed.

We will consider a similar scheme, but instead put ourselves in the high dimensional scenario where p is allowed

to be much larger than n and thus OLS may perform badly or be ill-posed. In this case we propose estimating the adjustment vectors using the Lasso [7]. The adjustment vectors would take the form

$$\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n_A} \sum_{i \in A} \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \boldsymbol{\beta} \right)^2 + \lambda_a \sum_{j=1}^{p} |\beta_j| \right], \tag{1}$$

$$\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{2n_B} \sum_{i \in B} \left(b_i - \bar{b}_B - (\mathbf{x}_i - \bar{\mathbf{x}}_B)^T \boldsymbol{\beta} \right)^2 + \lambda_b \sum_{j=1}^{p} |\beta_j| \right], \tag{2}$$

and the proposed Lasso adjusted ATE estimator is[1]

$$\widehat{ATE}_{\text{Lasso}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}} \right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B - \bar{\mathbf{x}})^T \hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}} \right].$$

λ_a and λ_b are regularization parameters for the Lasso which must be chosen by the experimenter; simulations show that cross-validation works well. In the next section, we study this estimator under the Neyman-Rubin model, and provide conditions on the potential outcomes, the covariates and the regularization parameters under which it enjoys similar asymptotic and finite-sample advantages as $\widehat{ATE}_{\text{OLS}}$.

We may also consider adjustments based on potential covariates $\mathbf{x}^{(a)}_i$ and $\mathbf{x}^{(b)}_i$ as in [15], so that $\mathbf{x}_i = \mathbf{x}^{(a)}_i T_i + \mathbf{x}^{(b)}_i (1 - T_i)$ are observed. We may still define $\widehat{ATE}_{\text{Lasso}}$ with adjustment vectors from two separate Lasso regressions, $\{a_i - \bar{a}_A, i \in A\}$ against $\{\mathbf{x}^{(a)}_i - \bar{\mathbf{x}}^{(a)}_A, i \in A\}$ and $\{b_i - \bar{b}_B, i \in B\}$ against $\{\mathbf{x}^{(b)}_i - \bar{\mathbf{x}}^{(b)}_A, i \in B\}$. Moreover, our analysis is carried out separately for the two potential populations $\{a_i, \mathbf{x}^{(a)}_i, i \le n\}$ and $\{b_i, \mathbf{x}^{(b)}_i, i \le n\}$, so that Theorems 1 and 2 presented below will prevail when their conditions hold for both potential populations. However, when $\mathbf{x}^{(a)}_i \neq \mathbf{x}^{(b)}_i$, the Lasso or OLS adjustments are no longer guaranteed to have smaller or equal asymptotic variance than the unadjusted one, even in the case of fixed p. In practice, one may still choose between the adjusted and unadjusted estimators based on the widths of the corresponding confidence intervals.

[1] To simplify the notation, we omit the dependence of $\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}}$, $\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}}$, λ_a and λ_b on the population size n.

4. Theoretical results

4.1. Notation

For a vector $\boldsymbol{\beta} \in R^p$ and a subset $S \subset \{1, ..., p\}$, denote S^c as the complement of S and define $\boldsymbol{\beta}_S = \{\beta_j : j \in S\}$, where β_j is the j-th component of $\boldsymbol{\beta}$. Let $|S|$ denote the cardinality of the set S. For any column vector $\mathbf{u} = (u_1, ..., u_m)^T$, we denote $\|\mathbf{u}\|_2^2 = \sum_{i=1}^m u_i^2$, $\|\mathbf{u}\|_1 = \sum_{i=1}^m |u_i|$, $\|\mathbf{u}\|_\infty = \max_{i=1,...,m} |u_i|$ and $\|\mathbf{u}\|_0 = |\{j : u_j \neq 0\}|$. For a given $m \times m$ matrix D, let $\lambda_{\min}(D)$ and $\lambda_{\max}(D)$ denote the smallest and largest eigenvalues of D respectively. D^{-1} denotes the inverse of the matrix D. Let $\xrightarrow{d}$ and $\xrightarrow{p}$ denote convergence in distribution and in probability, respectively.

4.2. Decomposition of the potential outcomes

The Neyman-Rubin model does not assume a linear relationship between the potential outcomes and the covariates. In order to study the properties of adjustment under this model, we decompose the potential outcomes into a term linear in the covariates and an error term. Given vectors of coefficients $\boldsymbol{\beta}^{(a)}, \boldsymbol{\beta}^{(b)} \in \mathbb{R}^p$, we write[2] for $i = 1, ..., n$,

$$a_i = \bar{a} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(a)} + e^{(a)}_i, \tag{3}$$

$$b_i = \bar{b} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(b)} + e^{(b)}_i. \tag{4}$$

Note that we have not added any assumptions to the model; we have simply defined unit-level residuals, $e^{(a)}_i$ and $e^{(b)}_i$, given the vectors $\boldsymbol{\beta}^{(a)}, \boldsymbol{\beta}^{(a)}$. All the quantities in 3 and 4 are fixed, deterministic numbers. It is easy to verify that $\bar{e}^{(a)} = \bar{e}^{(b)} = 0$. In order to pursue a theory for the Lasso, we will add assumptions on the populations of a_i's, b_i's, and $\mathbf{x}_i$'s, and we will assume the existence of $\boldsymbol{\beta}^{(a)}, \boldsymbol{\beta}^{(b)}$ such that the error terms satisfy certain assumptions.

4.3. Conditions

We will need the following to hold for both the treatment and control potential outcomes. The first set of assumptions (1-3) are similar to those found in [5].

Condition 1 *Stability of treatment assignment probability.*

$$n_A/n \to p_A, \text{ as } n \to \infty \tag{5}$$

for some $p_A \in (0, 1)$.

[2] Again, we omit the dependence of $\boldsymbol{\beta}^{(a)}$, $\boldsymbol{\beta}^{(b)}$, λ_a, λ_b, $e^{(a)}$ and $e^{(b)}$ on n.

Condition 2 *The centered moment conditions. There exists a fixed constant $L > 0$ such that, for all $n = 1, 2, ...$ and $j = 1, ..., p$,*

$$n^{-1} \sum_{i=1}^{n} (x_{ij} - (\bar{\mathbf{x}})_j)^4 \leq L; \tag{6}$$

$$n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^4 \leq L; \quad n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^4 \leq L. \tag{7}$$

Condition 3 *The means $n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^2$, $n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^2$ and $n^{-1} \sum_{i=1}^{n} e_i^{(a)} e_i^{(b)}$ converge to finite limits.*

Since we consider the high-dimensional setting where p is allowed to be much larger than n, we need additional assumptions to ensure that the Lasso is consistent for estimating $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$. Before stating them, we define several quantities.

Definition 1 *Given $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$, the sparsity measures for treatment and control groups, $s^{(a)}$ and $s^{(b)}$, are defined as the number of nonzero elements of $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$, i.e.,*

$$s^{(a)} = |\{j : \beta_j^{(a)} \neq 0\}|, \; s^{(b)} = |\{j : \beta_j^{(b)} \neq 0\}|, \tag{8}$$

respectively. We will allow $s^{(a)}$ and $s^{(b)}$ to grow with n, though the notation does not explicitly show this.

Definition 2 *Define δ_n to be the maximum covariance between the error terms and the covariates.*

$$\delta_n = \max_{\omega=a,b} \left\{ \max_j \left| \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - (\bar{\mathbf{x}})_j) \left(e_i^{(\omega)} - \bar{e}^{(\omega)} \right) \right| \right\}. \tag{9}$$

The following conditions will guarantee that the Lasso consistently estimates the adjustment vectors $\boldsymbol{\beta}^{(a)}, \boldsymbol{\beta}^{(b)}$ at a fast enough rate to ensure asymptotic normality of $\widehat{ATE}_{\text{Lasso}}$. It is an open question whether a weaker form of consistency would be sufficient for our results to hold.

Condition 4 *Decay and scaling. Let $s = \max\left\{s^{(a)}, s^{(b)}\right\}$.*

$$\delta_n = o\left(\frac{1}{s\sqrt{\log p}} \right). \tag{10}$$

$$(s \log p)/\sqrt{n} = o(1). \tag{11}$$

Condition 5 *Cone invertibility factor. Define the Gram matrix as $\Sigma = n^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$: There exist constants $C > 0$ and $\xi > 1$ not depending on n, such that*

$$\|\mathbf{h}_S\|_1 \leq Cs \|\Sigma \mathbf{h}\|_\infty, \; \forall \mathbf{h} \in \mathcal{C}, \tag{12}$$

with $\mathcal{C} = \{\mathbf{h} : \|\mathbf{h}_{S^c}\|_1 \leq \xi \|\mathbf{h}_S\|_1\}$, and

$$S = \{j : \boldsymbol{\beta}_j^{(a)} \neq 0 \text{ or } \boldsymbol{\beta}_j^{(b)} \neq 0\}. \tag{13}$$

Condition 6 *Let $\tau = \min\left\{1/70, (3p_A)^2/70, (3 - 3p_A)^2/70\right\}$. For constants $0 < \eta < \frac{\xi - 1}{\xi + 1}$ and $0 < M < \infty$, assume the regularization parameters of the Lasso belong to the sets*

$$\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_A} \sqrt{\frac{2 \log p}{n}} + \delta_n \right), \tag{14}$$

$$\lambda_b \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_B} \sqrt{\frac{2 \log p}{n}} + \delta_n \right). \tag{15}$$

Denote respectively the population variances of $e^{(a)}$ and $e^{(b)}$ and the population covariance between them by

$$\sigma^2_{e^{(a)}} = n^{-1} \sum_{i=1}^{n} (e_i^{(a)})^2, \quad \sigma^2_{e^{(b)}} = n^{-1} \sum_{i=1}^{n} (e_i^{(b)})^2,$$

$$\sigma_{e^{(a)} e^{(b)}} = n^{-1} \sum_{i=1}^{n} e_i^{(a)} e_i^{(b)}.$$

Theorem 1 *Assume conditions 1 through 6 hold for some $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$. Then*

$$\sqrt{n} \left(\widehat{ATE}_{\text{Lasso}} - ATE \right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \tag{16}$$

where

$$\sigma^2 = \lim_{n \to \infty} \left[\frac{1 - p_A}{p_A} \sigma^2_{e^{(a)}} + \frac{p_A}{1 - p_A} \sigma^2_{e^{(b)}} + 2\sigma_{e^{(a)} e^{(b)}} \right]. \tag{17}$$

The proof of Theorem 1 is given in the supplemental material. It is easy to show, as in the following corollary of Theorem 1, that the asymptotic variance of $\widehat{ATE}_{\text{Lasso}}$ is no worse than $\widehat{ATE}_{\text{unadj}}$ when $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$ are defined as coefficients of regressing potential outcomes on a subset of covariates. More specifically, suppose there exists a subset $J \subset \{1, ..., p\}$, such that

$$\boldsymbol{\beta}^{(a)} = ((\boldsymbol{\beta}_J^{(a)})^T, \mathbf{0})^T, \; \boldsymbol{\beta}^{(b)} = ((\boldsymbol{\beta}_J^{(b)})^T, \mathbf{0})^T, \tag{18}$$

where $\boldsymbol{\beta}_J^{(a)}$ and $\boldsymbol{\beta}_J^{(b)}$ are the population level OLS coefficients for regressing the potential outcomes a and b on the covariates in the subset J with intercept, respectively.

Corollary 1 *For $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$ defined in 18 and some λ_a and λ_b, assume conditions 1 through 6 hold. Then the asymptotic variance of $\sqrt{n}$ $\widehat{ATE}_{\text{Lasso}}$ is no greater than that of the $\sqrt{n}$ $\widehat{ATE}_{\text{unadj}}$. The difference is $\frac{1}{p_A(1-p_A)}\Delta$, where*

$$\Delta = -\lim_{n \to \infty} \|X \boldsymbol{\beta}_E\|_2^2 \leq 0, \tag{19}$$

$$\boldsymbol{\beta}_E = (1 - p_A)\boldsymbol{\beta}^{(a)} + p_A \boldsymbol{\beta}^{(b)}. \tag{20}$$

Remark 1. *If we instead assume that the covariates are uniformly bounded, i.e.,* $\max_{i,j}|x_{ij}| \leq L$*, then the fourth moment condition on the error terms, given in* **7***, can be weakened to a second moment condition. Our simulation studies show that the distribution of the unadjusted and the Lasso adjusted estimator may be non-normal when: (1) The covariates are generated from Gaussian distributions and the error terms do not satisfy second moment condition, e.g., being generated from t distribution with one degree of freedom; and (2) The covariates do not have bounded fourth moments, e.g., being generated from t distribution with three degrees of freedom. See the histograms in Figure 1 (the corresponding p-values of Kolmogorov–Smirnov testing for normality are less than* $2.2e-16$*). These findings indicate that our moment conditions cannot be dramatically weakened for asymptotic normality. However, we also find that the Lasso adjusted estimator still has smaller variance and mean squared error than the unadjusted estimator, even when these moment conditions do not hold. In practice, when the covariates do not have bounded fourth moments, one should perform some transformation—e.g., a logarithm transformation—to ensure that the transformed covariates have bounded fourth moments while having a large variance so as to retain useful information. We leave to future work to explore the properties of different transformations.*

Remark 2. *Statement* **26***, typically required in debiasing the Lasso [16], is stronger by a factor of* $\sqrt{\log p}$ *than the usual requirement for* l_1 *consistency of the Lasso.*

Remark 3. *Condition 5 is slightly weaker than the typical restricted eigenvalue condition for analyzing the Lasso.*

Remark 4. *If we assume* $\delta_n = O\left(\sqrt{\frac{\log p}{n}}\right)$ *which satisfies* **25***, then Condition 6 requires that the tuning parameters are proportional to* $\sqrt{\frac{\log p}{n}}$ *which is typically assumed for the Lasso in the high-dimensional linear regression model.*

Remark 5. *For fixed* p*,* $\delta_n = 0$ *in* **9***, Condition 4 holds automatically, and Condition 5 holds when the smallest eigenvalue of* Σ *is uniformly bounded away from 0. In this case, Corollary 1 reverts to Corollary 1.1. in [5]. When these conditions are not satisfied, we should set* λ_a *and* λ_b *to be large enough to cause the Lasso adjusted estimator to revert to the unadjusted one.*

5. Neyman-type conservative variance estimate

We note that the asymptotic variance in Theorem 1 involves the cross-product term $\sigma_{e^{(a)}e^{(b)}}$ which is not consistently estimable in the Neyman model as a_i and b_i are never simultaneously observed. However, we can give a

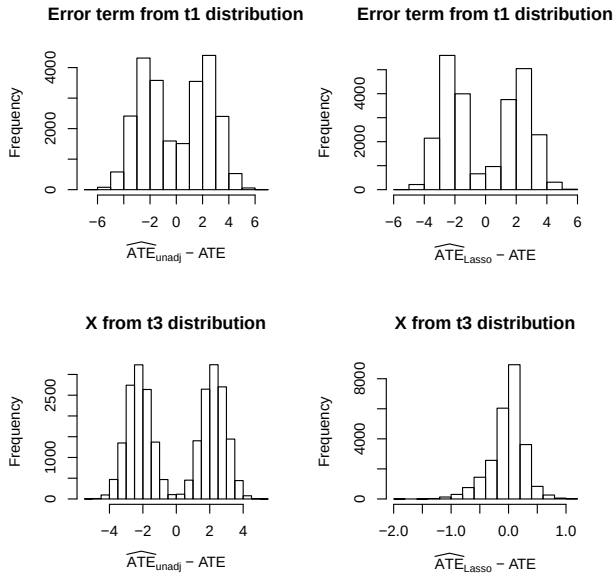

Figure 1: Histograms of the unadjusted estimator and the Lasso adjusted estimator when the moment conditions do not hold. The potential outcomes are simulated from linear regression model and then kept fixed, see more details in the supplemental material. For the upper two subplots, the error terms are generated from t distribution with one degree of freedom and therefore do not satisfy second moment condition; while for the lower two subplots, the covariates are generated from t distribution with there degrees of freedom and thus violate fourth moment condtion.

Neyman-type conservative estimate of the variance which can be used to construct a conservative confidence interval for the ATE. Let

$$\hat{\sigma}^2_{e^{(a)}} = \frac{1}{n_A - df^{(a)}} \sum_{i \in A} \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}}\right)^2, \tag{21}$$

$$\hat{\sigma}^2_{e^{(b)}} = \frac{1}{n_B - df^{(b)}} \sum_{i \in B} \left(b_i - \bar{b}_B - (\mathbf{x}_i - \bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}}\right)^2, \tag{22}$$

where $df^{(a)}$ and $df^{(b)}$ are degrees of freedom defined by

$$df^{(a)} = \hat{s}^{(a)} + 1 = ||\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}}||_0 + 1;$$

$$df^{(b)} = \hat{s}^{(b)} + 1 = ||\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}}||_0 + 1.$$

Condition 7 *For the Gram matrix* Σ *defined in Condition 5, the largest eigenvalue is bounded away from* ∞*, that is, there exists a constant* $\Lambda_{max} < \infty$ *such that*

$$\lambda_{max}(\Sigma) \leq \Lambda_{max}.$$

Define the variance estimate of $\sqrt{n}(\widehat{ATE}_{\text{Lasso}} - ATE)$ as follows:

$$\hat{\sigma}^2_{\text{Lasso}} = \frac{n}{n_A}\hat{\sigma}^2_{e^{(a)}} + \frac{n}{n_B}\hat{\sigma}^2_{e^{(b)}}. \tag{23}$$

Theorem 2 *Assume conditions in Theorem 1 and condition 7 hold. Then $\hat{\sigma}^2_{\text{Lasso}}$ converges in probability to*

$$\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{e^{(a)}} + \frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}},$$

which is greater than or equal to the asymptotic variance of $\sqrt{n}(\widehat{ATE}_{\text{Lasso}} - ATE)$. The difference is

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left[a_i - b_i - ATE - (\mathbf{x}_i - \bar{\mathbf{x}})^T(\boldsymbol{\beta}^{(a)} - \boldsymbol{\beta}^{(b)})\right]^2.$$

Remark 6. *Condition 7 is used to obtain the following bounds for the number of selected covariates by the Lasso: $\max(\hat{s}^{(a)}, \hat{s}^{(b)}) = o_p(\min(n_A, n_B))$. In fact, if we redefine $\hat{\sigma}^2_{e^{(a)}}$ and $\hat{\sigma}^2_{e^{(b)}}$ without adjusted degree of freedom, i.e.,*

$$(\hat{\sigma}^*)^2_{e^{(a)}} = \frac{1}{n_A}\sum_{i\in A}\left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}}\right)^2,$$

$$(\hat{\sigma}^*)^2_{e^{(b)}} = \frac{1}{n_B}\sum_{i\in B}\left(b_i - \bar{b}_B - (\mathbf{x}_i - \bar{\mathbf{x}}_B)^T\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}}\right)^2,$$

and define $(\hat{\sigma}^*)^2_{\text{Lasso}} = \frac{n}{n_A}(\hat{\sigma}^*)^2_{e^{(a)}} + \frac{n}{n_B}(\hat{\sigma}^*)^2_{e^{(b)}}$. It follows from the bounds for $\max(\hat{s}^{(a)}, \hat{s}^{(b)})$ that $(\hat{\sigma}^2_{e^{(a)}}, \hat{\sigma}^2_{e^{(b)}})$ and $((\hat{\sigma}^*)^2_{e^{(a)}}, (\hat{\sigma}^*)^2_{e^{(b)}})$ have the same asymptotic property. Hence, the conclusion of Theorem 2 still holds without Condition 7.

Remark 7. *With the conservative variance estimate in Theorem 2, the Lasso adjusted confidence interval is also valid for the PATE (Population Average Treatment Effect) if there is a super population of size N with $N > n$.*

Theorem 3 *Assume conditions in Theorem 1 hold. Then $(\hat{\sigma}^*)^2_{\text{Lasso}}$ converges in probability to*

$$\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{e^{(a)}} + \frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}}.$$

Remark 8. *Though $(\hat{\sigma}^*)^2_{\text{Lasso}}$ has the same limit as $\hat{\sigma}^2_{\text{Lasso}}$, our simulation experience shows that, for finite sample behaviors, the confidence intervals based on $(\hat{\sigma}^*)^2_{\text{Lasso}}$ may yield low coverage probabilities (e.g., the coverage probability for 95% confidence interval can be only 80%). Hence, we recommend readers to use $\hat{\sigma}^2_{\text{Lasso}}$ in practice.*

6. Related work

The Lasso has already made several appearances in the literature on treatment effect estimation. In the context of observational studies, [16] constructs confidence intervals for preconceived effects or their contrasts by debiasing the Lasso adjusted regression, [17] employs the Lasso as a formal method for selecting adjustment variables via a two-stage procedure which concatenates features from models for treatment and outcome, and similarly, [18] gives very general results for estimating a wide range of treatment effect parameters, including the case of instrumental variables estimation. In addition to the Lasso, [19] considers nonparametric adjustments in the estimation of ATE. The main difference between this prior work and ours is the sampling framework: we focus on a randomized trial, analyzed under the Neyman model with fixed potential outcomes, while previous work uses independent sampling with random error terms.

Our work is related to the estimation of heterogeneous or subgroup-specific treatment effects; including interaction terms to allow the imputed individual-level treatment effects to vary according to some linear combination of covariates. This is pursued in the high-dimensional setting in [20]; this work advocates solving the Lasso on a reduced set of modified covariates, rather than the full set of covariate × treatment interactions, and includes extensions to binary outcomes and survival data. The recent work in [21] considers the problem of designing multiple-testing procedures for detecting subgroup-specific treatment effects; they pose this as an optimization over testing procedures where constraints are added to enforce guarantees on type-I error rate and power to detect effects. Again, the sampling framework in these works is distinct from ours; they do not use the Neyman model as a basis for designing the methods or investigating their properties.

7. PAC data illustration and simulations

We now return to the PAC-man[3] study introduced earlier. We examine the data in more detail and explore the results of several adjustment procedures. There were 1013 patients in the PAC-man study: 506 treated (managed with PAC) and 507 control (managed without PAC, but retaining the option of using alternative devices). The outcome variable is quality-adjusted life years (QALYs). One QALY represents one year of life in full health; in-hospital death corresponds to a QALY of zero. Most of

[3]The PAC-man data is not public. We are allowed access to the data, but we do not have the right to redistribute the data.

the patients in the study died (67%). We have 59 covariates about each individual in the study; we include all main effects as well as 1113 two-way interactions, and form a design matrix $\mathbf{X}$ with 1172 columns and 1013 rows. See Appendix B for more details on the design matrix.

We first check the conditions required for asymptotic normality of the Lasso adjusted estimator; the conditions refer to a sequence of populations, but we will examine them in this specific finite-sample dataset. The crucial ones are the moment condition (Condition 2), the scaling condition (Condition 4) and the cone invertibility of the design matrix condition (Condition 5). In this data set, the covariates are reasonably bounded after standardization ($\max_{i,j}|x_{ij}| \le 11.2$), thus the fourth moment condition on the errors terms can likely be weakened to the second moment condition (see Remark 1). Since we do not know the full set of potential outcomes or the 'active set' S, we cannot precisely estimate these conditions. However, we use the bootstrap to estimate the active set of covariates S and the error terms $e^{(a)}$ and $e^{(b)}$. See Appendix C for more details. Our estimated S contains 21 covariates and the estimated second moments of $e^{(a)}$ and $e^{(b)}$ are 11.8 and 12.0, respectively. The estimated maximal covariance δ_n equals 0.34 and the scaling $(s\log p)/\sqrt{n}$ is 3.55. Condition 5 is computationally infeasible to check; as an approximation, we examine the largest and smallest eigenvalues of the sub-Gram matrix $(1/n)\mathbf{X}_S^T\mathbf{X}_S$, which are 3.23 and 0.04 respectively. Thus the quantity in condition 5 is reasonably bounded away from zero.

We now estimate the ATE using the unadjusted estimator, the Lasso adjusted estimator and the OLS adjusted estimator which is computed based on a sub-design matrix containing only the 59 main effects. We also present results for the two-step estimator $\widehat{ATE}_{\text{Lasso+OLS}}$ which adopts the Lasso to select covariates and then uses OLS to refit the regression coefficients. See [28–31] for statistical properties of Lasso+OLS estimator in linear regression model. Let $\hat{\boldsymbol{\beta}}^{(a)}$ be the Lasso estimator defined in 1 (we omit the subscript "Lasso" for the sake of simplicity) and let $\hat{S}^{(a)} = \{j : \hat{\boldsymbol{\beta}}_j^{(a)} \neq 0\}$ be the support of $\hat{\boldsymbol{\beta}}^{(a)}$. The Lasso+OLS adjustment vector $\hat{\beta}^{(a)}_{\text{Lasso+OLS}}$ for treatment group A is defined by

$$\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso+OLS}} = \underset{\boldsymbol{\beta}:\ \beta_j=0,\ \forall j\notin\hat{S}^{(a)}}{\arg\min} \frac{1}{2n_A}\sum_{i\in A}\left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T\boldsymbol{\beta}\right)^2.$$

We can define the Lasso+OLS adjustment vector $\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso+OLS}}$ for control group B similarly. Then

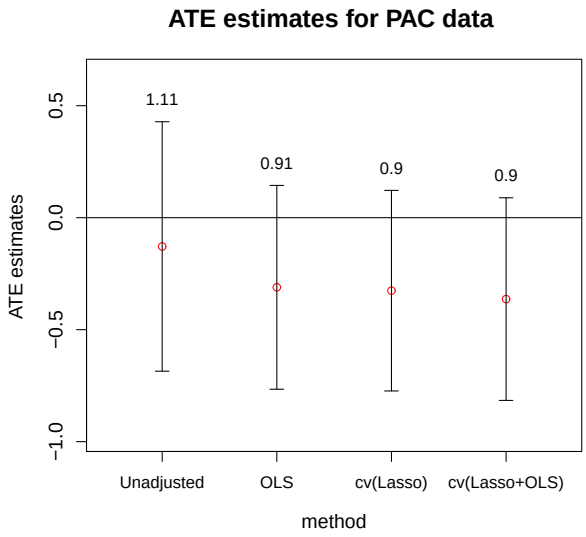

Figure 2: ATE estimates (red circles) and 95% confidence intervals (bars) for the PAC data. The numbers above each bar are the corresponding interval lengths.

$\widehat{ATE}_{\text{Lasso+OLS}}$ is given by

$$\widehat{ATE}_{\text{Lasso+OLS}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A - \bar{\mathbf{x}})^T\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso+OLS}}\right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B - \bar{\mathbf{x}})^T\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso+OLS}}\right].$$

We use the R package "glmnet" to compute the Lasso solution path and select the tuning parameters λ_a and λ_b by 10-fold Cross Validation (CV). To indicate the method of selecting tuning parameters, we denote the corresponding estimators as cv(Lasso) and cv(Lasso+OLS) respectively. We should mention that for the cv(Lasso+OLS) adjusted estimator, we compute the CV error for a given value of λ_a (or λ_b) based on the whole Lasso+OLS estimator instead of the Lasso estimator (see Algorithm 1 in the appendix). Therefore, the cv(Lasso+OLS) and the cv(Lasso) may select different covariates to do the adjustment. This type of cross validation requires more computation than the cross validation based on just the Lasso estimator since it needs to compute the OLS estimator for each fold and each given λ_a (or λ_b), but it can give better prediction and model selection performance.

Figure 2 presents the ATE estimates along with 95% confidence intervals (CI). The interval lengths are shown on top of each interval bar. All the methods give confidence intervals containing 0; hence, this experiment failed to provide sufficient evidence to reject the hypothesis that PAC did not have an effect on patient QALYs (either positive or negative). Since the caretakers of patients managed without PAC retained the option of using less

invasive cardiac output monitoring devices, such an effect may have been particularly hard to detect in this experiment.

However, it is interesting to note that, compared with the unadjusted estimator, the OLS adjusted estimator causes the ATE estimate to decrease (from -0.13 to -0.31), and shortens the confidence interval by about 20% (see Tabel 1). This is due mainly to the imbalance in the pre-treatment probability of death, which was highly predictive of the post-treatment QALYs. The cv(Lasso) adjusted estimator yields a comparable ATE estimate and confidence interval, but the fitted model is more interpretable and parsimonious than the OLS model: it selects 24 and 8 covariates for treated and control, respectively. The cv(Lasso+OLS) estimator selects even fewer covariates: 4 and 5 for treated and control, respectively, but performs a similar adjustment as the cv(Lasso) (see the comparison of fitted values in Figure 3). We also note that these adjustments agree with the one performed in [13], where the treatment effect was adjusted downwards to -0.27 after stratifying into 4 groups based on predicted probability of death.

The covariates selected by Lasso for adjustment are shown in Table 2, where "A·A" denote quadratic term of the covariate A and "A:B" denote two way interaction between two covariates A and B. Among them, patient's age and estimated probability of death (p_death), together with the quadratic term "age·age" and interactions "age:p_death" and "p_death:mech_vent[4]", are the most important covariates for the adjustment. The patients in control group are slightly older and have slightly higher risk of death. These covariates are important predictors of the outcome. Therefore, the unadjusted estimator overestimates the benefits of receiving PAC.

Since not all the potential outcomes are observed, we cannot know the true gains of adjustment methods. However, we can estimate the gains via building a simulated set of potential outcomes by matching treated units to control units on observed covariates. We use the matching method described in [22] which gives 1013 observations with all potential outcomes "known". We match on the 59 main effects. The ATE is -0.29. We then use this synthetic data set to calculate the biases, standard deviations (SD) and root-mean square errors ($\sqrt{\text{MSE}}$) of different ATE estimators based on 25000 replicates of completely randomized experiment which assigns 506 subjects to the treated group and the remainders to the control group.

Table 3 shows the results. For all the methods, the bias is substantially smaller (100 times smaller) than the SD. The SD and $\sqrt{\text{MSE}}$ of the OLS adjusted estimator are both 10.2% smaller than those of the unadjusted estimator, while the cv(Lasso) and cv(Lasso+OLS) adjust-

[4]mechanical ventilation at admission

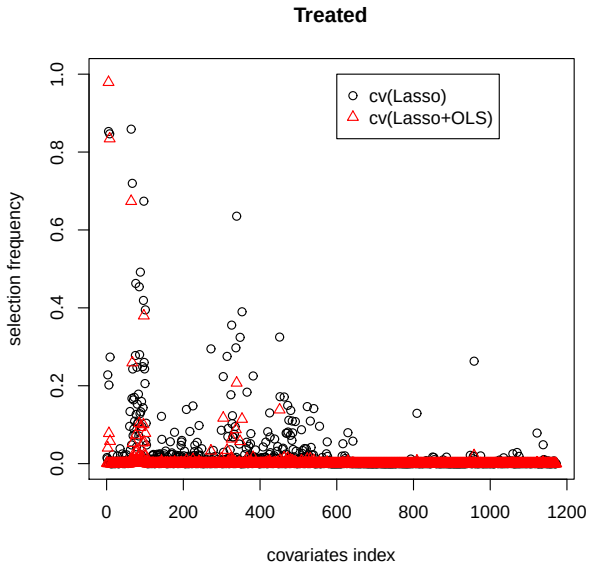

Figure 4: Selection stability comparison of cv(Lasso) and cv(Lasso+OLS) for treated group.

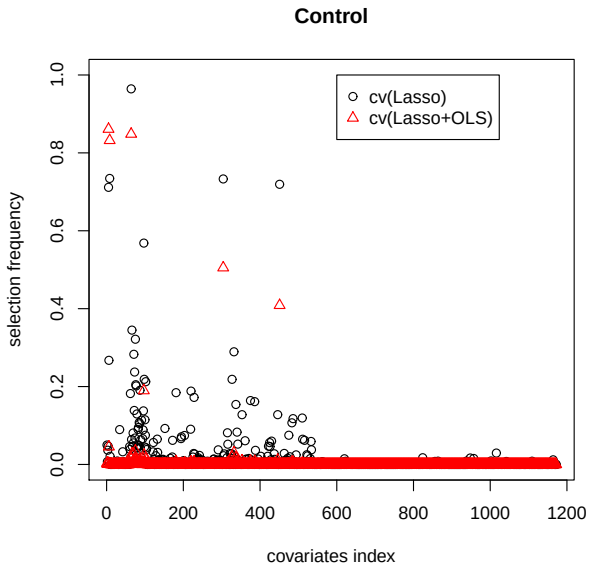

Figure 5: Selection stability comparison of cv(Lasso) and cv(Lasso+OLS) for control group.

Table 1: Statistics for the PAC illustration

Methods	$\widehat{ATE}$	$\hat{\sigma}^2_{\text{ATE}}$	95% confidence interval	No. of selected covariates	
				treated	control
Unadjusted	-0.13	0.081	[-0.69,0.43]	-	-
OLS	-0.31	0.054	[-0.77,0.14]	-	-
cv(Lasso)	-0.33	0.052	[-0.77,0.12]	24	8
cv(Lasso+OLS)	-0.36	0.053	[-0.82,0.09]	4	5

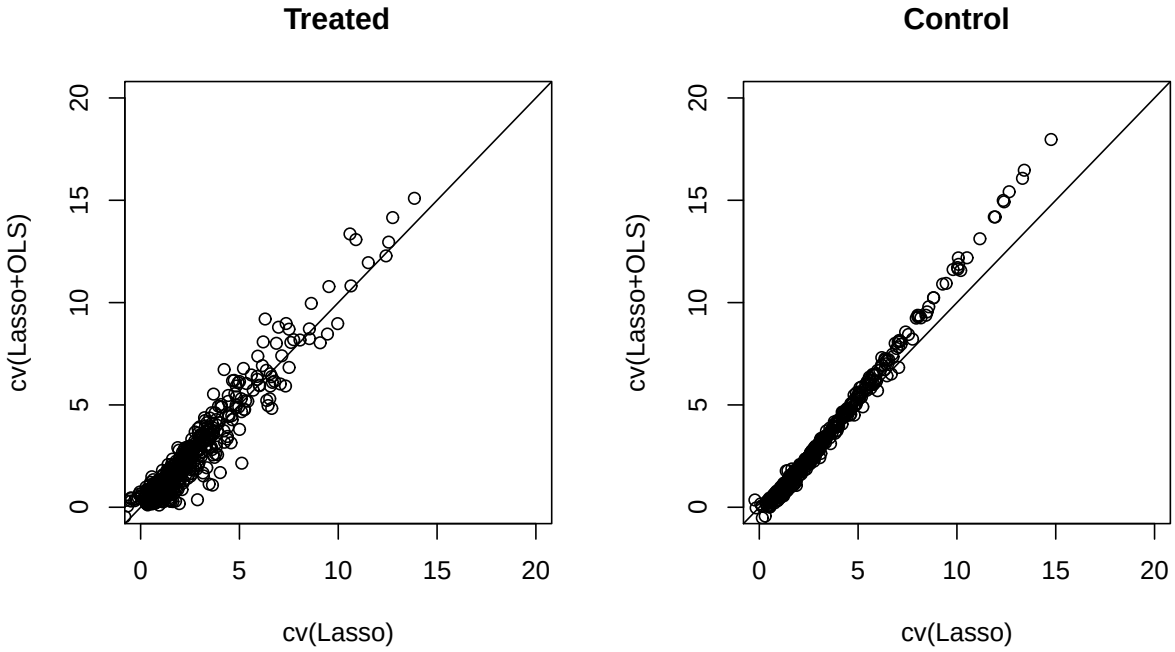

Figure 3: Adjustment (fitted) value comparison for cv(Lasso) and cv(Lasso+OLS).

ed estimators further improve the SD and $\sqrt{\text{MSE}}$ of the OLS adjusted estimator by approximately 4.7%. Moreover, all these methods provide conservative confidence intervals with coverage probabilities higher than 99%. However, the interval lengths of the OLS, cv(Lasso) and cv(Lasso+OLS) adjusted estimator are comparable and are approximately 10% shorter than that of the unadjusted estimator. The cv(Lasso+OLS) adjusted estimator is similar to the cv(Lasso) adjusted estimator in terms of mean squared error, confidence interval length and coverage probability, but outperforms the latter with much fewer and more stable covariates in the adjustment (see Figures 4 and 5 for the selection frequency of each covariate for treatment group and control group respectively).

We conduct additional simulation studies to evaluate the finite sample performance of $\widehat{ATE}_{\text{Lasso}}$ and compare it with that of the OLS adjusted estimator and the unadjusted estimator. A qualitative analysis of these simulations yields the same conclusions as presented above; however, for the sake of brevity, we defer the simulation details in Appendix A.

Acknowledgements

We thank Rebecca Barter for copyediting and suggestions for clarifying the text. This research is partially supported by NSF grants DMS-1106753, DMS-1107000, DMS-1129626, DMS-1209014, CDS&E-MSS 1228246, DMS-1160319 (FRG), AFOSR grant FA9550-14-1-0016, NSA Grant H98230-15-1-0040, ONR grant N00014-15-1-2367, the Center for Science of Information (CSoI), an US NSF Science and Technology Center, under grant agreement CCF-0939370, and the Department of Defense (DoD) through the National Defense Science & Engineering

Table 2: Selected covariates for adjustment

method	treatment	covariates
cv(Lasso+OLS)	treated	age, p_death, age·age, age:p_death
cv(Lasso+OLS)	control	age, p_death, age·age, age:p_death, p_death:mech_vent
cv(Lasso)	treated	pac_rate, age, p_death, age·age, p_death·p_death, region:im_score, region:systemnew, pac_rate:age, pac_rate:p_death, pac_rate:systemnew, im_score:interactnew, age:p_death, age:glasgow, age:systemnew, interactnew:systemnew, pac_rate:creatinine, age:mech_vent, age:respiratory, age:creatinine, interactnew:mech_vent, interactnew:male, glasgow:organ_failure, p_death:mech_vent, systemnew:male
cv(Lasso)	control	age, p_death, age·age, unitsize:p_death, pac_rate:systemnew, age:p_death, interactnew:mech_vent, p_death:mech_vent

Covariate definitions: age (patient's age); p_death (baseline probability of death); mech_vent (mechanical ventilation at admission); region (geographic region); pac_rate (PAC rate in unit); creatinine, respiratory, glasgow, interactnew, organ_failure, systemnew, im_score (various physiological indicators).

Table 3: Statistics for the PAC synthetic data set

						No. of selected covariates	
	Bias	SD	$\sqrt{\text{MSE}}$	Coverage (%)	Length	treated	control
Unadjusted	0.001(0)	0.20(0.02)	0.20(0.02)	99	1.06	-	-
OLS	0.002(0)	0.18(0.02)	0.18(0.02)	99	0.95	-	-
cv(Lasso)	0.001(0)	**0.17(0.02)**	**0.17(0.02)**	99	**0.94**	25(23)	15(14)
cv(Lasso+OLS)	**0.000(0)**	**0.17(0.02)**	**0.17(0.02)**	99	0.95	**6(6)**	**4(3)**

The numbers in parentheses are the corresponding standard errors estimated by using the bootstrap with $B = 500$ resamplings of the ATE estimates.

Graduate Fellowship (NDSEG) Program.

References

[1] Splawa-Neyman J, Dabrowska DM, Speed TP (1990) On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science* 5(4):465–472.

[2] Harvey S et al. (2005) Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (PAC-Man): a randomised controlled trial. *Lancet* 366(9484):472–477.

[3] Dalen JE (2001) The Pulmonary Artery Catheter — Friend, Foe, or Accomplice? *JAMA* 286(3):348.

[4] Connors AF et al. (1996) The effectiveness of right heart catheterization in the initial care of critically III patients. *Jama* 276(11):889–897.

[5] Lin W (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *The Annals of Applied Statistics* 7:295–318.

[6] Permutt T (1990) Testing for imbalance of covariates in controlled experiments. *Statistics in medicine* 9(12):1455–1462.

[7] Tibshirani R (1994) Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society B* 58:267–288.

[8] Bühlmann P, Van De Geer S (2011) *Statistics for high-dimensional data: methods, theory and applications.* (Springer Science & Business Media).

[9] Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies.

[10] Freedman DA (2008) On regression adjustments to experimental data. *Advances in Applied Mathematics* 40(2):180–193.

[11] Freedman DA (2008) On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics* 2(1):176–196.

[12] Freedman DA (2008) Randomization does not justify logistic regression. *Statistical Science* 23(2):237–249.

[13] Miratrix LW, Sekhon JS, Yu B (2013) Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 75:369–396.

[14] Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81:945–960.

[15] Pitkin E et al. (2013) Improved Precision in Estimating Average Treatment Effects. *arXiv preprint arXiv:1311.0291*.

[16] Zhang CH, Zhang SS (2014) 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.

[17] Belloni A, Chernozhukov V, Hansen C (2013) Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies* 81(2):608–650.

[18] Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2013) Program evaluation with high-dimensional data. *arXiv preprint arXiv:1311.2645*.

[19] Li L, Tchetgen Tchetgen E, van der Vaart A, Robins JM (2011) Higher order inference on a treatment effect under low regularity conditions. *Statistics & probability letters* 81(7):821–828.

[20] Tian L, Alizadeh A, Gentles A, Tibshirani R (2014) A simple method for detecting interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* just accep.

[21] Rosenblum M, Liu H, En-Hsu Y (2014) Optimal Tests of Treatment Effects for the Overall Population and Two Subpopulations in Randomized Trials, Using Sparse Linear Programming. *Journal of the American Statistical Association* 109(507):1216–1228.

[22] Diamond A, Sekhon JS (2013) Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *Review of Economics and Statistics* 95(3):932–945.

[23] Meinshausen N (2007) Relaxed Lasso. *Computational Statistics and Data Analysis* 52:374C393.

[24] Efron B, Hastie T and Tibshirani R (2004) Least angle regression. *Annals of Statistics* 32:407-499.

[25] Belloni A, Chernozhukov V (2009) Least Squares After Model Selection in High-dimensional Sparse Models. *Bernoulli* 19:521-547.

[26] Liu H, Yu B (2013) Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 7:3124C3169.

[27] Massart P (1986) in *Geometrical and Statistical Aspects of Probability in Banach Spaces*. (Springer), pp. 73–109.

[28] Meinshausen N (2007) Relaxed Lasso. *Computational Statistics and Data Analysis* 52:374C393.

[29] Efron B, Hastie T and Tibshirani R (2004) Least angle regression. *Annals of Statistics* 32:407-499.

[30] Belloni A, Chernozhukov V (2009) Least Squares After Model Selection in High-dimensional Sparse Models. *Bernoulli* 19:521-547.

[31] Liu H, Yu B (2013) Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 7:3124C3169.

[32] Massart P (1986) in *Geometrical and Statistical Aspects of Probability in Banach Spaces*. (Springer), pp. 73–109.

[33] Freedman DA (2008) On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics* 2(1):176–196.

A. Simulation

In this section we carry out simulation studies to evaluate the finite sample performance of the $\widehat{ATE}_{\text{Lasso}}$ estimator. We also present results for the $\widehat{ATE}_{\text{OLS}}$ estimator when $p < n$ and the two-step estimator $\widehat{ATE}_{\text{Lasso+OLS}}$.

Again, we use the R package "glmnet" to compute the Lasso solution path and select the tuning parameters λ_a and λ_b by 10-fold Cross Validation (CV) and denote the corresponding adjusted estimators as cv(Lasso) and cv(Lasso+OLS) respectively.

The potential outcomes a_i and b_i are generated from the following nonlinear model: for $i = 1, ..., n$,

$$a_i = \sum_{j=1}^{s} x_{ij}\beta_j^{(a1)} + \exp\left(\sum_{j=1}^{s} x_{ij}\beta_j^{(a2)}\right) + \epsilon_i^{(a)},$$

$$b_i = \sum_{j=1}^{s} x_{ij}\beta_j^{(b1)} + \exp\left(\sum_{j=1}^{s} x_{ij}\beta_j^{(b2)}\right) + \epsilon_i^{(b)},$$

where $\epsilon_i^{(a)}$ and $\epsilon_i^{(b)}$ are independent error terms. We set $n = 250$, $s = 10$, $p = 50$ and 500. For $p = 50$, we compute the OLS estimator and compare it with the Lasso. The vector of covariates $\mathbf{x}_i$ is generated from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$. We consider two different Toeplitz covariance matrices Σ which control the correlation among the covariates:

$$\Sigma_{ii} = 1;\ \Sigma_{ij} = \rho^{|i-j|}\ \forall i \neq j,$$

where $\rho = 0, 0.6$. The true coefficients $\beta_j^{(a1)}$, $\beta_j^{(a2)}$, $\beta_j^{(b1)}$, $\beta_j^{(b2)}$ are generated independently according to

$$\beta_j^{(a1)} \sim t_3; \quad \beta_j^{(a2)} \sim 0.1 * t_3, \quad j = 1, ..., s,$$

$$\beta_j^{(b1)} \sim \beta_j^{(a1)} + t_3; \quad \beta_j^{(b2)} \sim \beta_j^{(a2)} + 0.1 * t_3, \quad j = 1, ..., s,$$

where t_3 denotes the t distribution with three degrees of freedom. This ensures that the treatment effects are not not constant across individuals, and that the linear model does not hold in this simulation. The error terms $\epsilon_i^{(a)}$ and $\epsilon_i^{(b)}$ are generated according to the following linear model with some hidden covariates $\mathbf{z}_i$:

$$\epsilon_i^{(a)} = \sum_{j=1}^{s} z_{ij}\beta_j^{(a1)} + \tilde{\epsilon}_i^{(a)},$$

$$\epsilon_i^{(b)} = \sum_{j=1}^{s} z_{ij}\beta_j^{(b1)} + \tilde{\epsilon}_i^{(b)},$$

where $\tilde{\epsilon}_i^{(a)}$ and $\tilde{\epsilon}_i^{(b)}$ are drawn independently from standard normal distribution. The vector $\mathbf{z}_i$ is independent of $\mathbf{x}_i$ and also drawn independently from the multivariate normal distribution $\mathcal{N}(0, \Sigma)$. The values of $\mathbf{x}_i$, $\beta^{(a1)}$, $\beta^{(a2)}$, $\beta^{(b1)}$, $\beta^{(b2)}$, $\mathbf{z}_i$, $\tilde{\epsilon}_i^{(a)}$, $\tilde{\epsilon}_i^{(b)}$, a_i and b_i are generated once and then kept fixed.

After the potential outcomes are generated, a completely randomized experiment is simulated 25000 times, assigning $n_A = 100, 125, 150$ subjects to treatment A and the remainder to control B. There are 12 different combinations of (p, ρ, n_A) in total.

Figures 9, 10, 11 show boxplots of different ATE estimators with their standard deviations (computed from 25000 replicates of randomized experiments) presented on top of each box. Regardless of whether the design is balanced ($n_A = 125$) or not ($n_A = 100, 150$), the regression based estimators have much smaller variances than the unadjusted estimator and therefore improve the estimation precision.

To further compare the performance of these estimators, we present the bias, the standard deviation (SD) and the root-mean square error ($\sqrt{\text{MSE}}$) of the estimates in Table 4. Bias is reported as the absolute difference from the true treatment effect. We find that the bias of each method is substantially smaller (more than 10 times smaller) than the SD. The cv(Lasso) and cv(Lasso+OLS) adjusted estimators perform similar in terms of SD and $\sqrt{\text{MSE}}$: reducing those of the OLS adjusted estimator and the unadjusted estimator by $10\% - 15\%$ and $15\% - 31\%$ respectively. We also compare the number of covariates selected by cv(Lasso) and cv(Lasso+OLS) for the treatment group and control group separately; see Table 5. It is easy to see that the cv(Lasso+OLS) adjusted estimator uses many fewer (more than 44%) covariates in the adjustment to obtain similar improvement of SD and $\sqrt{\text{MSE}}$ of ATE estimate as the cv(Lasso) adjusted estimator. Moreover, we find that the covariates selected by the cv(Lasso+OLS) are more stable across different realizations of treatment assignment than the covariates selected by the cv(Lasso). Because of this, we would rank the performance of the estimators in decreasing order as follows: cv(Lasso+OLS) adjusted, the cv(Lasso) adjusted, the OLS adjusted, and unadjusted estimators.

We move now to study the finite sample performance of the Neyman-type conservative variance estimates. For each simulation example and each one of the 25000 completely randomized experiments, we calculate the ATE estimates ($\widehat{ATE}$) and the variance estimates ($\hat{\sigma}$) and then form the 95% confidence intervals $[\widehat{ATE} - 1.96 \cdot \hat{\sigma}/\sqrt{n}, \widehat{ATE} + 1.96 \cdot \hat{\sigma}/\sqrt{n}]$. Figures 6, 7, 8 present boxplots of the interval length with the coverage probability noted on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators. More results are showed in Table 6. We find that all the confidence intervals for the unadjusted estimator are conservative. The cv(Lasso) adjusted and the cv(Lasso+OLS0 adjusted estimators perform very similar: although their coverage probability (at least 92%) may be slightly less than the pre-assigned confidence level (95%), their mean interval length is much shorter ($26\% - 37\%$) than that of the unadjusted estimator. The OLS adjusted estimator has comparable interval length as the cv(Lasso) and cv(Lasso+OLS) adjusted estimator, but has slightly worse coverage probability ($90\% - 93\%$).

To further investigate how good the Neyman standard deviation (SD) estimate is, we compare them in Figure 12 with the "true" SD presented in Table 4 (the SD of the ATE estimates over 25000 randomized experiments). We find that Neyman SD estimate is very conservative for the unadjusted estimator (its mean is $5\% - 14\%$ larger than the "true" SD); however, for the OLS adjusted estimator, the mean of Neyman SD estimate can be $6\% - 100\%$ smaller than the "true" SD; this may be due to over-fitting. For the cv(Lasso) and cv(Lasso+OLS) adjusted estimator, the mean of Neyman SD estimator is within $1 \pm 7\%$ of the "true" SD. Thus, although the Neyman variance esti-

mate is asymptotically conservative, it can be too narrow in finite samples for the regression-based adjusted estimator. However, if we increase the sample size n to 1000, almost all the confidence intervals are conservative.

We conduct more simulation examples to evaluate the conditions assumed for asymptotic normality of the Lasso adjusted estimator. We use the same simulation setup as above, but for simplicity, we generate the potential outcomes from linear a model (set $\beta^{(a2)} = \beta^{(b2)} = 0$) and remove the effects of the hidden covariates z_i in generating the error terms $\epsilon_i^{(a)}$ and $\epsilon_i^{(b)}$ and set $\rho = 0,\ n_A = 125$. We find that the distribution of the cv(Lasso) adjusted estimator may be non-normal when:

(1) The covariates are generated from Gaussian distribution and the error terms do not satisfy second moment condition, e.g., being generated from t distribution with one degree of freedom, see the upper two subplots of Figure 1 (in the main text) for the histograms of unadjusted the cv(Lasso) adjusted estimators (the corresponding p-values of Kolmogorov–Smirnov testing for normality are less than $2.2e{-}16$).

(2) The covariates do not have bounded fourth moments, e.g., being generated from t distribution with three degrees of freedom, see the lower two subplots of Figure 1 (in the main text) for the histograms of unadjusted the cv(Lasso) adjusted estimators (again, the corresponding p-values of Kolmogorov–Smirnov testing for normality are less than $2.2e-16$).

These findings indicate that our moment condition (Condition 2 and Remark 1) cannot be dramatically weakened. However, we also find that the cv(Lasso) adjusted estimator still has smaller SD and $\sqrt{\text{MSE}}$ than the unadjusted estimator even when these moment conditions do not hold.

B. The design matrix of the PAC data

In the PAC data, there are 59 covariates (main effects), 50 of which are binary-valued indicators, which may be correlated with the outcomes. One of the main effects (called interactnew) has heavy tail, so we transform it with the function: $x \to \log(|x| + 1)$ to make it look normally distributed. We then centralize and standardize the non-indicator covariates. The quadratic terms (9 in total) of non-indicator covariates and two-way interactions between main effects (1711 in total) may also help in modeling the potential outcomes, so we included them in the design matrix. The quadratic terms and the interactions between non-indicator covariates are also centered and standardized. Some of the interactions are identical to other effects and we only retain one of them. We also remove the interactions which are highly correlated with the main effects (correlation larger than 0.95) and remove the indicators with very sparse entries (where the number of 1's is less than 20). In the end, we form a design matrix X with 1172 columns (covariates) and 1013 rows (subjects).

C. Checking conditions

Let $S^{(a)} = \{j : \boldsymbol{\beta}_j^{(a)} \neq 0\}$ and $S^{(b)} = \{j : \boldsymbol{\beta}_j^{(b)} \neq 0\}$ denote the sets of relevant covariates for treatment group and control group respectively. Denote $S = S^{(a)} \bigcup S^{(b)} = \{j : \boldsymbol{\beta}_j^{(a)} \neq 0 \text{ or } \boldsymbol{\beta}_j^{(b)} \neq 0\}$. We use bootstrap to get an estimation of the set of relevant covariates $S^{(a)}, S^{(b)}$. The approximation errors $e^{(a)}$ and $e^{(b)}$ are estimated by regressing the observed potential outcomes a and b on the covariates in S respectively. We only present how to estimate $S^{(a)}$ and $e^{(a)}$ in detail; the estimation of $S^{(b)}$ and $e^{(b)}$ is similar.

Let A, B be the set of treated subjects (using PAC) and control subjects (without using PAC) respectively. Denote $a_i, i \in A$ the potential outcomes (quality-adjusted life years (QALYs)) under treatment and $x_i \in \mathcal{R}^{1172}$ the covariate vector of the ith subject. For each $d = 1, ..., 1000$, we draw a bootstrap sample $\{(a_i^*(b), x_i^*(b)) : i \in A\}$ with replacement from the data points $\{(a_i, x_i) : i \in A\}$. Then we compute the cv(Lasso+OLS) adjusted vector $\hat{\boldsymbol{\beta}}(d)$ based on each bootstrap sample $\{(a_i^*(b), x_i^*(b)) : i \in A\}$. Let τ_j be the selection fraction of non-zero $\hat{\beta}_j(b)$ in the 1000 bootstrap estimators, i.e., $\tau_j = (1/1000)\sum_{b=1}^{1000} \mathbb{I}_{\{\hat{\beta}_j(b)\neq 0\}}$, where $\mathbb{I}$ is the indicator function. We form the set of relevant covariates $S^{(a)}$ by including the covariates whose selection fraction is larger than 0.5: $S^{(a)} = \{j : \tau_j > 0.5\}$.

To estimate the approximation error $e^{(a)}$, we regress a_i on the relevant covariates $x_{ij}, j \in S^{(a)}$ and compute OLS estimate and the corresponding residual. That is, let $T^{(a)}$ denote the complement set of $S^{(a)}$,

$$\boldsymbol{\beta}_{\text{OLS}}^{(a)} = \underset{\boldsymbol{\beta}:\ \beta_j=0,\ \forall j\in T^{(a)}}{\arg\min} \frac{1}{2n_A}\sum_{i\in A}\left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T\boldsymbol{\beta}\right)^2.$$

$$e_i^{(a)} = a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T\boldsymbol{\beta}_{\text{OLS}}^{(a)},\ i \in A.$$

The maximal covariance δ_n is estimated as:

$$\max\left\{\max_j\left|\frac{1}{n_A}\sum_{i\in A}(x_{ij} - (\bar{\mathbf{x}})_j)\left(e_i^{(a)} - \bar{e}_A^{(a)}\right)\right|,\right.$$
$$\left.\max_j\left|\frac{1}{n_B}\sum_{i\in B}(x_{ij} - (\bar{\mathbf{x}})_j)\left(e_i^{(b)} - \bar{e}_B^{(b)}\right)\right|\right\}.$$

D. Proofs of Theorem 1, 2, 3 and Corollary 1

In this section, we will prove Theorem 1 - 3 and Corollary 1 under weaker sparsity conditions.

Definition 3 *We define an approximate sparsity measure. Given the regularization parameter λ_a, λ_b and $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$, the sparsity measures for treatment and control groups, $s^{(a)}_{\lambda_a}$ and $s^{(b)}_{\lambda_b}$ are defined as*

$$s^{(a)}_{\lambda_a} = \sum_{j=1}^{p} \min\left\{\frac{|\beta^{(a)}_j|}{\lambda_a}, 1\right\}, \; s^{(b)}_{\lambda_b} = \sum_{j=1}^{p} \min\left\{\frac{|\beta^{(b)}_j|}{\lambda_b}, 1\right\}, \tag{24}$$

respectively. We will allow $s^{(a)}_{\lambda_a}$ and $s^{(b)}_{\lambda_b}$ to grow with n, though the notation does not explicitly show this. Note that this is weaker than strict sparsity, as it allows $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$ to have many small non-zero entries.

Condition (*). Suppose there exist $\boldsymbol{\beta}^{(a)}$, $\boldsymbol{\beta}^{(b)}$, λ_a and λ_b such that the conditions 1, 2, 3 and the following statements 1, 2, 3 hold simultaneously.

- **Statement 1**. Decay and scaling. Let $s_\lambda = \max\left\{s^{(a)}_{\lambda_a}, s^{(b)}_{\lambda_b}\right\}$,

$$\delta_n = o\left(\frac{1}{s_\lambda \sqrt{\log p}}\right), \tag{25}$$

$$(s_\lambda \log p)/\sqrt{n} = o(1). \tag{26}$$

- **Statement 2**. Cone invertibility factor. Define the Gram matrix as $\Sigma = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$: There exist constants $C > 0$ and $\xi > 1$ not depending on n, such that

$$\|\mathbf{h}_S\|_1 \le C s_\lambda \|\Sigma \mathbf{h}\|_\infty, \; \forall \mathbf{h} \in \mathcal{C}, \tag{27}$$

with $\mathcal{C} = \{\mathbf{h} : \|\mathbf{h}_{S^c}\|_1 \le \xi \|\mathbf{h}_S\|_1\}$, and

$$S = \{j : |\beta^{(a)}_j| > \lambda_a \text{ or } |\beta^{(b)}_j| > \lambda_b\}. \tag{28}$$

- **Statement 3**. Let $\tau = \min\left\{1/70, (3p_A)^2/70, (3-3p_A)^2/70\right\}$. For constants $0 < \eta < \frac{\xi-1}{\xi+1}$ and $0 < M < \infty$, assume the regularization parameters of the Lasso belong to the sets

$$\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}} + \delta_n\right), \tag{29}$$

$$\lambda_b \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_B}\sqrt{\frac{2\log p}{n}} + \delta_n\right). \tag{30}$$

It is easy to verify that Condition (*) is implied by conditions 1 - 6. In the following, we will prove Theorem 1 - 3 and Corollary 1 under the weaker Condition (*). For ease of notation, we will omit the subscript of $\hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}}$, $\hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}}$, s_λ, $s^{(a)}_{\lambda_a}$ and $s^{(b)}_{\lambda_b}$ from now on. Moreover, we can assume, without loss of generality, that

$$\bar{a} = 0, \; \bar{b} = 0, \; \bar{\mathbf{x}} = \mathbf{0}. \tag{31}$$

Otherwise, we can consider $\breve{a}_i = a_i - \bar{a}$, $\breve{b}_i = b_i - \bar{b}$ and $\breve{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$. Then, ATE $= \bar{a} - \bar{b} = 0$ and the definition of $\widehat{ATE}_{\text{Lasso}}$ becomes

$$\widehat{ATE}_{\text{Lasso}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}\right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)}\right]. \tag{32}$$

We will rely heavily on the following Massart concentration inequality for sampling without replacement.

Lemma 1 *Let $\{z_i, i = 1, ..., n\}$ be a finite population of real numbers. Let $A \subset \{i, \ldots, n\}$ be a subset of deterministic size $|A| = n_A$ that is selected randomly without replacement. Define $p_A = n_A/n$, $\sigma^2 = n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2$. Then, for any $t > 0$,*

$$P\left(\bar{z}_A - \bar{z} \ge t\right) \le \exp\left\{-\frac{p_A n_A t^2}{(1+\tau)^2\sigma^2}\right\}, \tag{33}$$

with $\tau = \min\left\{1/70, (3p_A)^2/70, (3-3p_A)^2/70\right\}$.

Remark. *Massart showed in his paper [32] that for sampling without replacement, the following concentration inequality holds:*

$$P\left(\bar{z}_A - \bar{z} \ge t\right) \le \exp\left\{-\frac{p_A n_A t^2}{\sigma^2}\right\}.$$

His proof required that n/n_A must be an integer. We extend the proof to allow n/n_A to be a non-integer but with a slightly larger constant factor $(1+\tau)^2$.

Proof. Assume $\bar{z} = 0$ without loss of generality. For $n_A \le n/2$, let $m \ge 2$ and $r \ge 0$ be integers satisfying $n - n_A m = r < n_A$. Let $u \ge 0$. We first prove that

$$\begin{aligned} &E\exp\left(u\sum_{i\in A} z_i\right) \\ &\le E\exp\left(u\delta\sum_{i\in B} z_i/\{m(m+1)\} + u^2 n\sigma^2/4\right) \end{aligned} \tag{34}$$

for a random subset $B \subset \{1, \ldots, n\}$ of fixed size $|B| \le n/2$ and a certain fixed $\delta \in \{-1, 1\}$. Let P_1 be the probability under which $\{i_1, \ldots, i_n\}$ is a random permutation of $\{1, \ldots, n\}$. Given $\{i_1, \ldots, i_n\}$, we divide the sequence into consecutive blocks $B_1, \ldots, B_{n_A}$ with $|B_j| = m + 1$ for

$j = 1, \ldots, r$ and $|B_j| = m$ for $j = r+1, ..., n_A$. Let $\bar{z}_k$ be the mean of $\{z_i : i \in B_k\}$ and P_2 be a probability conditionally on $\{i_1, \ldots, i_n\}$ under which w_k is a random element of $\{z_i : i \in B_k\}$, $k = 1, \ldots, n_A$. Then $\{w_1, \ldots, w_{n_A}\}$ is a random sample from $\{z_1, \ldots, z_n\}$ without replacement under $P = P_1 P_2$. Let $\Delta_k = \max_{i \in B_k} z_i - \min_{i \in B_k} z_i$ and denote E_2 the expectation under P_2. The Hoeffding inequality gives

$$E_2 \exp\left(u \sum_{k=1}^{n_A} w_k\right) \leq \exp\left(u \sum_{k=1}^{n_A} \bar{z}_k + (u^2/8) \sum_{k=1}^{n_A} \Delta_k^2\right). \tag{35}$$

As $\Delta_i^2 \leq 2 \sum_{i \in B_k} (z_i - \bar{z}_k)^2 \leq 2 \sum_{i \in B_k} z_i^2$,

$$E_2 \exp\left(u \sum_{k=1}^{n_A} w_k\right) \leq \exp\left(u \sum_{k=1}^{n_A} \bar{z}_k + u^2 n \sigma^2 / 4\right) \tag{36}$$

Let $B = \cup_{k=1}^{r} B_k$. As $\bar{z} = 0$,

$$\sum_{k=1}^{n_A} \bar{z}_k = \sum_{i \in B} z_i / \{m(m+1)\}. \tag{37}$$

This yields **34** with $\delta = 1$ when $|B| \leq n/2$. Otherwise, **34** holds with $\delta = -1$ when B is replaced by B^c, as $\sum_{i \in B} z_i = -\sum_{i \in B^c} z_i$ due to $\bar{z} = 0$.

Now, as $m(m+1) \geq 6$, repeated application of **34** yields

$$\begin{aligned}
& E \exp\left(u \sum_{i \in A} z_i\right) \\
\leq\ & E \exp\left[u\delta' \sum_{i \in B'} z_i / \{m(m+1)m'(m'+1)\}\right. \\
& \left. + \left(1 + \{m(m+1)\}^{-2}\right) u^2 n \sigma^2 / 4\right] \\
\leq\ & \exp\left[\left(1 + \{m(m+1)\}^{-2}(1 + 1/36 + 1/36^2 + \right.\right. \\
& \left.\left. \cdots)\right) u^2 n \sigma^2 / 4\right] \\
=\ & \exp\left[\left(1 + (36/35)\{m(m+1)\}^{-2}\right) u^2 n \sigma^2 / 4\right] \\
\leq\ & \exp\left[(1+\tau)^2 u^2 n \sigma^2 / 4\right]
\end{aligned} \tag{38}$$

with $\tau = (18/35)\{m(m+1)\}^{-2}$. The upper bound for τ follows from $2 \leq m < n/n_A < m+1$.

As $\bar{z} = 0$, we also have

$$E \exp\left(u \sum_{i \in A} z_i\right) \leq \exp\left[(1+\tau)^2 u^2 n \sigma^2 / 4\right] \tag{39}$$

for $n_A > n/2$. This yields **33** via the usual

$$\begin{aligned}
& P\{\bar{z}_A - \bar{z} > t\} \\
\leq\ & \exp\left[-ut + (1+\tau)^2 u^2 n \sigma^2 / (4 n_A^2)\right] \\
=\ & \exp\left[-2 \frac{p_A n_A t^2}{(1+\tau)^2 \sigma^2} + \frac{p_A n_A t^2}{(1+\tau)^2 \sigma^2}\right]
\end{aligned} \tag{40}$$

with $u = 2 p_A n_A t / \{\sigma(1+\tau)\}^2$.

D.1. Proof of Theorem 1

Proof. Recall the decompositions of the potential outcomes:

$$a_i = \bar{a} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(a)} + e_i^{(a)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(a)} + e_i^{(a)}, \tag{41}$$

$$b_i = \bar{b} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(b)} + e_i^{(b)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(b)} + e_i^{(b)}. \tag{42}$$

If we define $\mathbf{h}^{(a)} = \hat{\boldsymbol{\beta}}^{(a)} - \boldsymbol{\beta}^{(a)}$, $\mathbf{h}^{(b)} = \hat{\boldsymbol{\beta}}^{(b)} - \boldsymbol{\beta}^{(b)}$, by substitution, we have

$$\begin{aligned}
& \sqrt{n}(\widehat{ATE}_{\text{Lasso}} - ATE) \\
&= \underbrace{\sqrt{n}\left[\bar{e}_A^{(a)} - \bar{e}_B^{(b)}\right]}_{*} - \underbrace{\sqrt{n}\left[(\bar{\mathbf{x}}_A)^T \mathbf{h}^{(a)} - (\bar{\mathbf{x}}_B)^T \mathbf{h}^{(b)}\right]}_{**}.
\end{aligned}$$

We will analyze these two terms separately, showing that $(*)$ is asymptotically normal with mean 0 and variance given by **17**, and that $(**)$ is $o_p(1)$.

Asymptotic normality of $(*)$ follows from the Theorem 1 in [33] with a and b replaced by $e^{(a)}$ and $e^{(b)}$ respectively. To bound $(**)$, we will apply Hölder inequality to each of the terms. We will focus on the term involving the treatment group A, but exact same analysis is applied to the control group B. We have the bound

$$\left|(\bar{\mathbf{x}}_A)^T \mathbf{h}^{(a)}\right| \leq \|\bar{\mathbf{x}}_A\|_\infty \|\mathbf{h}^{(a)}\|_1. \tag{43}$$

We will bound the two terms on the right hand side of **43** by the following Lemma 2 and Lemma 3, respectively.

Lemma 2 *Under the moment condition of [6], if we let* $c_n = \frac{(1+\tau)L^{1/4}}{p_A}\sqrt{\frac{2\log p}{n}}$, *then as* $n \to \infty$,

$$P\left(\|\bar{\mathbf{x}}_A\|_\infty > c_n\right) \to 0.$$

Thus, $\|\bar{\mathbf{x}}_A\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$.

Lemma 3 *Assume the conditions of Theorem 1 hold. Then* $\|\mathbf{h}^{(a)}\|_1 = o_p\left(\frac{1}{\sqrt{\log p}}\right)$.

The proofs of these two Lemmas are below. Using these two Lemmas, it is easy to show that $(**) = \sqrt{n} \cdot O_p\left(\sqrt{\frac{\log p}{n}}\right) \cdot o_p\left(\frac{1}{\sqrt{\log p}}\right) = o_p(1)$.

D.2. Proof of Corollary 1

Proof. By Theorem 1 in [33], the asymptotic variance of $\sqrt{n}\ \widehat{ATE}_{\text{unadj}}$ is $\frac{1-p_A}{p_A}\lim_{n\to\infty}\sigma_a^2 + \frac{p_A}{1-p_A}\lim_{n\to\infty}\sigma_b^2 + 2\lim_{n\to\infty}\sigma_{ab}$, so the difference is

$$\begin{aligned}
& \frac{1-p_A}{p_A}\lim_{n\to\infty}\left(\sigma_{e^{(a)}}^2 - \sigma_a^2\right) + \frac{p_A}{1-p_A}\lim_{n\to\infty}\left(\sigma_{e^{(b)}}^2 - \sigma_b^2\right) \\
& + 2\lim_{n\to\infty}\left(\sigma_{e^{(a)}e^{(b)}} - \sigma_{ab}\right).
\end{aligned}$$

We will analyze these three terms separately. Since $X\boldsymbol{\beta}^{(a)}$ and $X\boldsymbol{\beta}^{(b)}$ are the orthogonal projections of a and b onto the same subspace, we have

$$(X\boldsymbol{\beta}^{(a)})^T e^{(a)} = (X\boldsymbol{\beta}^{(a)})^T e^{(b)}$$

$$= (X\boldsymbol{\beta}^{(b)})^T e^{(a)} = (X\boldsymbol{\beta}^{(b)})^T e^{(b)} = 0.$$

Simple calculations yield

$$\sigma^2_{e^{(a)}} - \sigma^2_a = ||e^{(a)}||_2^2 - ||a||_2^2 = -||X\boldsymbol{\beta}^{(a)}||_2^2,$$

$$\sigma^2_{e^{(b)}} - \sigma^2_b = ||e^{(b)}||_2^2 - ||b||_2^2 = -||X\boldsymbol{\beta}^{(b)}||_2^2,$$

$$\sigma_{e^{(a)}e^{(b)}} - \sigma_{ab} = (e^{(a)})^T(e^{(b)}) - a^T b = -(X\boldsymbol{\beta}^{(a)})^T(X\boldsymbol{\beta}^{(b)})$$

Combining the above three equalities, we obtain the corollary.

D.3. Proof of Theorem 2

Proof. To prove Theorem 2, it is enough to show that

$$\hat{\sigma}^2_{e^{(a)}} \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(a)}}, \tag{44}$$

$$\hat{\sigma}^2_{e^{(b)}} \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(b)}}. \tag{45}$$

We will only prove the statement 44 and omit the proof of the statement 45 since it is very similar.

We first state the following two lemmas. Lemma 4 bounds the number of selected covariates (covariates with a nonzero coefficient), while Lemma 5 establishes conditions under which the subsample mean (without replacement) converges in probability to the population mean.

Lemma 4 *Under the conditions in Theorem 2, there exists a constant C, such that the following holds with probability going to 1:*

$$\hat{s}^{(a)} \le Cs; \quad \hat{s}^{(b)} \le Cs. \tag{46}$$

The proof of Lemma 4 can be found below.

Lemma 5 *Let $\{z_i, i = 1, ..., n\}$ be a finite population of real numbers. Let $A \subset \{i, \ldots, n\}$ be a subset of deterministic size $|A| = n_A$ that is selected randomly without replacement. Suppose that the population mean of the z_i has a finite limit and that there exist constants $\epsilon > 0$ and $L < \infty$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} |z_i|^{1+\epsilon} \le L. \tag{47}$$

If $\frac{n_A}{n} \to p_A \in (0, 1)$, then

$$\bar{z}_A \xrightarrow{p} \lim_{n\to\infty} \bar{z}. \tag{48}$$

By definition 21 and simple calculations,

$$\begin{aligned}
&\hat{\sigma}^2_{e^{(a)}} \\
=\ & \frac{1}{n_A - df^{(a)}} \sum_{i\in A} \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}\right)^2 \\
=\ & \frac{1}{n_A - df^{(a)}} \sum_{i\in A} \Big(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \boldsymbol{\beta}^{(a)} \\
& + (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)} - \hat{\boldsymbol{\beta}}^{(a)})\Big)^2 \\
=\ & \frac{1}{n_A - df^{(a)}} \sum_{i\in A} \Big(a_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(a)} - (\bar{a}_A - (\bar{\mathbf{x}}_A)^T \boldsymbol{\beta}^{(a)}) \\
& + (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)} - \hat{\boldsymbol{\beta}}^{(a)})\Big)^2 \\
=\ & \frac{n_A}{n_A - df^{(a)}} \frac{1}{n_A} \sum_{i\in A} \Big(e_i^{(a)} - \bar{e}_A^{(a)} + (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \\
& (\boldsymbol{\beta}^{(a)} - \hat{\boldsymbol{\beta}}^{(a)})\Big)^2 \\
=\ & \frac{n_A}{n_A - df^{(a)}} \left\{ \frac{1}{n_A} \sum_{i\in A} \left(e_i^{(a)} - \bar{e}_A^{(a)}\right)^2 + \frac{1}{n_A} \sum_{i\in A} \right. \\
& \left. \left((\mathbf{x}_i - \bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)} - \hat{\boldsymbol{\beta}}^{(a)})\right)^2 \right\} + \frac{n_A}{n_A - df^{(a)}} \\
& \left\{ \frac{1}{n_A} \sum_{i\in A} (e_i^{(a)} - \bar{e}_A^{(a)})(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)} - \hat{\boldsymbol{\beta}}^{(a)}) \right\}.
\end{aligned}$$

The second to last equality is due to the decomposition of potential outcome a:

$$a_i = \mathbf{x}_i^T \boldsymbol{\beta}^{(a)} + e_i^{(a)}; \quad \bar{a}_A = (\bar{\mathbf{x}}_A)^T \boldsymbol{\beta}^{(a)} + \bar{e}_A^{(a)}.$$

It is easy to see that

$$\frac{1}{n_A} \sum_{i\in A} \left(e_i^{(a)} - \bar{e}_A^{(a)}\right)^2 = \frac{1}{n_A} \sum_{i\in A} (e_i^{(a)})^2 - (\bar{e}_A^{(a)})^2. \tag{49}$$

By the $4th$ moment condition on the approximation error $e^{(a)}$ (see 7), and applying Lemma 5 we get

$$\frac{1}{n_A} \sum_{i\in A} (e_i^{(a)})^2 \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(a)}}; \quad \bar{e}_A^{(a)} \xrightarrow{p} \lim_{n\to\infty} \bar{e}^{(a)} = 0.$$

Therefore,

$$\frac{1}{n_A} \sum_{i\in A} \left(e_i^{(a)} - \bar{e}_A^{(a)}\right)^2 \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(a)}}. \tag{50}$$

Simple algebra operations give

$$\begin{aligned}
&\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\\
=\;&(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})^T\left[\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\right]\\
&(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\\
\le\;&||\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)}||_1^2\cdot||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T||_\infty.
\end{aligned}\tag{51}$$

We next show that 51 converges to 0 in probability. By Lemma 3 and Lemma 7, we have

$$||\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)}||_1=\|\mathbf{h}^{(a)}\|_1=o_p\left(\frac{1}{\sqrt{\log p}}\right),\tag{52}$$

$$||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T||_\infty=O_p(1).\tag{53}$$

Therefore,

$$\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\xrightarrow{p}0.\tag{54}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\frac{1}{n_A}\sum_{i\in A}(e_i^{(a)}-\bar{e}_A^{(a)})(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})|\\
\le\;&\left[\frac{1}{n_A}\sum_{i\in A}\left(e_i^{(a)}-\bar{e}_A^{(a)}\right)^2\right]^{\frac{1}{2}}\cdot\\
&\left[\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\right]^{\frac{1}{2}}
\end{aligned}\tag{55}$$

which converges to 0 in probability because of 50 and 54.

By Lemma 4 and Condition 4, we have

$$\frac{n_A}{n_A-df^{(a)}}=\frac{n_A}{n_A-\hat{s}^{(a)}-1}\xrightarrow{p}1.\tag{56}$$

Combining 50, 54, 55 and 56, we conclude that

$$\hat{\sigma}^2_{e^{(a)}}\xrightarrow{p}\lim_{n\to\infty}\sigma^2_{e^{(a)}}.$$

The remaining part of the proof is to study the difference between the conservative variance estimate and the true asymptotic variance:

$$\begin{aligned}
&\left(\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}}\right)-\left(\frac{1-p_A}{p_A}\right.\\
&\left.\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\frac{p_A}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}}+2\lim_{n\to\infty}\sigma_{e^{(a)}e^{(b)}}\right)\\
=\;&\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\lim_{n\to\infty}\sigma^2_{e^{(b)}}-2\lim_{n\to\infty}\sigma_{e^{(a)}e^{(b)}}\\
=\;&\lim_{n\to\infty}\sigma^2_{e^{(a)}-e^{(b)}}\\
=\;&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left(a_i-b_i-\mathbf{x}_i^T(\boldsymbol{\beta}^{(a)}-\boldsymbol{\beta}^{(b)})\right)^2.
\end{aligned}\tag{57}$$

D.4. Proof of Theorem 3

Proof. By Lemma 4, $\max(\hat{s}^{(a)},\hat{s}^{(b)})=o_p(\min(n_A,n_B))$. Therefore, $(\hat{\sigma}^2_{e^{(a)}},\hat{\sigma}^2_{e^{(b)}})$ and $((\hat{\sigma}^*)^2_{e^{(a)}},(\hat{\sigma}^*)^2_{e^{(b)}})$ have the same limits. The conclusion follows from Theorem 2.

E. Proofs of Lemmas

In this section, we will drop the superscript on $\mathbf{h}$, e and $\hat{\boldsymbol{\beta}}$ and focus on the proof for treatment group A, as the same analysis can be applied to control group B.

E.1. Proof of Lemma 2

Proof. Let $c_n=\frac{(1+\tau)L^{1/4}}{p_A}\sqrt{\frac{2\log p}{n}}$. By the union bound,

$$\begin{aligned}
P\left(\|\bar{\mathbf{x}}_A\|_\infty>c_n\right)&=P\left(\max_{j=1,\dots,p}\left|\frac{1}{n_A}\sum_{i\in A}x_{ij}\right|>c_n\right)\\
&\le\sum_{j=1}^pP\left(\left|\frac{1}{n_A}\sum_{i\in A}x_{ij}\right|>c_n\right).
\end{aligned}\tag{58}$$

By Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\sum_{i=1}^nx_{ij}^2\le\left(\frac{1}{n}\sum_{i=1}^nx_{ij}^4\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^n1^2\right)^{\frac{1}{2}}\le\sqrt{L}.\tag{59}$$

Substituting the concentration inequality 33 into 58,

$$\begin{aligned}
P\left(\|\bar{\mathbf{x}}_A\|_\infty>c_n\right)&\le2\exp\left\{\log p-\frac{p_An_Ac_n^2}{(1+\tau)^2L^{1/2}}\right\}\\
&=2\exp\{-\log p\}\to0.
\end{aligned}$$

E.2. Proof of Lemma 3

Proof. We start with the KKT condition, which characterizes the solution to the Lasso. Recall the definition of the Lasso estimator $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}\frac{1}{2n_A}\sum_{i\in A}\left(a_i-\bar{a}_A-(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\boldsymbol{\beta}\right)^2+\lambda_a\|\boldsymbol{\beta}\|_1.$$

The KKT condition for $\hat{\boldsymbol{\beta}}$ is

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)\left(a_i-\bar{a}_A-(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\hat{\boldsymbol{\beta}}\right)=\lambda_a\kappa, \quad (60)$$

where κ is the subgradient of $||\boldsymbol{\beta}||_1$ taking value at $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$, i.e.,

$$\kappa\in\partial||\boldsymbol{\beta}||_1\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \quad \text{with} \quad \begin{cases}\kappa_j\in[-1,1] \text{ for } j \text{ s.t. } \hat{\beta}_j=0\\ \kappa_j=\text{sign}(\hat{\beta}_j) \text{ otherwise}\end{cases} \quad (61)$$

Substituting a_i by the decomposition 3, 60 becomes

$$\begin{aligned}&\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\\&+\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)=\lambda_a\kappa.\end{aligned} \quad (62)$$

Multiplying both sides of 62 by $-\mathbf{h}^T=(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T$, we have

$$\begin{aligned}&\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right)^2-\mathbf{h}^T\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)\\&=\lambda_a(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T\kappa\le\lambda_a\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1\right)\end{aligned}$$

where the last inequality holds because

$$\boldsymbol{\beta}^T\kappa\le||\boldsymbol{\beta}||_1||\kappa||_\infty\le||\boldsymbol{\beta}||_1 \quad \text{and} \quad \hat{\boldsymbol{\beta}}^T\kappa=||\hat{\boldsymbol{\beta}}||_1.$$

Rearranging, and applying Hölder's inequality, we have

$$\begin{aligned}&\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right)^2\\&\le\lambda_a\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1\right)+\mathbf{h}^T\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)\\&\le\lambda_a\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1\right)+\|\mathbf{h}\|_1\underbrace{\left\|\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)\right\|_\infty}_{*}\end{aligned}$$

To control the term $(*)$, we define the event $\mathcal{L}=\{*\le\eta\lambda_a\}$. The following Lemma 6 shows that, with λ_a defined appropriately, $\mathcal{L}$ holds with probability approaching 1. We will prove this Lemma later.

Lemma 6 *Define*
$\mathcal{L}=\left\{\left\|\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)\right\|_\infty\le\eta\lambda_a\right\}$.
Then under the conditions of Theorem 1, $P(\mathcal{L})\to 1$.

On $\mathcal{L}$

$$\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right)^2\le\lambda_a\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1+\eta\|\mathbf{h}\|_1\right). \quad (63)$$

By substituting the defiition of $\mathbf{h}$, and several applications of the triangle inequality, we have

$$\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1\le\|\mathbf{h}_S\|_1-\|\mathbf{h}_{S^c}\|_1+2\|\boldsymbol{\beta}_{S^c}\|_1.$$

Therefore,

$$\begin{aligned}&\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right)^2\\&\le\lambda_a\left(\|\mathbf{h}_S\|_1-\|\mathbf{h}_{S^c}\|_1+2\|\boldsymbol{\beta}_{S^c}\|_1+\eta\|\mathbf{h}\|_1\right)\\&\le\lambda_a\left((\eta-1)\|\mathbf{h}_{S^c}\|_1+(1+\eta)\|\mathbf{h}_S\|_1+2\|\boldsymbol{\beta}_{S^c}\|_1\right).\end{aligned}$$

Because $\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right)^2\ge 0$, we obtain

$$\begin{aligned}&(1-\eta)\|\mathbf{h}_{S^c}\|_1\\&\le(1+\eta)\|\mathbf{h}_S\|_1+2\|\boldsymbol{\beta}_{S^c}\|_1\le(1+\eta)\|\mathbf{h}_S\|_1+2s\lambda_a.\end{aligned} \quad (64)$$

where the last inequality holds because of the definition of s in 24 and S in 28.

Consider the following two cases:

(I) If $(1+\eta)\|\mathbf{h}_S\|_1+2s\lambda_a\ge(1-\eta)\xi\|\mathbf{h}_S\|_1$ then by 64,

$$\begin{aligned}\|\mathbf{h}\|_1&=\|\mathbf{h}_S\|_1+\|\mathbf{h}_{S^c}\|_1\\&\le\left(\frac{1+\eta}{1-\eta}+1\right)\|\mathbf{h}_S\|_1+\frac{2s\lambda_a}{1-\eta}\\&\le\frac{2s\lambda_a}{1-\eta}\left(\frac{2}{(1-\eta)\xi-(1+\eta)}+1\right).\end{aligned}$$

By the definition of λ_a and the scaling assumptions 25, 26, we have that $s\lambda_a=o\left(\frac{1}{\sqrt{\log p}}\right)$.

(II) If $(1+\eta)\|\mathbf{h}_S\|_1+2s\lambda_a<(1-\eta)\xi\|\mathbf{h}_S\|_1$ then by 64 we have $\|\mathbf{h}_{S^c}\|_1\le\xi\|\mathbf{h}_S\|_1$. Applying the cone invertibility condition on the design matrix 27,

$$\begin{aligned}\|\mathbf{h}\|_1&=\|\mathbf{h}_S\|_1+\|\mathbf{h}_{S^c}\|_1\\&\le(1+\xi)\|\mathbf{h}_S\|_1\le(1+\xi)Cs\left\|\frac{1}{n}X^TX\mathbf{h}\right\|_\infty\end{aligned} \quad (65)$$

Before applying this inequality we will revisit the KKT condition 61, but this time we will take the l_∞-norm, yielding

$$\begin{aligned}&\left\|\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{h}\right\|_\infty\\&\le\lambda_a+\left\|\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(e_i-\bar{e}_A)\right\|_\infty\le(1+\eta)\lambda_a,\end{aligned} \quad (66)$$

where the latter inequality holds on the set $\mathcal{L}$. The final step is to control the deviation of the subsampled covariance matrix from the population covariance matrix, so

that we can apply 65. We define another event with constant $C_1 = \frac{2(1+\tau)L^{1/2}}{p_A}$

$$\mathcal{M} = \left\{ \left\| \frac{1}{n_A} \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T - \frac{1}{n} X^T X \right\|_\infty \leq C_1 \sqrt{\frac{\log p}{n}} \right\}$$

Lemma 7 *Assume stability of treatment assignment probability condition 1 and moment condition 6 hold. Then $P(\mathcal{M}) \to 1$.*

We will prove Lemma 7 later. Continuing our inequalities, on the event $\mathcal{L} \cap \mathcal{M}$,

$$\begin{aligned} & s \left\| \frac{1}{n} X^T X \mathbf{h} \right\|_\infty \\ \leq & C_1 s \sqrt{\frac{\log p}{n}} \|\mathbf{h}\|_1 + s \left\| \frac{1}{n_A} \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \mathbf{h} \right\|_\infty \\ \leq & o(1) \|\mathbf{h}\|_1 + s(1+\eta)\lambda_a, \end{aligned}$$

where we have applied the scaling assumption 26 and 66 in the second line. Hence, by 65,

$$\|\mathbf{h}\|_1 \leq (1+\xi) C \left[o(1) \|\mathbf{h}\|_1 + s(1+\eta)\lambda_a \right].$$

Again, applying the scaling assumptions 25 and 26, we get $\|\mathbf{h}\|_1 = o_p\left(\frac{1}{\sqrt{\log p}}\right)$.

E.3. Proof of Lemma 4

Proof. In the proof of Lemma 3, we have shown that, on $\mathcal{L}$ defined in Lemma 6,

$$\begin{aligned} & \frac{1}{n_A} \sum_{i \in A} \left((\mathbf{x}_i - \bar{\mathbf{x}}_A)^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right)^2 \qquad (67) \\ \leq & \lambda_a \left(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1 + \eta ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1 \right). \\ \leq & \lambda_a (1+\eta) ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1. \qquad (68) \end{aligned}$$

Let $\mathbf{x}^j$ be the j-th column of the design matrix X and $\bar{\mathbf{x}}_A^j = n_A^{-1} \sum_{i \in A} x_{ij}$. Again, by KKT conditon, we have

$$\left| \frac{1}{n_A} \sum_{i \in A} (x_{ij} - \bar{\mathbf{x}}_A^j) \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}} \right) \right| = \lambda_a,$$

$$\text{if } \hat{\boldsymbol{\beta}}_j \neq 0.$$

Substituting a_i by the decomposition 3 yields

$$\begin{aligned} & \left| \frac{1}{n_A} \sum_{i \in A} (x_{ij} - \bar{\mathbf{x}}_A^j)(e_i - \bar{e}_A) + \frac{1}{n_A} \sum_{i \in A} (x_{ij} - \bar{\mathbf{x}}_A^j) \right. \\ & \left. (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right| = \lambda_a. \end{aligned}$$

Combining with the definition of the event $\mathcal{L}$, we have if $\hat{\boldsymbol{\beta}}_j \neq 0$

$$\Delta_j := \left| \frac{1}{n_A} \sum_{i \in A} (x_{ij} - \bar{\mathbf{x}}_A^j)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right| \geq (1-\eta)\lambda_a. \qquad (69)$$

Let $Z = (\mathbf{z}_1, ..., \mathbf{z}_n) \in R^{p \times n}$ with $\mathbf{z}_i = \mathbf{x}_i - \bar{\mathbf{x}}_A \in R^p$ and denote $\mathbf{w} = Z^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$, then

$$\begin{aligned} \frac{1}{n_A} ||\mathbf{w}_A||_2^2 &= \frac{1}{n_A} \sum_{i \in A} \left((\mathbf{x}_i - \bar{\mathbf{x}}_A)^T (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right)^2 \\ &\leq \lambda_a (1+\eta) ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1. \end{aligned}$$

Let $Z_A = (\mathbf{z}_i : i \in A)$; since the largest eigenvalues of $Z_A^T Z_A$ and $Z_A Z_A^T$ are the same,

$$\begin{aligned} & \frac{1}{n_A^2} \mathbf{w}_A^T Z_A^T Z_A \mathbf{w}_A \\ \leq & \frac{1}{n_A^2} \lambda_{\max}(Z_A^T Z_A) ||\mathbf{w}_A||_2^2 \\ \leq & \frac{1}{n_A} \lambda_{\max}(Z_A Z_A^T) \lambda_a (\eta + 1) ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1 \\ \leq & \Lambda_{\max} \frac{n}{n_A} \lambda_a (1+\eta) ||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1. \end{aligned}$$

The last inequality holds because

$$\begin{aligned} & \lambda_{\max}(Z_A Z_A^T) \\ = & \max_{\mathbf{u}: ||\mathbf{u}||_2 = 1} \mathbf{u}^T Z_A Z_A^T \mathbf{u} \\ = & \max_{\mathbf{u}: ||\mathbf{u}||_2 = 1} \mathbf{u}^T \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \mathbf{u} \\ = & \max_{\mathbf{u}: ||\mathbf{u}||_2 = 1} \mathbf{u}^T \sum_{i \in A} \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} - n_A \mathbf{u}^T (\bar{\mathbf{x}}_A)(\bar{\mathbf{x}}_A)^T \mathbf{u} \\ \leq & \max_{\mathbf{u}: ||\mathbf{u}||_2 = 1} \mathbf{u}^T \sum_{i \in A} \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} \leq n \Lambda_{\max}. \qquad (70) \end{aligned}$$

On the other hand,

$$\frac{1}{n_A^2} \mathbf{w}_A^T Z_A^T Z_A \mathbf{w}_A = \sum_{j=1}^p \Delta_j^2 \geq \sum_{j: \hat{\beta}_j \neq 0} \Delta_j^2 \geq (1-\eta)^2 \lambda_a^2 \hat{s}. \qquad (71)$$

Combining 69, 71 and the fact that with probability going to 1 (see the proof of Lemma 3)

$$||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||_1 \leq C s (1+\eta) \lambda_a,$$

where C is a constant, we conclude that with probability going to 1,

$$\begin{aligned} \hat{s} &\leq \frac{1}{(1-\eta)^2} \frac{1}{\lambda_a^2} \Lambda_{\max} \frac{n}{n_A} \lambda_a (1+\eta) C s (1+\eta) \lambda_a \\ &\leq \frac{C(1+\eta)^2}{p_A (1-\eta)^2} s. \end{aligned}$$

E.4. Proof of Lemma 5

Proof. For any $t > 0$, we have

$$P(|\bar{z}_A - \lim_{n\to\infty} \bar{z}| > t) \le P(|\bar{z}_A - \bar{z}| > t/2) + P(|\bar{z} - \lim_{n\to\infty} \bar{z}| > t/2). \tag{72}$$

The second term in the right hand side of 72 obviously converges to 0 as $n \to \infty$. To bound the first term, we apply the concentration inequality 33. By 47, it is easy to show

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n z_i^2 &= \frac{1}{n}\sum_{i=1}^n |z_i|^{1-\epsilon}|z_i|^{1+\epsilon} \\ &\le (nL)^{\frac{1-\epsilon}{1+\epsilon}} \frac{1}{n}\sum_{i=1}^n |z_i|^{1+\epsilon} \le L^{\frac{2}{1+\epsilon}} n^{\frac{1-\epsilon}{1+\epsilon}}.\end{aligned}$$

Concentration inequality 33 yields

$$P(|\bar{z}_A - \bar{z}| > t/2) \le 2\exp\left\{-\frac{p_A n_A t^2}{4(1+\tau)^2 L^{\frac{2}{1+\epsilon}} n^{\frac{1-\epsilon}{1+\epsilon}}}\right\} \to 0.$$

E.5. Proof of Lemma 6

Proof. It is easy to verify that

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A) = \frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - (\bar{\mathbf{x}}_A)(\bar{e}_A).$$

Hence,

$$\begin{aligned}&||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A)||_\infty \\ &\le ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty + ||(\bar{\mathbf{x}}_A)(\bar{e}_A)||_\infty.\end{aligned} \tag{73}$$

We analyze these two terms on the right hand side of the inequality separately. For the first term, by the triangle inequality and the definition of δ_n in 9,

$$||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty \tag{74}$$

$$\le ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i||_\infty + ||\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i||_\infty$$

$$\le ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i||_\infty + \delta_n. \tag{75}$$

We will again bound 74 by the concentration inequality 33 in Lemma 1. By the Cauchy-Schwarz inequality, we have for any $j = 1, .., p$,

$$\frac{1}{n}\sum_{i=1}^n x_{ij}^2 e_i^2 \le \left(\frac{1}{n}\sum_{i=1}^n x_{ij}^4\right)^{\frac{1}{2}} \left(\frac{1}{n}\sum_{i=1}^n e_i^4\right)^{\frac{1}{2}} \le L.$$

Let $t_n = \frac{(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}}$, then by the union bound and the concentration inequality 33,

$$\begin{aligned}&P\left(||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i e_i||_\infty > t_n\right) \\ &\le 2\exp\left\{\log p - \frac{p_A n_A t_n^2}{(1+\tau)^2 L)}\right\} \\ &= 2\exp\{-\log p\} \to 0.\end{aligned}$$

Taking this back to 74, we have

$$P\left(||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty \le t_n + \delta_n\right) \to 1. \tag{76}$$

For the second term, by Lemma 2, we have shown that,

$$P\left(\|\bar{\mathbf{x}}_A\|_\infty \le \frac{(1+\tau)L^{1/4}}{p_A}\sqrt{\frac{2\log p}{n}}\right) \to 1.$$

A similar proof yields

$$P\left(\|\bar{e}_A\|_\infty \le \frac{(1+\tau)L^{1/4}}{p_A}\sqrt{\frac{2\log p}{n}}\right) \to 1.$$

Hence, under the scaling condition 26,

$$P\left(\|(\bar{\mathbf{x}}_A)(\bar{e}_A)\|_\infty \le \frac{(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}}\right) \to 1. \tag{77}$$

Combining 76 and 77 yields

$$\begin{aligned}P\Bigg(&||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A)||_\infty \\ &\le \frac{2(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}} + \delta_n\Bigg) \to 1.\end{aligned}$$

The conclusion follows from the condition $\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}} + \delta_n\right)$.

E.6. Proof of Lemma 7

Proof. It is easy to see that

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T = \frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T - (\bar{\mathbf{x}}_A)(\bar{\mathbf{x}}_A)^T.$$

Then, by triangle inequality,

$$||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T - \frac{1}{n}X^T X||_\infty \tag{78}$$

$$\le \underbrace{||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T - \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T||_\infty}_{*} + \underbrace{||(\bar{\mathbf{x}}_A)(\bar{\mathbf{x}}_A)^T||_\infty}_{**}. \tag{79}$$

We control the first term (∗) again using the concentration inequality **33** and the union bound. By the wayCauchy-Schwarz inequality, for $j, k = 1, ..., p$,

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 x_{ik}^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} x_{ij}^4\right)^{\frac{1}{2}} \left(\frac{1}{n}\sum_{i=1}^{n} x_{ik}^4\right)^{\frac{1}{2}} \leq L.$$

Then,

$$\begin{aligned} & P\left(||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T||_\infty \geq \frac{(1+\tau)L^{1/2}}{p_A}\right. \\ & \qquad \left.\sqrt{\frac{3\log p}{n}}\right) \\ \leq \;& 2\exp\left\{2\log p - \frac{3p_A n_A(1+\tau)^2 L\log p}{(1+\tau)^2 L p_A^2 n}\right\} \\ = \;& 2\exp\{-\log p\} \to 0. \end{aligned} \tag{80}$$

The second term (∗∗) is bounded by again observing that, by Lemma 2 and the scaling condition **26**,

$$(**) \leq ||\bar{\mathbf{x}}_A||_\infty^2 = o_p(\sqrt{\frac{\log p}{n}}). \tag{81}$$

Combining **80** and **81** yields the conclusion.

F. Tables and Figures

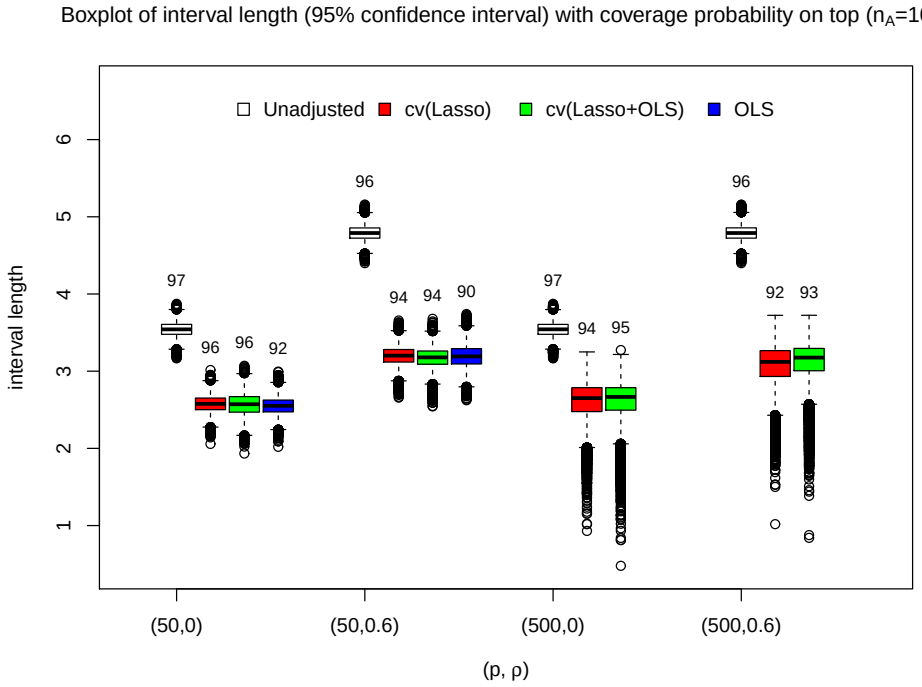

Figure 6: Boxplot of the interval length with coverage probability (%) on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 100$.

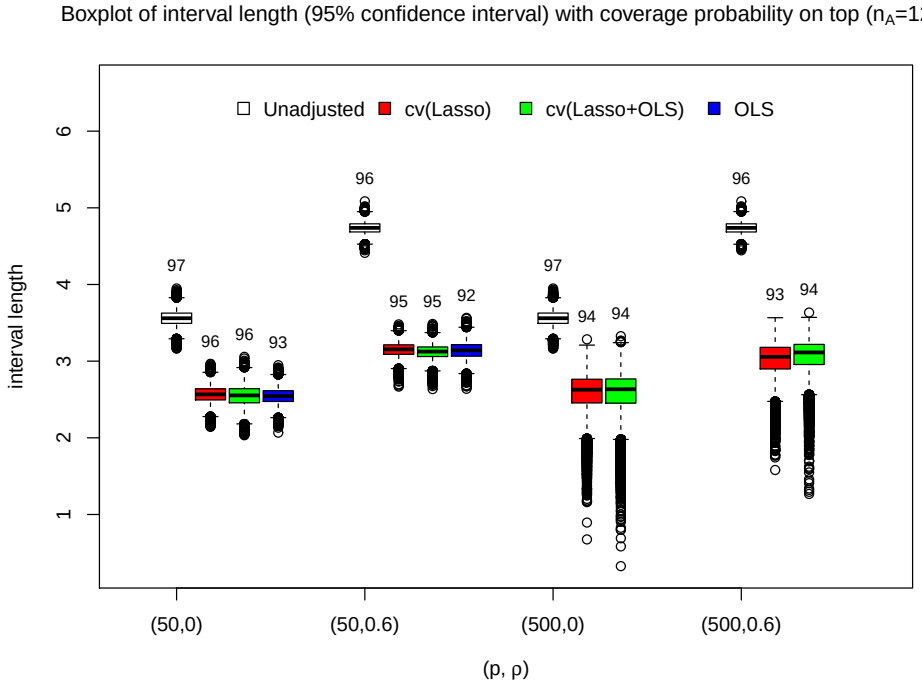

Figure 7: Boxplot of the interval length with coverage probability (%) on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 125$.

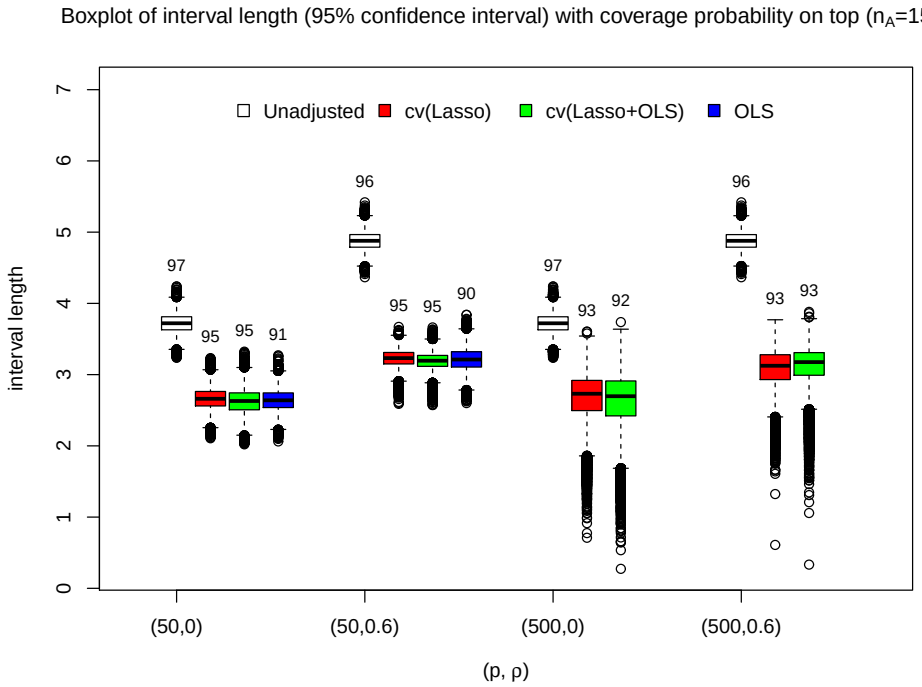

Figure 8: Boxplot of the interval length with coverage probability (%) on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 150$.

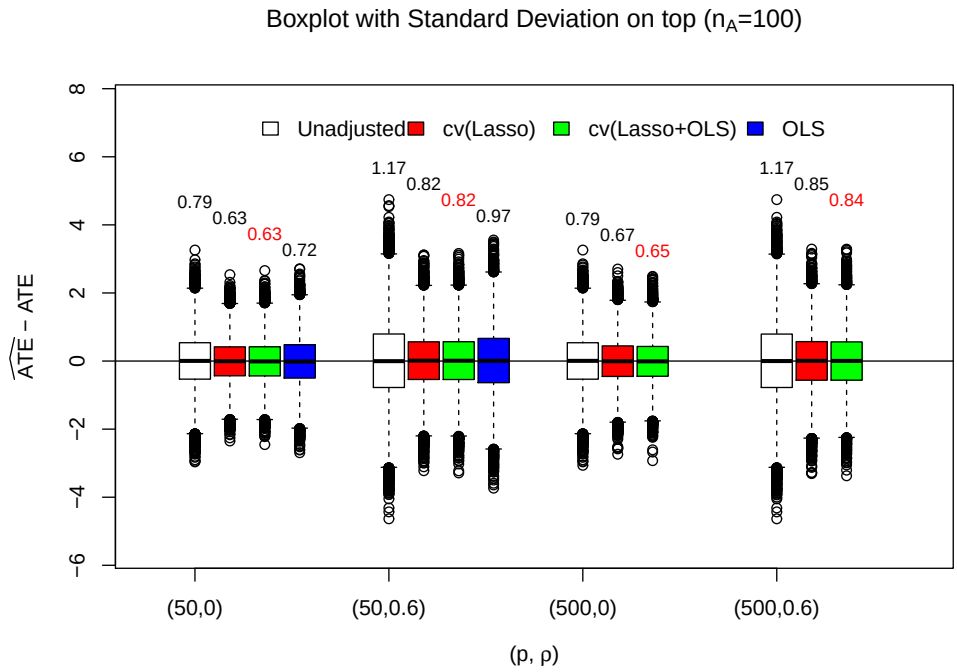

Figure 9: Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 100$.

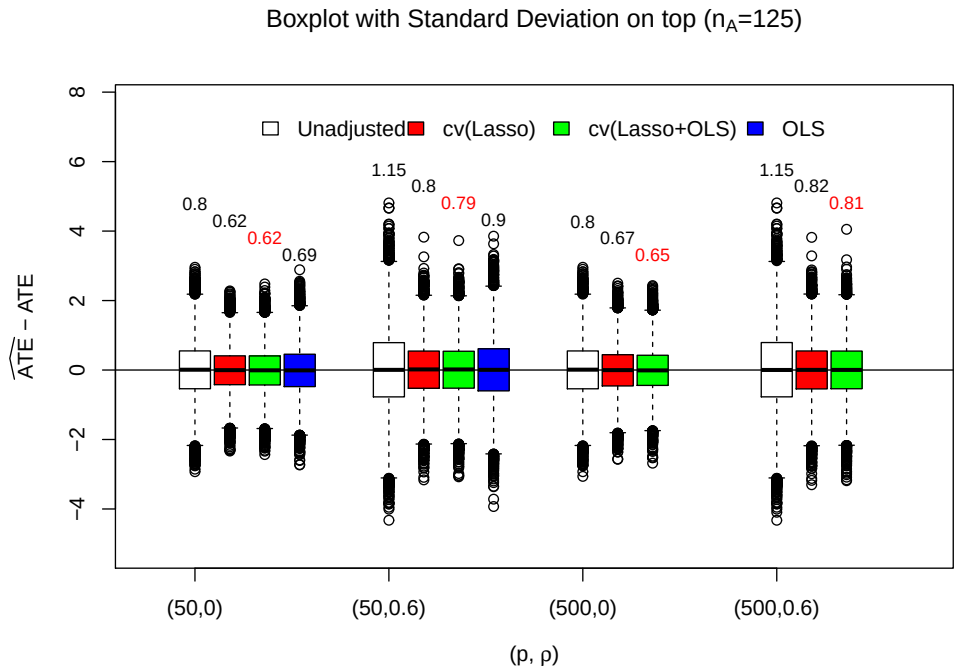

Figure 10: Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 125$.

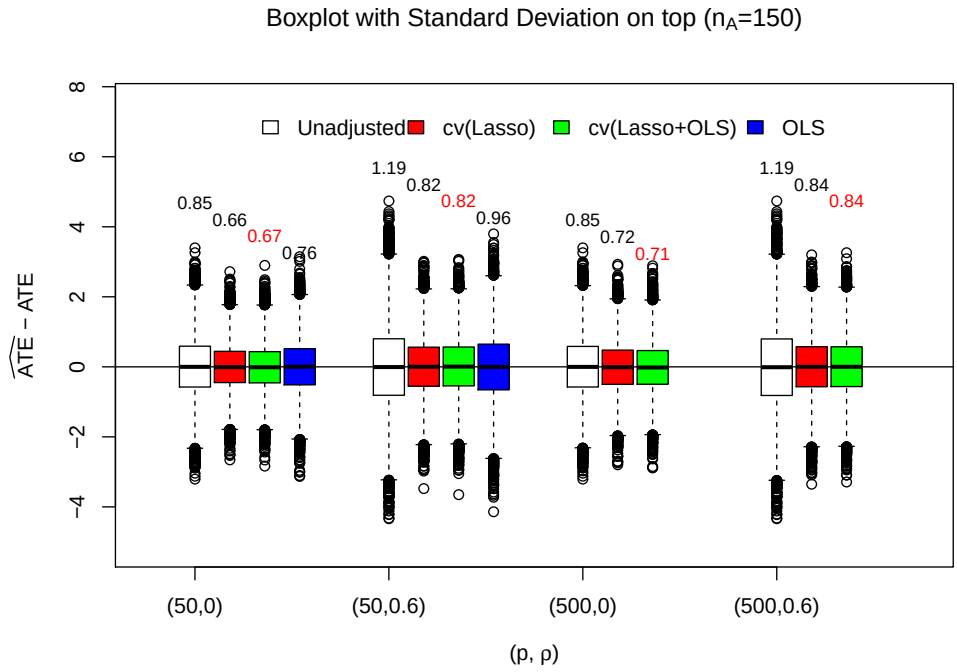

Figure 11: Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 150$.

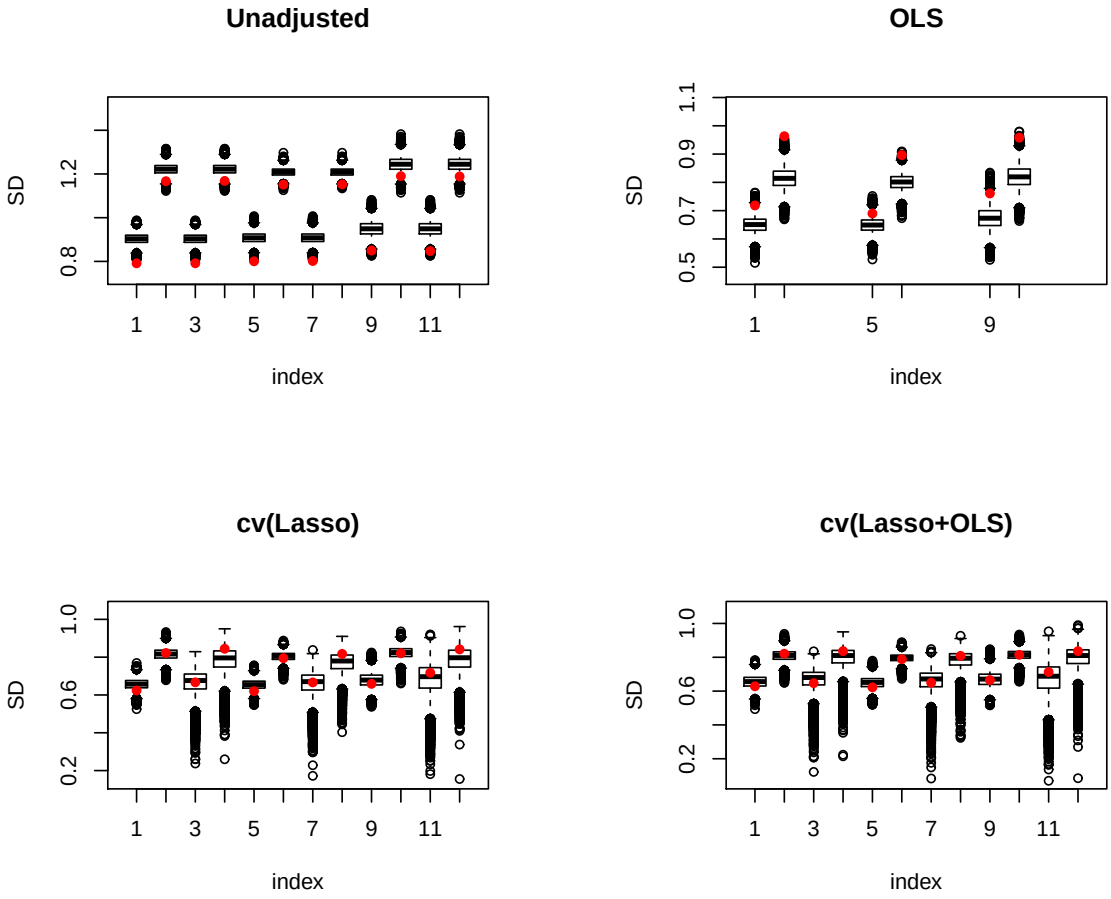

Figure 12: Boxplot of Neyman SD estimate with the "true" SD presented as red dot.

Table 4: Bias, standard deviation (SD) and root-mean square error $\sqrt{\text{MSE}}$ of ATE estimates

				(p, ρ)	
Statistic	Method	(50,0)	(50,0.6)	(500,0)	(500,0.6)
			$n_A = 100$		
	Unadjusted	**0.003(0.004)**	**0.005(0.005)**	**0.002(0.003)**	**0.003(0.005)**
bias	OLS	0.014(0.005)	0.013(0.006)	-	-
	cv(Lasso)	0.007(0.004)	0.014(0.005)	0.006(0.004)	0.005(0.004)
	cv(Lasso+OLS)	0.011(0.004)	0.013(0.005)	0.009(0.004)	**0.003(0.004)**
	Unadjusted	0.79(0.08)	1.17(0.11)	0.79(0.07)	1.17(0.11)
SD	OLS	0.72(0.07)	0.96(0.09)	-	-
	cv(Lasso)	**0.62(0.06)**	**0.82(0.08)**	0.67(0.06)	**0.84(0.08)**
	cv(Lasso+OLS)	0.63(0.06)	**0.82(0.08)**	**0.65(0.06)**	**0.84(0.08)**
	Unadjusted	0.79(0.08)	1.17(0.11)	0.79(0.07)	1.17(0.11)
$\sqrt{\text{MSE}}$	OLS	0.72(0.07)	0.97(0.09)	-	-
	cv(Lasso)	**0.63(0.06)**	**0.82(0.08)**	0.67(0.06)	0.85(0.08)
	cv(Lasso+OLS)	**0.63(0.06)**	**0.82(0.08)**	**0.65(0.06)**	**0.84(0.08)**
			$n_A = 125$		
	Unadjusted	0.008(0.005)	0.011(0.007)	**0.006(0.004)**	0.01(0.007)
bias	OLS	0.008(0.004)	**0.005(0.005)**	-	-
	cv(Lasso)	**0.005(0.003)**	0.012(0.005)	0.007(0.004)	0.004(0.004)
	cv(Lasso+OLS)	0.012(0.004)	0.012(0.005)	0.011(0.004)	**0.003(0.003)**
	Unadjusted	0.80(0.08)	1.15(0.11)	0.8(0.08)	1.15(0.11)
SD	OLS	0.69(0.06)	0.90(0.09)	-	-
	cv(Lasso)	**0.62(0.06)**	**0.79(0.07)**	0.67(0.06)	0.82(0.08)
	cv(Lasso+OLS)	**0.62(0.06)**	**0.79(0.07)**	**0.65(0.06)**	**0.81(0.08)**
	Unadjusted	0.80(0.07)	1.15(0.11)	0.8(0.07)	1.15(0.11)
$\sqrt{\text{MSE}}$	OLS	0.69(0.07)	0.90(0.09)	-	-
	cv(Lasso)	**0.62(0.06)**	0.80(0.08)	0.67(0.06)	0.82(0.08)
	cv(Lasso+OLS)	**0.62(0.06)**	**0.79(0.07)**	**0.65(0.06)**	**0.81(0.08)**
			$n_A = 150$		
	Unadjusted	0.004(0.004)	**0.000(0.005)**	**0.002(0.003)**	0.005(0.005)
bias	OLS	**0.002(0.003)**	0.006(0.005)	-	-
	cv(Lasso)	0.003(0.003)	0.002(0.004)	0.01(0.005)	0.002(0.003)
	cv(Lasso+OLS)	0.011(0.004)	0.006(0.004)	0.017(0.005)	**0.001(0.003)**
	Unadjusted	0.85(0.08)	1.19(0.11)	0.85(0.08)	1.19(0.11)
SD	OLS	0.76(0.07)	0.96(0.09)	-	-
	cv(Lasso)	**0.66(0.06)**	0.82(0.08)	0.72(0.07)	**0.84(0.08)**
	cv(Lasso+OLS)	0.67(0.06)	**0.81(0.07)**	**0.71(0.07)**	**0.84(0.08)**
	Unadjusted	0.85(0.08)	1.19(0.11)	0.85(0.08)	1.19(0.11)
$\sqrt{\text{MSE}}$	OLS	0.76(0.07)	0.96(0.09)	-	-
	cv(Lasso)	**0.66(0.06)**	**0.82(0.08)**	0.72(0.07)	**0.84(0.08)**
	cv(Lasso+OLS)	0.67(0.06)	**0.82(0.08)**	**0.71(0.07)**	**0.84(0.08)**

The numbers in parentheses are the corresponding standard errors estimated by using the bootstrap with $B = 500$ resamplings of the ATE estimates.

Table 5: Mean number of selected covariates for treated and control group

		(p,ρ)			
Group	Method	(50,0)	(50,0.6)	(500,0)	(500,0.6)
			$n_A = 100$		
treated	cv(Lasso)	16	13	22	22
	cv(Lasso+OLS)	6	6	7	7
control	cv(Lasso)	20	11	32	28
	cv(Lasso+OLS)	8	6	7	7
			$n_A = 125$		
treated	cv(Lasso)	17	13	25	24
	cv(Lasso+OLS)	7	6	6	6
control	cv(Lasso)	19	11	32	27
	cv(Lasso+OLS)	8	6	9	8
			$n_A = 150$		
treated	cv(Lasso)	18	13	29	26
	cv(Lasso+OLS)	8	7	6	6
control	cv(Lasso)	19	12	30	25
	cv(Lasso+OLS)	8	6	11	8

Table 6: Coverage probability (%) and mean interval length (in parentheses) for 95% confidence interval

		(p,ρ)		
Methods	(50,0)	(50,0.6)	(500,0)	(500,0.6)
		$n_A = 100$		
Unadjusted	97.3(3.54)	95.8(4.79)	97.3(3.54)	95.8(4.79)
OLS	92.2(2.55)	90.0(3.19)	-	-
cv(Lasso)	95.8(2.58)	94.5(3.20)	94.3(2.61)	92.4(3.07)
cv(Lasso+OLS)	95.6(2.57)	94.4(3.17)	94.8(2.60)	93.0(3.11)
		$n_A = 125$		
Unadjusted	97.4(3.56)	96.0(4.74)	97.3(3.56)	95.9(4.74)
OLS	93.3(2.54)	91.6(3.14)	-	-
cv(Lasso)	96.0(2.56)	95.0(3.15)	94.1(2.59)	92.9(3.02)
cv(Lasso+OLS)	95.7(2.55)	94.9(3.12)	94.4(2.58)	93.6(3.06)
		$n_A = 150$		
Unadjusted	97.1(3.72)	95.8(4.88)	97.1(3.72)	95.8(4.88)
OLS	91.4(2.64)	90.4(3.21)	-	-
cv(Lasso)	95.4(2.66)	94.9(3.23)	92.9(2.68)	92.6(3.08)
cv(Lasso+OLS)	94.7(2.63)	94.8(3.19)	92.0(2.63)	93.1(3.11)

The numbers in parentheses are the corresponding mean interval lengths.

Algorithm 1 K-fold Cross Validation (CV) for the Lasso+OLS estimator

Input: Design matrix X, response Y and a sequence of tuning parameter $\lambda_1, ..., \lambda_J$; Number of folds K.

Output: The optimal tuning parameter selected by CV: $\lambda_{optimal}$.

1: Divide randomly the data $z = (X, Y)$ into K roughly equal parts $z_k, k = 1, ..., K$;

2: For each $k = 1, ..., K$, denote $\hat{S}^{(k)}(\lambda_0) = \emptyset$ and $\hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_0) = 0$.

- Fit the model with parameters $\lambda_j, j = 1, ..., J$ to the other $K-1$ parts $z_{-k} = z \setminus z_k$ of the data, giving the Lasso solution path $\hat{\beta}^{(k)}(\lambda_j), j = 1, ..., J$ and compute the selected covariates set $\hat{S}^{(k)}(\lambda_j) = \{l : \hat{\beta}^{(k)}_l(\lambda_j) \neq 0\}, j = 1, ..., J$ on the path;

- For each $j = 1, ..., J$, compute the Lasso+OLS estimator:

$$\hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_j) = \begin{cases} \underset{\boldsymbol{\beta}:\ \beta_j=0,\ \forall j \notin \hat{S}^{(k)}(\lambda_j)}{\arg\min} \left\{ \dfrac{1}{2|z_{-k}|} \displaystyle\sum_{i \in z_{-k}} (y_i - x_i^T \beta)^2 \right\}, & \text{if} \quad \hat{S}^{(k)}(\lambda_j) \neq \hat{S}^{(k)}(\lambda_{j-1}), \\ \hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_{j-1}), & \text{otherwise;} \end{cases} \tag{82}$$

- Compute the error in predicting the kth part of the data $PE^{(k)}$:

$$PE^{(k)}(\lambda_j) = \frac{1}{|z_k|} \sum_{i \in z_k} \left(y_i - x_i^T \hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_j) \right)^2;$$

3: Compute cross validation error $CV(\lambda_j)$, $j = 1, ..., J$:

$$CV(\lambda_j) = \frac{1}{K} \sum_{k=1}^{K} PE^{(k)}(\lambda_j);$$

4: Compute the optimal λ selected by CV;

$$\lambda_{optimal} = \underset{\lambda_j:\ j=1,...,J}{argmin}\ CV(\lambda_j);$$

5: **return** $\lambda_{optimal}$.