Supporting Information (Appendix)

Neyman-type conservative variance estimate

We have given a Neyman-type conservative estimate of the variance in the main text which can be used to construct a conservative confidence interval for the ATE. In this section, we will study the asymptotic behavior of this variance estimate. Recall that

$$\hat{\sigma}^2_{e^{(a)}} = \frac{1}{n_A - df^{(a)}} \sum_{i \in A} \left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)}_{\text{Lasso}} \right)^2, \tag{S1}$$

$$\hat{\sigma}^2_{e^{(b)}} = \frac{1}{n_B - df^{(b)}} \sum_{i \in B} \left(b_i - \bar{b}_B - (\mathbf{x}_i - \bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)}_{\text{Lasso}} \right)^2, \tag{S2}$$

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 S0.1. 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

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

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

$$\hat{\sigma}^2_{\text{Lasso}} = \frac{n}{n_A} \hat{\sigma}^2_{e^{(a)}} + \frac{n}{n_B} \hat{\sigma}^2_{e^{(b)}}. \tag{S3}$$

Theorem S1. *Assume conditions in Theorem 1 and condition S0.1 hold. Then $\hat{\sigma}^2_{\text{Lasso}}$ converges in probability to*

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

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. *The Neyman-type conservative variance estimate for the unadjusted estimator is given by*

$$\hat{\sigma}^2_{\text{unadj}} = \frac{n}{n_A} \frac{1}{n_A - 1} \sum_{i \in A} (a_i - \bar{a}_A)^2 + \frac{n}{n_B} \frac{1}{n_B - 1} \sum_{i \in B} \left(b_i - \bar{b}_B \right)^2,$$

which, under second moment conditions of potential outcomes a and b, converges in probability to

$$\frac{1}{p_A} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (a_i - \bar{a})^2 + \frac{1}{1-p_A} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (b_i - \bar{b})^2.$$

Therefore, for the $\boldsymbol{\beta}^{(a)}$ and $\boldsymbol{\beta}^{(b)}$ defined in [18], the limit of $\hat{\sigma}^2_{\text{Lasso}}$ is no greater than that of $\hat{\sigma}^2_{\text{unadj}}$ and the difference is

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

Remark 7. *With the conservative variance estimate in Theorem S1, 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 $N > n$.*

Remark 8. *The extra Condition S0.1 is used to obtain the following bounds for the number of selected covariates by the Lasso: $\max(\hat{s}^{(a)}, \hat{s}^{(b)}) = o_p(\min(n_A, n_B))$. Condition S0.1 can be removed from Theorem S1 if we redefine $\hat{\sigma}^2_{e^{(a)}}$ and $\hat{\sigma}^2_{e^{(b)}}$ without adjusting the degrees 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.

Theorem S2. *Assume the 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 9. *Though* $(\hat{\sigma}^*)^2_{\text{Lasso}}$ *has the same limit as* $\hat{\sigma}^2_{\text{Lasso}}$*, our simulation experience shows that, in finite samples, 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.*

Simulation

In this section we carry out simulation studies to evaluate the finite sample performance of $\widehat{ATE}_{\text{Lasso}}$ estimator. We also present results for the $\widehat{ATE}_{\text{OLS}}$ estimator when $p<n$ and the two-step estimator $\widehat{ATE}_{\text{Lasso+OLS}}$ which adopts Lasso to select covariates and then uses OLS to refit the regression coefficients, see [1, 2, 3, 4] 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 $\hat{S}^{(a)}=\{j:\hat{\boldsymbol{\beta}}^{(a)}_j\neq 0\}$ 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 $\widehat{ATE}_{\text{Lasso+OLS}}$ 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. We 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. We should mention that for the cv(Lasso+OLS) adjusted estimator, we compute the CV error for a given value of the λ_a (or λ_b) based on the whole Lasso+OLS estimator instead of the Lasso estimator, see Algorithm 1 for details. Therefore, the cv(Lasso+OLS) adjusted estimator and the cv(Lasso) adjusted estimator may select different covariates to do the adjustment. This type of cross validation for cv(Lasso+OLS) requires more computation effort than the cross validation based on just the Lasso estimator since it needs to compute the OLS estimator for each fold and for each λ_a (or λ_b), but it can give better prediction and covariates selection performance.

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 can compute OLS estimator and compare it with the Lasso. The covariates vector $\mathbf{x}_i$ 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 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 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 a standard normal distribution. The vector $\mathbf{z}_i$ is independent of $\mathbf{x}_i$ and is 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 S4, S5, S6 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 and than that of 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 S1. 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 selected covariates by cv(Lasso) and cv(Lasso+OLS) for treatment group and control group separately, see Table S2. 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). Overall, the cv(Lasso+OLS) adjusted, the cv(Lasso) adjusted, the OLS adjusted and the unadjusted estimators perform from best to worst.

We move now to study the finite sample performance of 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 Neyman 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 S1, S2, S3 present the boxplot 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 S3. 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 with 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 S7 with the "true" SD presented in Table S1 (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); while for the OLS adjusted estimator, the mean of Neyman SD estimate can be $6\% - 100\%$ smaller than the "true" SD which may be because of 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. Although the Neyman variance estimate is asymptotically conservative, in small samples the variance estimate may be too small and hence the confidence intervals are too narrow. 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 a linear 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.

The design matrix of the PAC data

In the PAC data, there are 59 covariates (main effects) including 50 indicators which may be correlated with the outcomes. One of the main effects (called interactnew) has heavy tail, so we perform the transform: $x \to \log(|x| + 1)$ to make it more 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 predict the potential outcomes, so we included them in the design matrix. The quadratic terms and the interactions between non-indicator covariates and the interactions between indicator covariates and 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 correlation larger than 0.95) with the main effects and remove the indicators with very sparse entries (where the number of 1's is less than 20). Finally, we form a design matrix X with 1172 columns (covariates) and 1013 rows (subjects).

Estimation of constants in the 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 relevant covariates sets $S^{(a)}, S^{(b)}$ and then 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 and the estimation of $S^{(b)}$ and $e^{(b)}$ are 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 $x_i \in R^{1172}$ the covariates vector of the ith subject. For each $d = 1, ..., 1000$, we draw a bootstrap sample $\{(a_i^*(d), x_i^*(d)) : i \in A\}$ with replacement from the data points $\{(a_i, x_i) : i \in A\}$. We then compute the LassoOLS(CV) adjusted vector $\hat{\boldsymbol{\beta}}(d)$ based on each bootstrap sample $\{(a_i^*(d), x_i^*(d)) : i \in A\}$. Let τ_j be the selection fraction of non-zero $\hat{\beta}_j(d)$ in the 1000 bootstrap estimators, i.e., $\tau_j = (1/1000)\sum_{d=1}^{1000} \mathbb{I}_{\{\hat{\beta}_j(d) \neq 0\}}$, where $\mathbb{I}$ is the indicator function. We form the relevant covariates $S^{(a)}$ by the covariates whose selection fraction are 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:

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

Proofs of Theorems 1, S1, S2 and Corollary 1

In this section, we will prove Theorem s1, S1, S2, and Corollary 1 under weaker sparsity conditions than those given in the main text.

Definition 1. 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, $s_{\lambda_a}^{(a)}$ and $s_{\lambda_b}^{(b)}$ are defined as

$$s_{\lambda_a}^{(a)} = \sum_{j=1}^{p} \min\left\{\frac{|\beta_j^{(a)}|}{\lambda_a}, 1\right\},\ s_{\lambda_b}^{(b)} = \sum_{j=1}^{p} \min\left\{\frac{|\beta_j^{(b)}|}{\lambda_b}, 1\right\}, \tag{S4}$$

respectively. We will allow $s_{\lambda_a}^{(a)}$ and $s_{\lambda_b}^{(b)}$ 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 $s_\lambda = \max\left\{s_{\lambda_a}^{(a)}, s_{\lambda_b}^{(b)}\right\}$,

$$\delta_n = o\left(\frac{1}{s_\lambda \sqrt{\log p}}\right), \tag{S5}$$

$$(s_\lambda \log p)/\sqrt{n} = o(1). \tag{S6}$$

- **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

$$\|\mathbf{h}_S\|_1 \leq C s_\lambda \|\Sigma \mathbf{h}\|_\infty,\ \forall \mathbf{h} \in \mathcal{C}, \tag{S7}$$

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

$$S = \{j : |\beta_j^{(a)}| > \lambda_a \text{ or } |\beta_j^{(b)}| > \lambda_b\}. \tag{S8}$$

- **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

$$\lambda_a \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}} + \delta_n\right), \tag{S9}$$

$$\lambda_b \in (\frac{1}{\eta}, M] \times \left(\frac{2(1+\tau)L^{1/2}}{p_B} \sqrt{\frac{2\log p}{n}} + \delta_n \right). \tag{S10}$$

It is easy to verify that Condition (*) is implied by Conditions 1 - 6 of the main text. We will prove Theorems 1, S1, S2, 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 $s^{(b)}_{\lambda_b}$. Note that we can assume, without loss of generality, that

$$\bar{a} = 0,\ \bar{b} = 0,\ \bar{\mathbf{x}} = \mathbf{0}. \tag{S11}$$

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

$$\widehat{ATE}_{\text{Lasso}} = \left[\bar{a}_A - (\bar{\mathbf{x}}_A)^T \hat{\boldsymbol{\beta}}^{(a)} \right] - \left[\bar{b}_B - (\bar{\mathbf{x}}_B)^T \hat{\boldsymbol{\beta}}^{(b)} \right]. \tag{S12}$$

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

Lemma S1. *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 $t > 0$,*

$$P\left(\bar{z}_A - \bar{z} \geq t\right) \leq \exp\left\{ -\frac{p_A n_A t^2}{(1+\tau)^2 \sigma^2} \right\}, \tag{S13}$$

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

Remark. *Massart showed in his paper [5] that for sampling without replacement, the following concentration inequality holds:*

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

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 $n_A \leq n/2$, let $m \geq 2$ and $r \geq 0$ be integers satisfying $n - n_A m = r < n_A$. Let $u \geq 0$, we first prove that

$$E \exp\left(u \sum_{i \in A} z_i \right) \leq E \exp\left(u\delta \sum_{i \in B} z_i / \{m(m+1)\} + u^2 n \sigma^2 / 4 \right) \tag{S14}$$

for a random subset $B \subset \{1, \ldots, n\}$ of fixed size $|B| \leq n/2$ 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

$$E_2 \exp\left(u \sum_{k=1}^{n_A} w_k \right) \leq \exp\left(u \sum_{k=1}^{n_A} \bar{z}_k + (u^2/8) \sum_{k=1}^{n_A} \Delta_k^2 \right). \tag{S15}$$

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

$$E_2 \exp\left(u \sum_{k=1}^{n_A} w_k \right) \leq \exp\left(u \sum_{k=1}^{n_A} \bar{z}_k + u^2 n \sigma^2 / 4 \right) \tag{S16}$$

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

$$\sum_{k=1}^{n_A} \bar{z}_k = \sum_{i \in B} z_i / \{m(m+1)\}. \tag{S17}$$

This yields (S14) with $\delta = 1$ when $|B| \leq n/2$. Otherwise, (S14) 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 (S14) yields

$$\begin{aligned} E \exp\left(u \sum_{i \in A} z_i \right) &\leq E \exp\left[u\delta' \sum_{i \in B'} z_i / \{m(m+1)m'(m'+1)\} + \left(1 + \{m(m+1)\}^{-2}\right) u^2 n \sigma^2 / 4 \right] \\ &\leq \exp\left[\left(1 + \{m(m+1)\}^{-2}(1 + 1/36 + 1/36^2 + \cdots)\right) u^2 n \sigma^2 / 4 \right] \\ &= \exp\left[\left(1 + (36/35)\{m(m+1)\}^{-2}\right) u^2 n \sigma^2 / 4 \right] \\ &\leq \exp\left[(1+\tau)^2 u^2 n \sigma^2 / 4 \right] \end{aligned} \tag{S18}$$

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

$$E \exp\left(u \sum_{i \in A} z_i\right) \leq \exp\left[(1+\tau)^2 u^2 n \sigma^2 / 4\right] \tag{S19}$$

for $n_A > n/2$. This yields (S13) via the usual

$$\begin{aligned} P\left\{\bar{z}_A - \bar{z} > t\right\} &\leq \exp\left[-ut + (1+\tau)^2 u^2 n \sigma^2/(4n_A^2)\right] \\ &= \exp\left[-2\frac{p_A n_A t^2}{(1+\tau)^2\sigma^2} + \frac{p_A n_A t^2}{(1+\tau)^2\sigma^2}\right] \end{aligned} \tag{S20}$$

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

Proof of Theorem 1.

Proof. Recall the decompositions of the potential outcomes:

$$a_i = \bar{a} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(a)} + e_i^{(a)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(a)} + e_i^{(a)}, \tag{S21}$$

$$b_i = \bar{b} + (\mathbf{x}_i - \bar{\mathbf{x}})^T \boldsymbol{\beta}^{(b)} + e_i^{(b)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(b)} + e_i^{(b)}. \tag{S22}$$

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

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

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 [6] with a and b replaced by $e^{(a)}$ and $e^{(b)}$ respectively. To bound (∗∗), we will apply Hölder's 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

$$\left|(\bar{\mathbf{x}}_A)^T \mathbf{h}^{(a)}\right| \leq \|\bar{\mathbf{x}}_A\|_\infty \|\mathbf{h}^{(a)}\|_1. \tag{S23}$$

We bound the two terms on the right hand side of (S23) by the following Lemma S2 and Lemma S3, respectively.

Lemma S2. *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 S3. *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 (∗∗)$= \sqrt{n} \cdot O_p\left(\sqrt{\frac{\log p}{n}}\right) \cdot o_p\left(\frac{1}{\sqrt{\log p}}\right) = o_p(1)$.

□

Proof of Corollary 1.

Proof. By Theorem 1 in [6], 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

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

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

$$\sigma_{e^{(a)}}^2 - \sigma_a^2 = ||e^{(a)}||_2^2 - ||a||_2^2 = -||X\boldsymbol{\beta}^{(a)}||_2^2, \tag{S24}$$

$$\sigma_{e^{(b)}}^2 - \sigma_b^2 = ||e^{(b)}||_2^2 - ||b||_2^2 = -||X\boldsymbol{\beta}^{(b)}||_2^2, \tag{S25}$$

$$\sigma_{e^{(a)}e^{(b)}} - \sigma_{ab} = (e^{(a)})^T(e^{(b)}) - a^T b = -(X\boldsymbol{\beta}^{(a)})^T(X\boldsymbol{\beta}^{(b)}) \tag{S26}$$

Combining (S24), (S25) and (S26), we obtain the corollary. □

Proof of Theorem S1.

Proof. To prove Theorem S1, it is enough to show that

$$\hat{\sigma}^2_{e^{(a)}} \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(a)}}, \tag{S27}$$

$$\hat{\sigma}^2_{e^{(b)}} \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(b)}}. \tag{S28}$$

We will only prove the statement (S27) and omit the proof of the statement (S28) since it is identical.

We first state the following two lemmas. Lemma S4 bounds the number of selected covariates (covariates with a nonzero coefficient), while Lemma S5 establishes conditions under which the subsample mean (without replacement) converges in probability to the population mean.

Lemma S4. *Under conditions in Theorem S1, there exists a constant C, such that the following holds with probability going to 1:*

$$\hat{s}^{(a)} \le Cs; \quad \hat{s}^{(b)} \le Cs. \tag{S29}$$

The proof of Lemma S4 can be found below.

Lemma S5. *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*

$$\frac{1}{n}\sum_{i=1}^{n} |z_i|^{1+\epsilon} \le L. \tag{S30}$$

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

$$\bar{z}_A \xrightarrow{p} \lim_{n\to\infty} \bar{z}. \tag{S31}$$

By definition (S1) and simple calculations,

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

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

$$\frac{1}{n_A} \sum_{i\in A} \left(e_i^{(a)} - \bar{e}_A^{(a)}\right)^2 = \frac{1}{n_A} \sum_{i\in A} (e_i^{(a)})^2 - (\bar{e}_A^{(a)})^2. \tag{S32}$$

By the $4th$ moment condition on the approximation error $e^{(a)}$ (see (7)), and applying Lemma S5 gives

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

Therefore,

$$\frac{1}{n_A} \sum_{i\in A} \left(e_i^{(a)} - \bar{e}_A^{(a)}\right)^2 \xrightarrow{p} \lim_{n\to\infty} \sigma^2_{e^{(a)}}. \tag{S33}$$

By simple algebra,

$$\begin{aligned}
&\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\\
=\;&(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})^T\left[\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\right](\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\\
\leq\;&||\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)}||_1^2\cdot||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T||_\infty. \qquad \text{(S34)}
\end{aligned}$$

We next show that (S34) converges to 0 in probability. By Lemma S3 and Lemma S7, we have

$$||\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)}||_1=\|\mathbf{h}^{(a)}\|_1=o_p\left(\frac{1}{\sqrt{\log p}}\right), \qquad \text{(S35)}$$

$$||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T||_\infty=O_p(1). \qquad \text{(S36)}$$

Therefore,

$$\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\xrightarrow{p}0. \qquad \text{(S37)}$$

By Cauchy-Schwarz inequality,

$$\begin{aligned}
&|\frac{1}{n_A}\sum_{i\in A}(e_i^{(a)}-\bar{e}_A^{(a)})(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})|\\
\leq\;&\left[\frac{1}{n_A}\sum_{i\in A}\left(e_i^{(a)}-\bar{e}_A^{(a)}\right)^2\right]^{\frac{1}{2}}\left[\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}^{(a)}-\hat{\boldsymbol{\beta}}^{(a)})\right)^2\right]^{\frac{1}{2}} \qquad \text{(S38)}
\end{aligned}$$

which converges to 0 in probability because of (S33) and (S37).

By Lemma S4 and Condition 4, we have

$$\frac{n_A}{n_A-df^{(a)}}=\frac{n_A}{n_A-\hat{s}^{(a)}-1}\xrightarrow{p}1. \qquad \text{(S39)}$$

Combining (S33), (S37), (S38) and (S39), 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:

$$\begin{aligned}
&\left(\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}}\right)-\left(\frac{1-p_A}{p_A}\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\frac{p_A}{1-p_A}\lim_{n\to\infty}\sigma^2_{e^{(b)}}+2\lim_{n\to\infty}\sigma_{e^{(a)}e^{(b)}}\right)\\
=\;&\lim_{n\to\infty}\sigma^2_{e^{(a)}}+\lim_{n\to\infty}\sigma^2_{e^{(b)}}-2\lim_{n\to\infty}\sigma_{e^{(a)}e^{(b)}}\\
=\;&\lim_{n\to\infty}\sigma^2_{e^{(a)}-e^{(b)}}\\
=\;&\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\left(a_i-b_i-\mathbf{x}_i^T(\boldsymbol{\beta}^{(a)}-\boldsymbol{\beta}^{(b)})\right)^2. \qquad \text{(S40)}
\end{aligned}$$

□

Proof of Theorem S2.

Proof. By Lemma S4, $\max(\hat{s}^{(a)},\hat{s}^{(b)})=o_p(\min(n_A,n_B))$. 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 S1. □

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.

Proof of Lemma S2.

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

$$P\left(\|\bar{\mathbf{x}}_A\|_\infty > c_n\right) = P\left(\max_{j=1,\ldots,p}\left|\frac{1}{n_A}\sum_{i\in A}x_{ij}\right| > c_n\right) \le \sum_{j=1}^{p} P\left(\left|\frac{1}{n_A}\sum_{i\in A}x_{ij}\right| > c_n\right). \tag{S41}$$

By Cauchy-Schwarz inequality, we have

$$\frac{1}{n}\sum_{i=1}^{n}x_{ij}^2 \le \left(\frac{1}{n}\sum_{i=1}^{n}x_{ij}^4\right)^{\frac{1}{2}}\left(\frac{1}{n}\sum_{i=1}^{n}1^2\right)^{\frac{1}{2}} \le \sqrt{L}. \tag{S42}$$

Substituting the concentration inequality (S13) into (S41),

$$P\left(\|\bar{\mathbf{x}}_A\|_\infty > c_n\right) \le 2\exp\left\{\log p - \frac{p_A n_A c_n^2}{(1+\tau)^2 L^{1/2}}\right\} = 2\exp\{-\log p\} \to 0.$$

□

Proof of Lemma S3.

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

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)\left(a_i - \bar{a}_A - (\mathbf{x}_i - \bar{\mathbf{x}}_A)^T\hat{\boldsymbol{\beta}}\right) = \lambda_a\kappa, \tag{S43}$$

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

$$\kappa \in \partial||\boldsymbol{\beta}||_1\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \quad \text{with} \quad \begin{cases}\kappa_j \in [-1,1] \text{ for } j \text{ s.t. } \hat{\beta}_j = 0\\ \kappa_j = \text{sign}(\hat{\beta}_j) \text{ otherwise}\end{cases} \tag{S44}$$

Substituting a_i by the decomposition (3), (S43) becomes

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A) = \lambda_a\kappa. \tag{S45}$$

Multiplying both sides of (S45) 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 is 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 by 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 S6 shows that, with λ_a defined appropriately, $\mathcal{L}$ holds with probability approaching 1. We will prove this Lemma later.

Lemma S6. *Define* $\mathcal{L} = \left\{ \left\| \frac{1}{n_A} \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A) \right\|_\infty \leq \eta \lambda_a \right\}$. *Then under the conditions of Theorem 1,* $P(\mathcal{L}) \to 1$.

On $\mathcal{L}$

$$\frac{1}{n_A} \sum_{i \in A} \left((\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \mathbf{h} \right)^2 \leq \lambda_a \left(\|\boldsymbol{\beta}\|_1 - \|\hat{\boldsymbol{\beta}}\|_1 + \eta \|\mathbf{h}\|_1 \right). \tag{S46}$$

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

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

Therefore,

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

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

$$(1 - \eta) \|\mathbf{h}_{S^c}\|_1 \leq (1 + \eta) \|\mathbf{h}_S\|_1 + 2 \|\boldsymbol{\beta}_{S^c}\|_1 \leq (1 + \eta) \|\mathbf{h}_S\|_1 + 2 s \lambda_a. \tag{S47}$$

where the last inequality is because of the definition of s in (S4) and S in (S8).

Consider the following two cases:

(I) If $(1 + \eta)\|\mathbf{h}_S\|_1 + 2 s \lambda_a \geq (1 - \eta)\xi \|\mathbf{h}_S\|_1$ then by (S47),

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

By the definition of λ_a and the scaling assumptions (S5), (S6), we have that $s \lambda_a = o\left(\frac{1}{\sqrt{\log p}} \right)$.

(II) If $(1+\eta)\|\mathbf{h}_S\|_1 + 2 s \lambda_a < (1-\eta)\xi \|\mathbf{h}_S\|_1$ then by (S47) we have $\|\mathbf{h}_{S^c}\|_1 \leq \xi \|\mathbf{h}_S\|_1$. Applying the cone invertibility condition on the design matrix (S7),

$$\|\mathbf{h}\|_1 = \|\mathbf{h}_S\|_1 + \|\mathbf{h}_{S^c}\|_1 \leq (1 + \xi) \|\mathbf{h}_S\|_1 \leq (1 + \xi) C s \left\| \frac{1}{n} X^T X \mathbf{h} \right\|_\infty \tag{S48}$$

Before applying this inequality we will revisit the KKT condition (S44), but this time we will take the l_∞-norm, yielding

$$\left\| \frac{1}{n_A} \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(\mathbf{x}_i - \bar{\mathbf{x}}_A)^T \mathbf{h} \right\|_\infty \leq \lambda_a + \left\| \frac{1}{n_A} \sum_{i \in A} (\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A) \right\|_\infty \leq (1 + \eta) \lambda_a \tag{S49}$$

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 (S48). 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 S7. *Assume stability of treatment assignment probability condition 1 and moment condition 6 hold. Then* $P(\mathcal{M}) \to 1$.

We will prove Lemma S7 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 (S6) and (S49) in the second line. Hence, by (S48),

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

Again, applying the scaling assumptions (S5) and (S6), we get $\|\mathbf{h}\|_1 = o_p\left(\frac{1}{\sqrt{\log p}} \right)$. □

Proof of Lemma S4.

Proof. In the proof of Lemma S3, we have shown that, on $\mathcal{L}$ defined in Lemma S6,

$$\frac{1}{n_A}\sum_{i\in A}\left((\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right)^2 \le \lambda_a\left(\|\boldsymbol{\beta}\|_1-\|\hat{\boldsymbol{\beta}}\|_1+\eta||\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}||_1\right).$$
$$\le \lambda_a(1+\eta)||\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}||_1. \tag{S50}$$

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

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

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

$$\Delta_j := \left|\frac{1}{n_A}\sum_{i\in A}(x_{ij}-\bar{\mathbf{x}}_A^j)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right| \ge (1-\eta)\lambda_a. \tag{S51}$$

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

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

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

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

The last inequality holds because

$$\begin{aligned}\lambda_{\max}(Z_AZ_A^T) &= \max_{\mathbf{u}:||\mathbf{u}||_2=1}\mathbf{u}^TZ_AZ_A^T\mathbf{u}\\ &= \max_{\mathbf{u}:||\mathbf{u}||_2=1}\mathbf{u}^T\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T\mathbf{u}\\ &= \max_{\mathbf{u}:||\mathbf{u}||_2=1}\mathbf{u}^T\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T\mathbf{u}-n_A\mathbf{u}^T(\bar{\mathbf{x}}_A)(\bar{\mathbf{x}}_A)^T\mathbf{u}\\ &\le \max_{\mathbf{u}:||\mathbf{u}||_2=1}\mathbf{u}^T\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T\mathbf{u}\le n\Lambda_{\max}.\end{aligned} \tag{S52}$$

On the other hand,

$$\frac{1}{n_A^2}\mathbf{w}_A^TZ_A^TZ_A\mathbf{w}_A=\sum_{j=1}^p\Delta_j^2\ge\sum_{j:\hat{\beta}_j\neq 0}\Delta_j^2\ge(1-\eta)^2\lambda_a^2\hat{s}. \tag{S53}$$

Combining (S51), (S53) and the fact that with probability going to 1 (see the proof of Lemma S3)

$$||\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}||_1\le Cs(1+\eta)\lambda_a,$$

where C is a constant. We conclude that with probability going to 1

$$\hat{s}\le\frac{1}{(1-\eta)^2}\frac{1}{\lambda_a^2}\Lambda_{\max}\frac{n}{n_A}\lambda_a(1+\eta)Cs(1+\eta)\lambda_a\le\frac{C(1+\eta)^2}{p_A(1-\eta)^2}s.$$

□

Proof of Lemma S5.

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

$$P(|\bar{z}_A - \lim_{n\to\infty} \bar{z}| > t) \leq P(|\bar{z}_A - \bar{z}| > t/2) + P(|\bar{z} - \lim_{n\to\infty} \bar{z}| > t/2). \tag{S54}$$

The second term in the right hand side of (S54) obviously converges to 0 as $n \to \infty$. To bound the first term, we apply the concentration inequality (S13). By (S30), it is easy to show

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\sum_{i=1}^{n} |z_i|^{1-\epsilon}|z_i|^{1+\epsilon} \leq (nL)^{\frac{1-\epsilon}{1+\epsilon}} \frac{1}{n}\sum_{i=1}^{n} |z_i|^{1+\epsilon} \leq L^{\frac{2}{1+\epsilon}} n^{\frac{1-\epsilon}{1+\epsilon}}.$$

Concentration inequality (S13) yields

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

□

Proof of Lemma S6.

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,

$$||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i - \bar{\mathbf{x}}_A)(e_i - \bar{e}_A)||_\infty \leq ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty + ||(\bar{\mathbf{x}}_A)(\bar{e}_A)||_\infty. \tag{S55}$$

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

$$\begin{aligned} ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty &\leq ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i e_i||_\infty + ||\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i e_i||_\infty \\ &\leq ||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i e_i||_\infty + \delta_n. \end{aligned} \tag{S56}$$

We will again bound (S56) by the concentration inequality (S13) in Lemma S1. By Cauchy-Schwarz inequality, we have for any $j = 1, .., p$,

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

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 (S13),

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

Taking this back to (S56), we have

$$P\left(||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i e_i||_\infty \leq t_n + \delta_n\right) \to 1. \tag{S57}$$

For the second term, by Lemma S2, we have shown that,

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

Similar proof yields

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

Hence, under the scaling condition (S6),

$$P\left(\|(\bar{\mathbf{x}}_A)(\bar{e}_A)\|_\infty \leq \frac{(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{2\log p}{n}}\right) \to 1. \tag{S58}$$

Combining (S57) and (S58) yields

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

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

Proof of Lemma S7.

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,

$$||\frac{1}{n_A}\sum_{i\in A}(\mathbf{x}_i-\bar{\mathbf{x}}_A)(\mathbf{x}_i-\bar{\mathbf{x}}_A)^T-\frac{1}{n}X^TX||_\infty \tag{S59}$$

$$\le \underbrace{||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T-\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T||_\infty}_{*}+\underbrace{||(\bar{\mathbf{x}}_A)(\bar{\mathbf{x}}_A)^T||_\infty}_{**}. \tag{S60}$$

We control the first term $(*)$ again using the concentration inequality (S13) and the union bound. By Cauchy-Schwarz inequality, for $j,k=1,...,p$,

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

Then,

$$\begin{aligned}&P\left(||\frac{1}{n_A}\sum_{i\in A}\mathbf{x}_i\mathbf{x}_i^T-\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i\mathbf{x}_i^T||_\infty\ge\frac{(1+\tau)L^{1/2}}{p_A}\sqrt{\frac{3\log p}{n}}\right)\\ \le\ & 2\exp\left\{2\log p-\frac{3p_An_A(1+\tau)^2L\log p}{(1+\tau)^2Lp_A^2n}\right\}=2\exp\{-\log p\}\to 0.\end{aligned} \tag{S61}$$

The second term $(**)$ is bounded by again observing that, by Lemma S2 and the scaling condition (S6),

$$(**)\le||\bar{\mathbf{x}}_A||_\infty^2=o_p(\sqrt{\frac{\log p}{n}}). \tag{S62}$$

Combining (S61) and (S62) yields the conclusion. □

Tables and Figures

References

1. Meinshausen N (2007) Relaxed Lasso. *Computational Statistics and Data Analysis* 52(1):374-393.
2. Efron B, Hastie T and Tibshirani R (2004) Least angle regression. *Annals of Statistics* 32(2):407-499.
3. Belloni A, Chernozhukov V (2009) Least Squares After Model Selection in High-dimensional Sparse Models. *Bernoulli* 19(2):521-547.
4. Liu H, Yu B (2013) Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 7(2013):3124-3169.
5. Massart P (1986) in *Geometrical and Statistical Aspects of Probability in Banach Spaces.* (Springer), pp. 73-109.
6. Freedman DA (2008) On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics* 2(1):176-196.

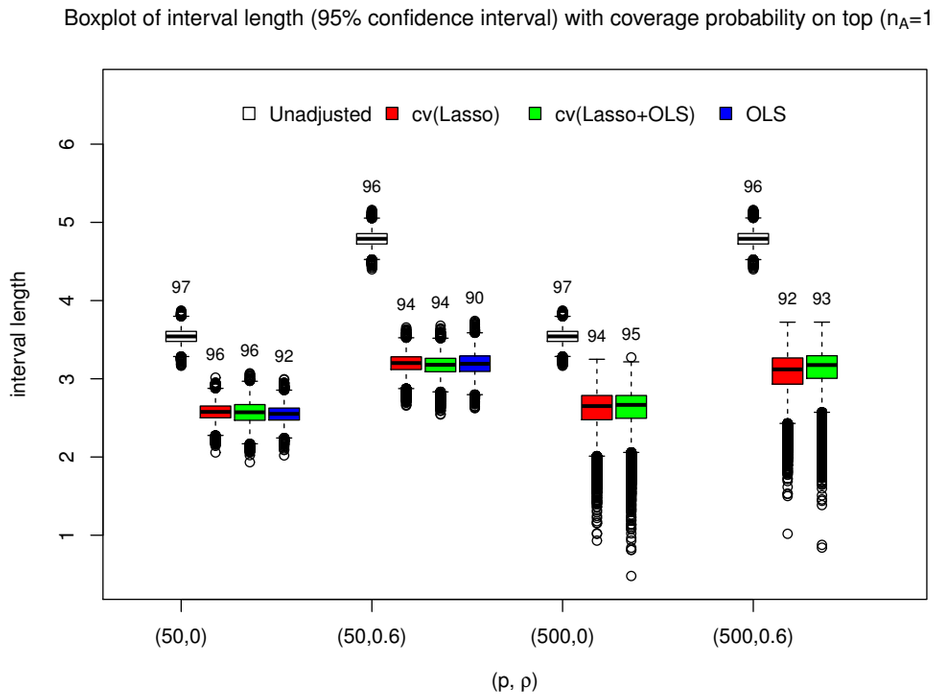

Fig. S1. Boxplot of the interval length with coverage probability $(\%)$ on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 100$.

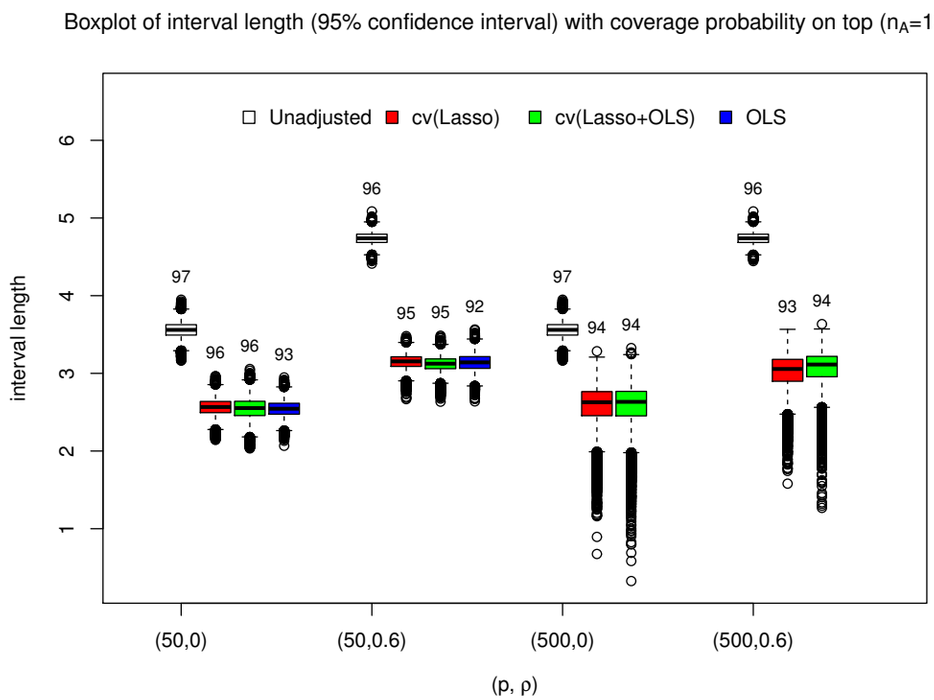

Fig. S2. Boxplot of the interval length with coverage probability $(\%)$ on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 125$.

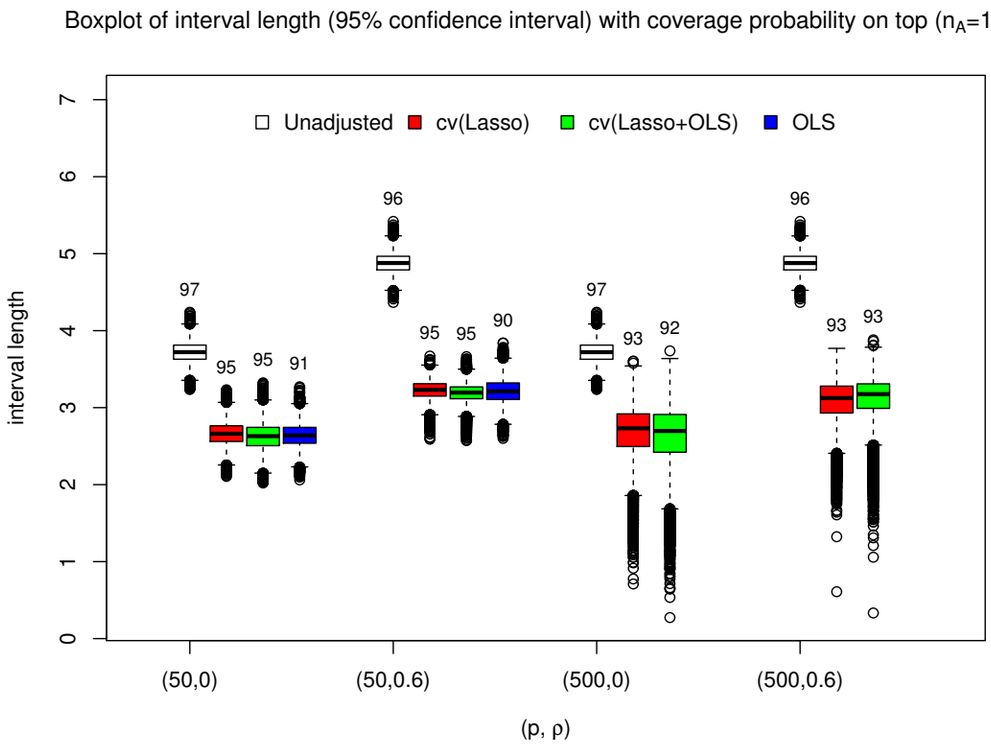

Fig. S3. Boxplot of the interval length with coverage probability $(\%)$ on top of each box for the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) adjusted and cv(Lasso+OLS) adjusted estimators with $n_A = 150$.

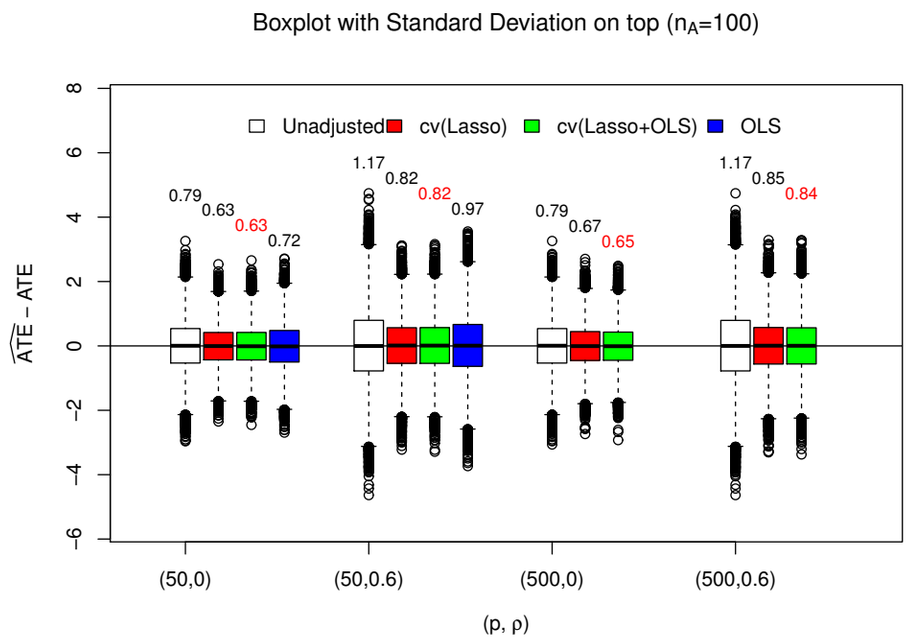

Fig. S4. Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 100$.

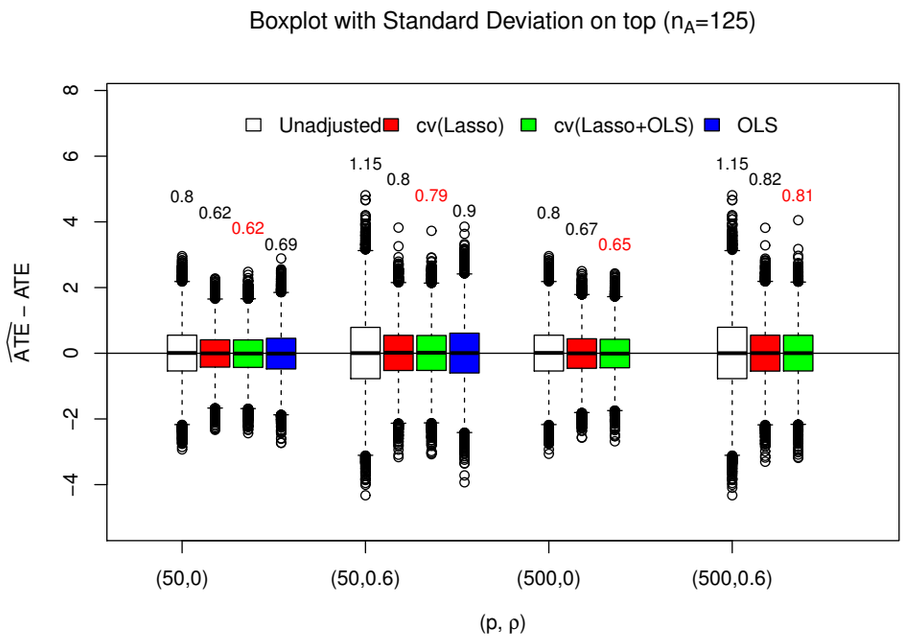

Fig. S5. Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 125$.

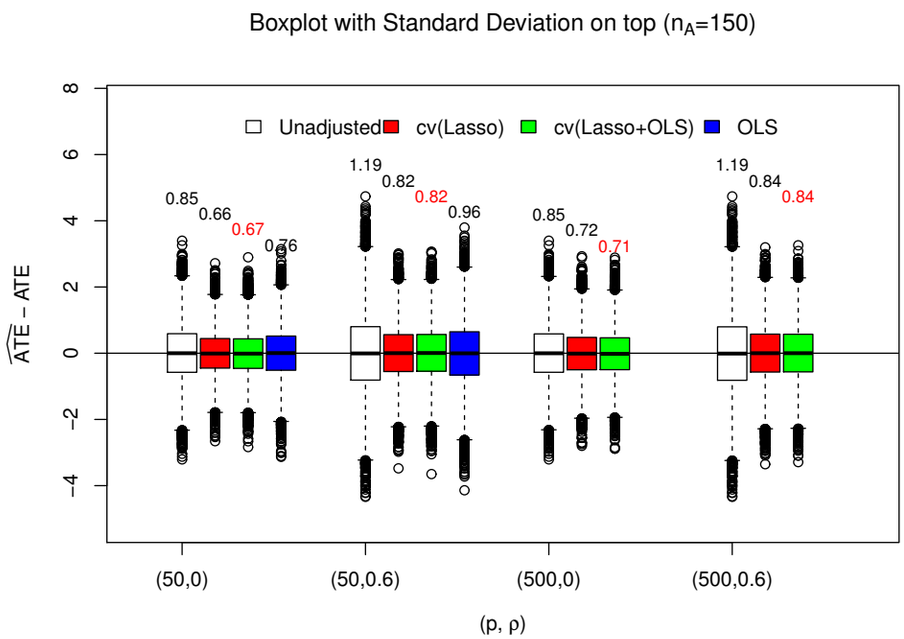

Fig. S6. Boxplot of the unadjusted, OLS adjusted (only computed when $p = 50$), cv(Lasso) and cv(Lasso+OLS) adjusted estimators with their standard deviations presented on top of each box for $n_A = 150$.

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:

$$\hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_j) = \begin{cases} \underset{\boldsymbol{\beta}:\ \beta_j=0,\ \forall j \notin \hat{S}^{(k)}(\lambda_j)}{\arg\min} \left\{ \frac{1}{2|z_{-k}|} \sum_{i \in z_{-k}} (y_i - x_i^T \beta)^2 \right\}, & \text{if} \quad \hat{S}^{(k)}(\lambda_j) \neq \hat{S}^{(k)}(\lambda_{j-1}), \\ \hat{\beta}^{(k)}_{\text{Lasso+OLS}}(\lambda_{j-1}), & \text{otherwise;} \end{cases} \tag{S63}$$

- 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** $\lambda_{optimal}$.

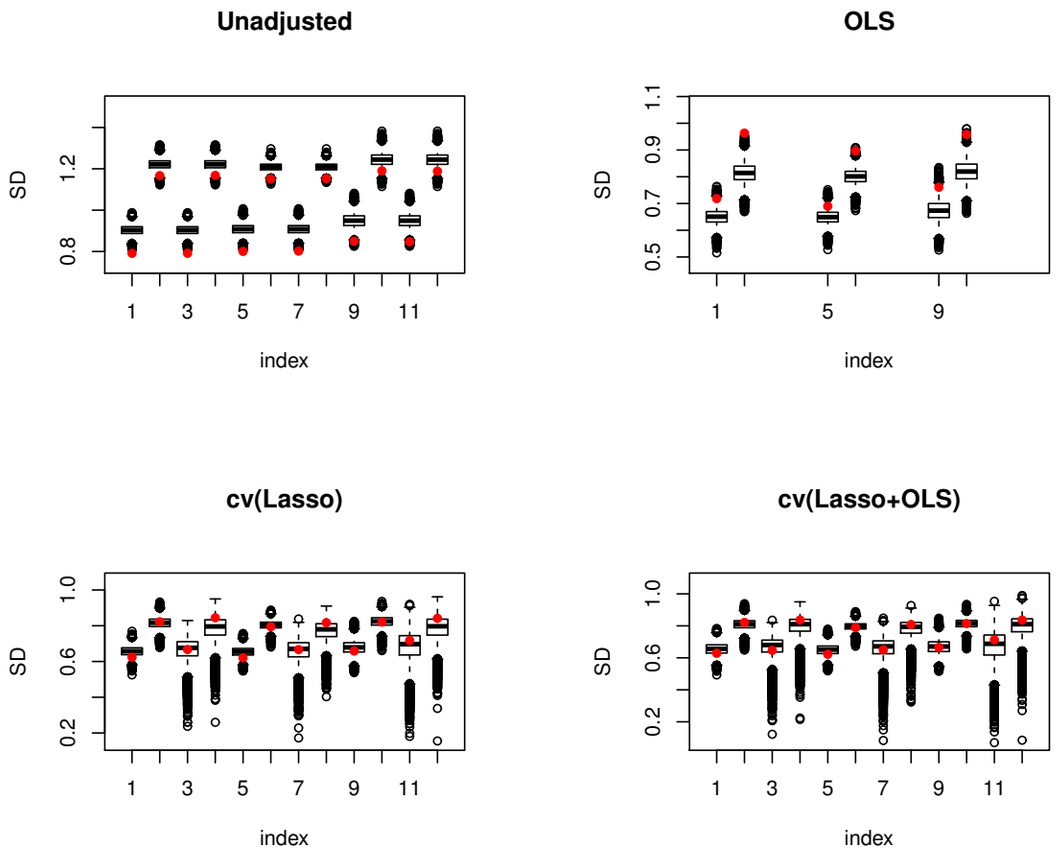

Fig. S7. Boxplot of Neyman SD estimate with the "true" SD presented as red dot.

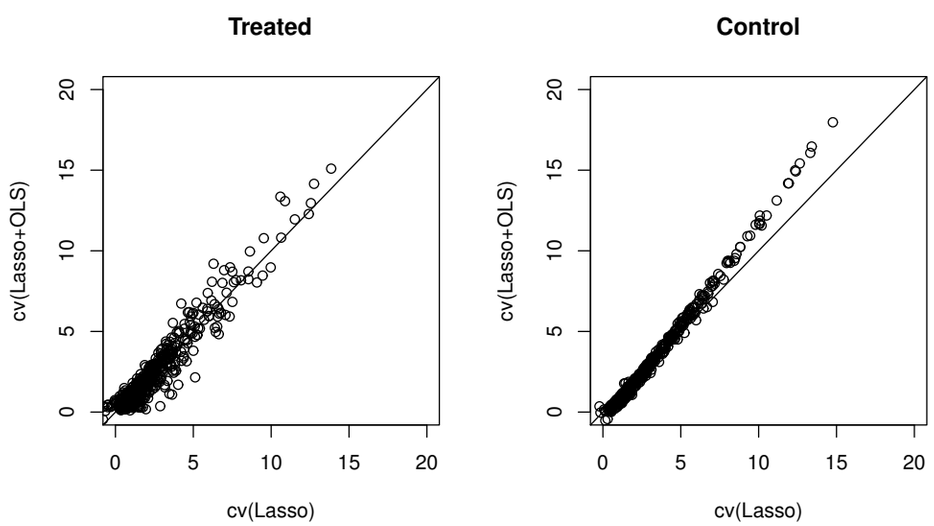

Fig. S8. Adjustment (fitted) value comparison for cv(Lasso) and cv(Lasso+OLS).

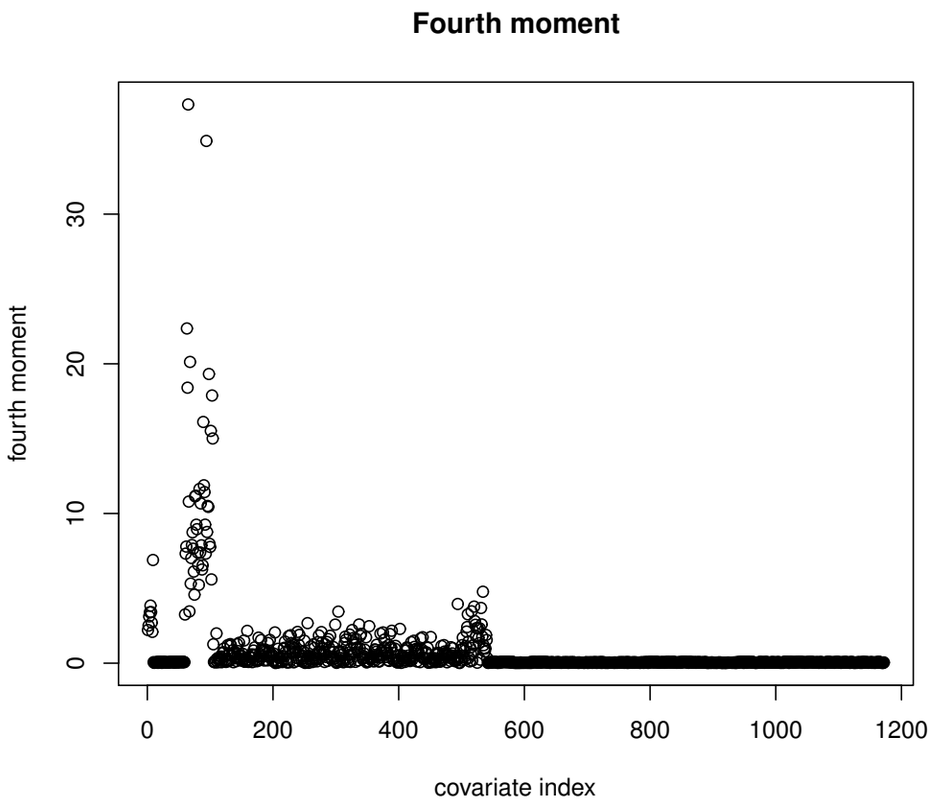

Fig. S9. Fourth moment of each covariate. The covariates with the largest two fourth moments (37.3 and 34.9 respectively) are quadratic term $interactnew^2$ and interaction term $IMscorerct : systemnew$ respectively. Neither of them are selected by the Lasso to do the adjustment. All the fourth moments of the main effects are less than 7.

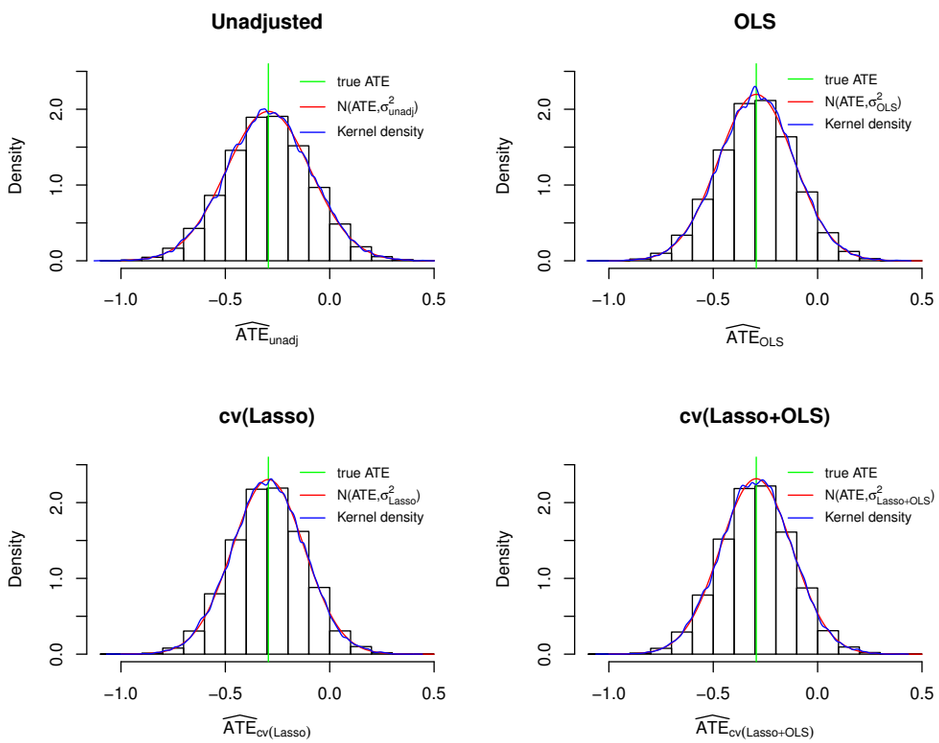

Fig. S10. Histograms of ATE estimates. The green vertical lines are the true ATE; the red curves are the density of normal distribution; the blue curves are the kernel density estimate. The blue curves are very close to the red ones meaning that all the ATE estimates follow normal distribution.

Table S1. 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 S2. Mean number of selected covariates for treated and control group

Group	Method	(p,ρ) (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 S3. Coverage probability $(\%)$ and mean interval length (in parentheses) for 95% confidence interval

Methods	(p,ρ) (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.

Table S4. Statistics for the PAC illustration

Methods	$\widehat{ATE}$	$\hat{\sigma}_{\mathrm{ATE}}$	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

Table S5. Statistics for the PAC synthetic data set

	Bias	SD	$\sqrt{\text{MSE}}$	Coverage (%)	Length	No. of selected covariates: treated	No. of selected covariates: 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.