Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 12: Double LASSO for Linear Causal Model with High-dimensional Controls

Instructor: Ma, Jun

Renmin University of China

December 8, 2021

Post-LASSO estimation

- The LASSO estimator is always biased if $\lambda \neq 0$.
- Recall that when $\mathbf{X}^\top \mathbf{X}/n = \mathbf{I}_k$, the LASSO estimator $\widehat{\beta}_{j,\lambda} = \operatorname{sgn}\left(\widetilde{\beta}_j\right)\left(\left|\widetilde{\beta}_j\right| - \lambda\right)_+$ shrinks the OLS estimator $\widetilde{\beta}_j$ towards zero.
- We can use post-LASSO:
  - Select regressors using LASSO;
  - Regress the dependent variable against regressors that survived LASSO selection (i.e., nonzero LASSO regression coefficients in the first step).
- The post-LASSO procedure uses the first-stage LASSO as a model selection step.

Linear model with high-dimensional controls

- Consider the model:

$$Y_i = \alpha D_i + X_i^\top \beta + U_i,$$

  where $X_i = \left(X_{i,1}, X_{i,2}, ..., X_{i,k}\right)^\top$ and
  - D_i: the main explanatory variable of interest which is always included;
  - X_i: potential control variables which are included to avoid the omitted variable bias.
- When the dimension of X_i is high (possibly $k \approx n$ or even $k > n$), we are forced to do model selection, since otherwise the OLS estimator of α is of low precision (high variance) and can not be computed if $k > n$.
- Under the sparse model assumption $\beta_j \neq 0$ for only a small number of j's, we can use LASSO to select the variables in the list X_i of potential variables and then do post-LASSO.

- Let $\mathcal{A} = \left\{j : \beta_j \neq 0\right\}$ denote the list of relevant controls. Note that $\mathcal{A}$ is unknown.
- Let

$$\left(\widehat{\alpha}_\lambda, \widehat{\beta}_{1,\lambda}, ..., \widehat{\beta}_{k,\lambda}\right) = \underset{a,b_1,...,b_k}{\operatorname{argmin}} \left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - aD_i - \sum_{j=1}^{k} b_j X_{i,j}\right)^2 + \lambda \sum_{j=1}^{k}\left|b_j\right|\right\}.$$

- $\left(\widehat{\alpha}_\lambda, \widehat{\beta}_{1,\lambda}, ..., \widehat{\beta}_{k,\lambda}\right)$ are biased.
- The selected controls are $\widehat{\mathcal{A}} = \left\{j : \widehat{\beta}_{j,\lambda} \neq 0\right\}$. Let $X_{i,\widehat{\mathcal{A}}}$ denote the sub-vector of X_i with only the controls in $\widehat{\mathcal{A}}$. Similarly, $X_{i,\mathcal{A}}$ denotes the vector of controls in $\mathcal{A}$.
- A post-LASSO estimator $\widehat{\alpha}\left(\widehat{\mathcal{A}}\right)$ of α is the OLS regression coefficient of D_i of the regression of Y_i against $\left(D_i, X_{i,\widehat{\mathcal{A}}}\right)$.
- Let $\widehat{\alpha}\left(\mathcal{A}\right)$ denote the oracle estimator when $\mathcal{A}$ is known: the OLS regression coefficient of D_i of the regression of Y_i against $\left(D_i, X_{i,\mathcal{A}}\right)$.

- If we are concerned with only the standard asymptotic normality theory, $\widehat{\alpha}\left(\widehat{\mathcal{A}}\right)$ can be as good as $\widehat{\alpha}\left(\mathcal{A}\right)$.
- $\widehat{\alpha}\left(\mathcal{A}\right)$ is asymptotically normal:
$$\sqrt{n}\left(\widehat{\alpha}\left(\mathcal{A}\right)-\alpha\right)\rightarrow_d \mathrm{N}\left(0,\omega^2\left(\mathcal{A}\right)\right),$$
where $\omega^2\left(\mathcal{A}\right)>0$ denotes the asymptotic variance.
- Under proper choice of the penalty parameter λ, e.g., in a homoskedastic model,
$$\lambda = 2\sigma\sqrt{\frac{2\log\left(kn\right)}{n}},$$
we have model selection consistency: $\Pr\left[\widehat{\mathcal{A}}=\mathcal{A}\right]\rightarrow 1$ as $n\uparrow\infty$.
- We can show that if $\widehat{\mathcal{A}}$ consistently estimates $\mathcal{A}$, where $\widehat{\mathcal{A}}$ is constructed by LASSO or other high-quality model selection procedure (e.g., the square root LASSO), we have the oracle property
$$\sqrt{n}\left(\widehat{\alpha}\left(\widehat{\mathcal{A}}\right)-\alpha\right)\rightarrow_d \mathrm{N}\left(0,\omega^2\left(\mathcal{A}\right)\right).$$

- Can we ignore the error in $\widehat{\mathcal{A}}$ and proceed as if we know the true model $\mathcal{A}$? The oracle property may not be reliable for the purpose of statistical inference on α, in real applications where the sample size n is fixed.
- The oracle property states that $\sqrt{n}\left(\widehat{\alpha}\left(\widehat{\mathcal{A}}\right)-\alpha\right) \overset{a}{\sim} \mathrm{N}\left(0, \omega^2\left(\mathcal{A}\right)\right)$ or $\widehat{\alpha}\left(\widehat{\mathcal{A}}\right) \overset{a}{\sim} \mathrm{N}\left(\alpha, \omega^2\left(\mathcal{A}\right)/n\right)$, when n is large. But in real applications, the exact distribution of $\sqrt{n}\left(\widehat{\alpha}\left(\widehat{\mathcal{A}}\right)-\alpha\right)$ may be very different from $\mathrm{N}\left(0, \omega^2\left(\mathcal{A}\right)\right)$.
- Typically, this happens when some of the true coefficients β are nonzero but close to zero. This is the case when there are many potential controls and some of them have small effects on the explained variable.
- Note the potential conflict: it is hard to shrink regression coefficients of irrelevant regressors to zero (large λ) while detect relevant regressors with small coefficients (small λ) and leave them out.

Problem with small coefficients and naive post-LASSO

- The oracle property is based on the fact of model selection consistency, which requires LASSO to detect the relevant controls with probability approaching one as $n \uparrow \infty$.
- Suppose that $k < n$, $\mathbf{X}^\top \mathbf{X}/n = \mathbf{I}_k$ and $\widehat{\beta}_{j,\lambda} = \operatorname{sgn}\left(\widetilde{\beta}_j\right)\left(\left|\widetilde{\beta}_j\right| - \lambda\right)_+$ with
$$\lambda = 2\sigma\sqrt{\frac{2\log(kn)}{n}}.$$
- We use alternative asymptotic theory as a tool to illustrate the problem. In the asymptotic analysis framework, the magnitude of the coefficient β_j should be made relative to the sample size n. We model "small coefficient" as
$$\beta_j = \frac{c}{\sqrt{n}},$$
where $c \neq 0$ is a constant.
- The notation $\beta_j \propto \xi_n$ means that β_j is equal to a nonzero constant multiplied by ξ_n.

- In the asymptotic analysis framework, we formally take $\beta_j = 0$, $\beta_j \propto n^{-1/2}$ and $\beta_j \propto 1$ as the definitions of zero, small and large coefficients.
- In reality, n is fixed. The assumption $\beta_j = c/\sqrt{n}$ is a tautology: we can always find c such that $\beta_j = c/\sqrt{n}$ holds.
- Under $\beta_j = c/\sqrt{n}$, we may derive different limiting distribution or probability that better approximates the exact distribution or probability. We use this assumption as a tool to illustrate the problem.
- Note that when $\beta_j = 0$,

$$\Pr\left[\widehat{\beta}_{j,\lambda} = 0\right] = \Pr\left[\left|\widetilde{\beta}_j\right| < 2\sigma\sqrt{\frac{2\log(kn)}{n}}\right]$$
$$= \Pr\left[\left|\sqrt{n}\widetilde{\beta}_j\right| < 2\sigma\sqrt{2\log(kn)}\right] \to 1,$$

since $\sqrt{n}\widetilde{\beta}_j$ behaves like a normal random variable when n is large.

- When $\beta_j \neq 0$, since $\left|\widetilde{\beta}_j - \beta_j\right| + \left|\widetilde{\beta}_j\right| \geq \left|\beta_j\right|$ and

$$\frac{2\sigma\sqrt{2\log(kn)} + \left|\sqrt{n}\left(\widetilde{\beta}_j - \beta_j\right)\right|}{\sqrt{n}} \to_p 0,$$

$$\begin{aligned} 0 \leq \Pr\left[\widehat{\beta}_{j,\lambda} = 0\right] &= \Pr\left[\left|\widetilde{\beta}_j\right| < 2\sigma\sqrt{\frac{2\log(kn)}{n}}\right] \\ &\leq \Pr\left[\left|\beta_j\right| < \frac{2\sigma\sqrt{2\log(kn)} + \left|\sqrt{n}\left(\widetilde{\beta}_j - \beta_j\right)\right|}{\sqrt{n}}\right] \to 0. \end{aligned}$$

- LASSO detects a large β_j with high probability.

- However, when β_j is small, it is possible that the exact probability $\Pr\left[\widehat{\beta}_{j,\lambda} = 0\right]$ corresponding to a fixed n is not close to zero, as illustrated by the limit of $\Pr\left[\widehat{\beta}_{j,\lambda} = 0\right]$ with the assumption $\beta_j = c/\sqrt{n}$ imposed: since $\left|\widetilde{\beta}_j - \beta_j\right| + \left|\beta_j\right| \geq \left|\widetilde{\beta}_j\right|$ and $\sqrt{2\log(kn)} \uparrow \infty$,

$$\begin{aligned}
\Pr\left[\widehat{\beta}_{j,\lambda} = 0\right] &= \Pr\left[\left|\widetilde{\beta}_j\right| < 2\sigma\sqrt{\frac{2\log(kn)}{n}}\right] \\
&\geq \Pr\left[\left|\widetilde{\beta}_j - \beta_j\right| + \left|\beta_j\right| < 2\sigma\sqrt{\frac{2\log(kn)}{n}}\right] \\
&\geq \Pr\left[\left|\sqrt{n}\left(\widetilde{\beta}_j - \beta_j\right)\right| < 2\sigma\sqrt{2\log(kn)} - |c|\right] \to 1.
\end{aligned}$$

- When β_j is small, the probability of $\widehat{\beta}_{j,\lambda} = 0$ so that LASSO fails to detect it can be large, since it shows that $\Pr\left[\widehat{\beta}_{j,\lambda} = 0\right]$ can be close to the limit 1 under $\beta_j = c/\sqrt{n}$ rather than 0.

- Consider the simple example $Y_i = \alpha D_i + \beta X_i + U_i$ with a single potential control X_i and a small coefficient β (the true model is $\mathcal{A} = \{X_i\}$).
- Let $\widehat{\mathcal{A}}$ denote the LASSO estimator of $\mathcal{A}$. Then,

$$\widehat{\alpha}\left(\widehat{\mathcal{A}}\right) = 1\left(\widehat{\mathcal{A}} = \emptyset\right)\widehat{\alpha}\left(\emptyset\right) + 1\left(\widehat{\mathcal{A}} = \{X_i\}\right)\widehat{\alpha}\left(\{X_i\}\right).$$

- Suppose that $\beta = c/\sqrt{n}$. With a non-negligible probability in finite samples, LASSO leaves X_i out and estimate $\widehat{\mathcal{A}} = \emptyset$. In this case, there is omitted variable bias. The post-LASSO estimator of α is

$$\widehat{\alpha}\left(\emptyset\right) = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i^2} = \alpha + \beta\frac{\sum_{i=1}^{n} D_i X_i}{\sum_{i=1}^{n} D_i^2} + \frac{\sum_{i=1}^{n} D_i U_i}{\sum_{i=1}^{n} D_i^2}$$

and then

$$\sqrt{n}\left(\widehat{\alpha}\left(\emptyset\right) - \alpha\right) = c\frac{\sum_{i=1}^{n} D_i X_i}{\sum_{i=1}^{n} D_i^2} + \frac{n^{-1/2}\sum_{i=1}^{n} D_i U_i}{n^{-1}\sum_{i=1}^{n} D_i^2}.$$

- Note that
$$\widehat{\rho} = \frac{\sum_{i=1}^{n} D_i X_i}{\sum_{i=1}^{n} D_i^2}$$
is the OLS estimator in the simple regression of X_i against D_i and $\widehat{\rho} \to_p \rho = \mathrm{E}\left[D_i X_i\right] / \mathrm{E}\left[D_i^2\right]$.
- When n is large,
$$\sqrt{n}\left(\widehat{\alpha}\left(\emptyset\right) - \alpha\right) \overset{a}{\sim} \mathrm{N}\left(c\rho, \frac{\mathrm{E}\left[D_i^2 U_i^2\right]}{\left(\mathrm{E}\left[D_i^2\right]\right)^2}\right)$$
$$\Longleftrightarrow \widehat{\alpha}\left(\emptyset\right) \overset{a}{\sim} \mathrm{N}\left(\alpha + c\rho, \frac{\mathrm{E}\left[D_i^2 U_i^2\right]}{n\left(\mathrm{E}\left[D_i^2\right]\right)^2}\right).$$
- The distribution of $\sqrt{n}\left(\widehat{\alpha}\left(\widehat{\mathcal{A}}\right) - \alpha\right)$ is close to a mixture of $\mathrm{N}\left(c\rho, \mathrm{E}\left[D_i^2 U_i^2\right] / \left(\mathrm{E}\left[D_i^2\right]\right)^2\right)$ and the limiting distribution of $\sqrt{n}\left(\widehat{\alpha}\left(\mathcal{A}\right) - \alpha\right)$.

- The the asymptotic bias is $c \times \rho$:
  - c: the coefficient of the omitted control in the linear structural/causal model;
  - ρ: the coefficient of the the omitted control X_i in the linear projection of D_i against X_i, i.e.,

$$\rho = \operatorname*{argmin}_{r \in \mathbb{R}} \mathrm{E}\left[(D_i - rX_i)^2\right].$$

- When ρ is large, the asymptotic bias $c\rho$ of the post LASSO estimator can be substantial.
- If ρ is small (i.e., $\rho \propto n^{-1/2}$) or zero, the asymptotic bias is negligible.

Double LASSO

- The double LASSO procedure of Belloni, Chernozhukov and Hansen (2014): since the bias of the naive post-Lasso depends on the magnitude of the correlation between the main regressor D_i and the controls X_i, one can run LASSO of D_i against X_i to detect correlated controls.
  - Adaptive LASSO tries to simultaneously estimate the causal effects well and identify relevant regressors in the classical low dimensional context;
  - Double LASSO pursues high-quality estimation of the effect of the main regressor with a large number of potential controls but does not pursue precise variable selection for the controls.
- Consider the linear projection model:

$$D_i = \sum_{j=1}^{k} \rho_j X_{i,j} + \eta_i,$$

where $(\rho_1, ..., \rho_k) = \text{argmin}_r \, \mathrm{E}\left[\left(D_i - X_i^\top r\right)^2\right]$ and η_i is defined to be the difference $D_i - \sum_{j=1}^{k} \rho_j X_{i,j}$ so that the equation above holds automatically.

- We run a LASSO regression of D_i against X_i: let

$$\left(\widehat{\rho}_{1,\lambda}, ..., \widehat{\rho}_{k,\lambda}\right) = \underset{b_1,...,b_k}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(D_i - \sum_{j=1}^{k} b_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^{k} \left| b_j \right| \right\}.$$

  - Large $\rho_j \Longrightarrow X_{i,j}$ will be detected by LASSO (assigned a nonzero coefficient);
  - Small or zero $\rho_j \Longrightarrow X_{i,j}$ will be dropped by LASSO (assigned a zero coefficient).
  - We should keep $X_{i,j}$ with large ρ_j for robustness to avoid omitted variable bias.
- We write a reduced-form equation:

$$\begin{aligned} Y_i &= \alpha D_i + \sum_{j=1}^{k} \beta_j X_{i,j} + U_i \\ &= \alpha \left(\sum_{j=1}^{k} \rho_j X_{i,j} + \eta_i \right) + \sum_{j=1}^{k} \beta_j X_{i,j} + U_i = \sum_{j=1}^{k} \pi_j X_{i,j} + \epsilon_i, \end{aligned}$$

where we define $\pi_j = \alpha \rho_j + \beta_j$ and $\epsilon_i = \alpha \eta_i + U_i$.

- We run LASSO regression of Y_i against X_i:

$$\left(\widehat{\pi}_{1,\lambda}, ..., \widehat{\pi}_{k,\lambda}\right) = \underset{b_1,...,b_k}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} b_j X_{i,j} \right)^2 + \lambda \sum_{j=1}^{k} \left| b_j \right| \right\}.$$

  - If ρ_j is small, π_j is large only if β_j is large. $X_{i,j}$ will be detected by LASSO.
  - If ρ_j is small, π_j is small only if β_j is small. $X_{i,j}$ will be dropped by LASSO.
- $X_{i,j}$ is dropped in both LASSO regressions, only if ρ_j is small and β_j is small. In such a case, the bias is negligible.

Double LASSO procedure

1. Run LASSO regression of D_i against X_i. Let $\widehat{\mathcal{A}}_D = \left\{ j : \widehat{\rho}_{j,\lambda} \neq 0 \right\}$ be the selected controls.
2. Run LASSO regression of Y_i against X_i. Let $\widehat{\mathcal{A}}_Y = \left\{ j : \widehat{\pi}_{j,\lambda} \neq 0 \right\}$ be the selected controls.
3. Estimate α by OLS regression of Y_i against D_i and controls in $\widehat{\mathcal{A}}_D \cup \widehat{\mathcal{A}}_Y$.

An alternative method: partialling out

- $\mathbf{X}_{\mathcal{A}}$: the $n \times |\mathcal{A}|$ matrix of observations only on the relevant controls; $\mathbf{D}$: the vector of $(D_1, ..., D_n)^\top$.
- By the partition regression theorem, we have

$$\widehat{\alpha}(\mathcal{A}) = \frac{\mathbf{D}^\top \mathbf{M}_{\mathcal{A}} \mathbf{Y}}{\mathbf{D}^\top \mathbf{M}_{\mathcal{A}} \mathbf{D}} = \frac{\widetilde{\mathbf{D}}^\top \widetilde{\mathbf{Y}}}{\widetilde{\mathbf{D}}^\top \widetilde{\mathbf{D}}},$$

  where $\mathbf{M}_{\mathcal{A}} = \mathbf{I}_n - \mathbf{X}_{\mathcal{A}} \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top$, $\widetilde{\mathbf{Y}} = \mathbf{M}_{\mathcal{A}} \mathbf{Y}$ and $\widetilde{\mathbf{D}} = \mathbf{M}_{\mathcal{A}} \mathbf{D}$.
- $\widetilde{\mathbf{Y}}$ and $\widetilde{\mathbf{D}}$ are regression residuals:

$$\begin{aligned} \widetilde{\mathbf{D}} &= \mathbf{D} - \mathbf{X}_{\mathcal{A}} \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \mathbf{D} \\ \widetilde{\mathbf{Y}} &= \mathbf{Y} - \mathbf{X}_{\mathcal{A}} \left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \mathbf{Y}, \end{aligned}$$

  where $\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \mathbf{D}$ and $\left(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}} \right)^{-1} \mathbf{X}_{\mathcal{A}}^\top \mathbf{Y}$ are OLS coefficients.
- In case of unknown $\mathcal{A}$, we use LASSO and post-LASSO to create residuals.

The partialling out procedure

1. Run LASSO regression of D_i against X_i. Let $\widehat{\mathcal{A}}_D = \left\{ j : \widehat{\rho}_{j,\lambda} \neq 0 \right\}$ be the selected controls.
2. Run post-LASSO of D_i against $X_{i,\widehat{\mathcal{A}}_D}$ and generate the OLS residual $\widetilde{D}_i$.
3. Run LASSO regression of Y_i against X_i. Let $\widehat{\mathcal{A}}_Y = \left\{ j : \widehat{\pi}_{j,\lambda} \neq 0 \right\}$ be the selected controls.
4. Run post-LASSO of Y_i against $X_{i,\widehat{\mathcal{A}}_Y}$ and generate the OLS residual $\widetilde{Y}_i$.
5. Estimate α by the OLS regression of $\widetilde{Y}_i$ against $\widetilde{D}_i$.

Comparison with naive post-LASSO

- Let $\left(\widehat{\alpha}^{\textsf{naive}}, \widehat{\beta}_1^{\textsf{naive}}, ..., \widehat{\beta}_k^{\textsf{naive}}\right)$ denote the naive post-LASSO estimator:

$$\left(\widehat{\alpha}^{\textsf{naive}}, \widehat{\beta}_1^{\textsf{naive}}, ..., \widehat{\beta}_k^{\textsf{naive}}\right)$$

$$= \underset{a,b_1,...,b_k: b_j=0, j\notin\widehat{\mathcal{A}}}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - aD_i - \sum_{j=1}^{k} b_j X_{i,j}\right)^2.$$

- By the first-order condition,

$$\sum_{i=1}^{n} D_i \left(Y_i - \widehat{\alpha}^{\textsf{naive}} D_i - \sum_{j=1}^{k} \widehat{\beta}_j^{\textsf{naive}} X_{i,j}\right) = 0$$

$$\implies \widehat{\alpha}^{\textsf{naive}} = \frac{\sum_{i=1}^{n} D_i \left(Y_i - \sum_{j=1}^{k} \widehat{\beta}_j^{\textsf{naive}} X_{i,j}\right)}{\sum_{i=1}^{n} D_i^2}$$

$$= \frac{\sum_{i=1}^{n} D_i \left(U_i - \sum_{j=1}^{k} \left(\widehat{\beta}_j^{\textsf{naive}} - \beta_j\right) X_{i,j}\right)}{\sum_{i=1}^{n} D_i^2}.$$

- Then,

$$\sqrt{n}\left(\widehat{\alpha}^{\mathsf{naive}} - \alpha\right) = \frac{n^{-1/2}\sum_{i=1}^{n} D_i U_i}{n^{-1}\sum_{i=1}^{n} D_i^2} + \frac{1}{n^{-1}\sum_{i=1}^{n} D_i^2}\sum_{j=1}^{k}\sqrt{n}\left(\widehat{\beta}_j^{\mathsf{naive}} - \beta_j\right)\left(\frac{1}{n}\sum_{i=1}^{n} D_i X_{i,j}\right).$$

- If $\beta_j = c/\sqrt{n}$, LASSO drops $X_{i,j}$ so that $j \notin \widehat{\mathcal{A}}$ and $\widehat{\beta}_j^{\mathsf{naive}} = 0$. Then,

$$\sqrt{n}\left(\widehat{\beta}_j^{\mathsf{naive}} - \beta_j\right)\left(\frac{1}{n}\sum_{i=1}^{n} D_i X_{i,j}\right) = -c\left(n^{-1}\sum_{i=1}^{n} D_i X_{i,j}\right)$$

and $n^{-1}\sum_{i=1}^{n} D_i X_{i,j} \to_p \mathrm{E}\left[D_i X_{i,j}\right]$, which is the source of the asymptotic bias when $\mathrm{E}\left[D_i X_{i,j}\right] \neq 0$.

- Let $\widetilde{\beta}^{\mathsf{po}}$ denote the post-LASSO OLS estimator of Y_i against $X_{i,\widehat{\mathcal{A}}_Y}$:

$$\left(\widetilde{\beta}_1^{\mathsf{po}}, ..., \widetilde{\beta}_k^{\mathsf{po}}\right) = \underset{b_1,...,b_k : b_j = 0, j \notin \widehat{\mathcal{A}}_Y}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} b_j X_{i,j}\right)^2 .$$

  and let $\widehat{\alpha}^{\mathsf{po}}$ denote the OLS estimator of $\widetilde{Y}_i$ against $\widetilde{D}_i$.
- Then,

$$\sqrt{n}\left(\widehat{\alpha}^{\mathsf{po}} - \alpha\right) = \frac{n^{-1/2} \sum_{i=1}^{n} \widetilde{D}_i U_i}{n^{-1} \sum_{i=1}^{n} \widetilde{D}_i^2} + \frac{1}{n^{-1} \sum_{i=1}^{n} \widetilde{D}_i^2} \sum_{j=1}^{k} \sqrt{n}\left(\widetilde{\beta}_j^{\mathsf{po}} - \beta_j\right) \left(\frac{1}{n} \sum_{i=1}^{n} \widetilde{D}_i X_{i,j}\right).$$

- If β_j is small, $\widetilde{\beta}_j^{\mathsf{po}}$ is constrained to be zero if and only if ρ_j is small or zero. In this case, it can be shown that $\sum_{i=1}^n \widetilde{D}_i X_{i,j} / \sum_{i=1}^n \widetilde{D}_i^2$ is negligible.
  - For example, in the case of $k = 1$, if ρ is small or zero, the first step LASSO drops X_i and $\widetilde{D}_i = D_i$. Then,

$$\frac{\sum_{i=1}^n \widetilde{D}_i X_i}{\sum_{i=1}^n \widetilde{D}_i^2} = \frac{\sum_{i=1}^n D_i X_i}{\sum_{i=1}^n D_i^2} \approx \frac{\mathrm{E}\left[D_i X_i\right]}{\mathrm{E}\left[D_i^2\right]} = \rho.$$

- If β_j is large, $\widetilde{\beta}_j^{\mathsf{po}}$ is constrained to be zero if and only if ρ_j is large. In this case, $X_{i,j}$ is selected in the first step ($\widehat{\rho}_{j,\lambda} \neq 0$) with high probability and by construction, $\sum_{i=1}^n \widetilde{D}_i X_{i,j} = 0$.

Comparison with double LASSO

- By the partition regression theorem, the double LASSO estimator is
$$\widehat{\alpha}^{\mathsf{dl}} = \frac{\sum_{i=1}^{n} \ddot{D}_i \ddot{Y}_i}{\sum_{i=1}^{n} \ddot{D}_i^2},$$
where $\ddot{D}_i$ and $\ddot{Y}_i$ are regression residuals from OLS regressions of D_i and Y_i against controls in $\widehat{\mathcal{A}}_D \cup \widehat{\mathcal{A}}_Y$.
- Partialling out:
$$\widehat{\alpha}^{\mathsf{po}} = \frac{\sum_{i=1}^{n} \widetilde{D}_i \widetilde{Y}_i}{\sum_{i=1}^{n} \widetilde{D}_i^2},$$
where $\widetilde{D}_i$ and $\widetilde{Y}_i$ may be constructed using different controls, since in general $\widehat{\mathcal{A}}_D \neq \widehat{\mathcal{A}}_Y$.
- Double LASSO is more conservative, since more controls are used to construct residuals.

Standard errors

- The asymptotic variance of $\widehat{\alpha}^{\mathsf{dl}}$ is

$$\sigma^2 = \frac{\mathrm{E}\left[\left(Y_i - \alpha D_i - X_i^\top \beta\right)^2 \left(D_i - X_i^\top \rho\right)^2\right]}{\left(\mathrm{E}\left[\left(D_i - X_i^\top \rho\right)^2\right]\right)^2},$$

  i.e., $\sqrt{n}\left(\widehat{\alpha}^{\mathsf{dl}} - \alpha\right) \to_d \mathrm{N}\left(0, \sigma^2\right)$.
- The standard error $\widehat{\sigma}/\sqrt{n}$ can be constructed by replacing α, β, ρ with their post-LASSO estimators $\left(\widehat{\alpha}^{\mathsf{pl}}, \widehat{\beta}^{\mathsf{pl}}, \widehat{\rho}^{\mathsf{pl}}\right)$:

$$\sigma^2 = \frac{n^{-1}\sum_{i=1}^n \left(Y_i - \widehat{\alpha}^{\mathsf{pl}} D_i - X_i^\top \widehat{\beta}^{\mathsf{pl}}\right)^2 \left(D_i - X_i^\top \widehat{\rho}^{\mathsf{pl}}\right)^2}{\left(n^{-1}\sum_{i=1}^n \left(D_i - X_i^\top \widehat{\rho}^{\mathsf{pl}}\right)^2\right)^2}.$$