# 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 1, 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 $\alpha$ 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 $\alpha$ 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 $\lambda$, 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 $\alpha$, 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 $\beta$ 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 $\lambda$) while detect relevant regressors with small coefficients (small $\lambda$) 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 $\beta_j$ 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.
- In the asymptotic analysis framework, we formally take $\beta_j = 0$, $\beta_j = c/\sqrt{n}$ and $\beta_j \neq 0$ 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 $\beta_j$ with high probability.

- However, when $\beta_j$ 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 $\beta_j$ 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 $\beta$ (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 $\alpha$ 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,

$$\sqrt{n}\left(\widehat{\alpha}\left(\emptyset\right) - \alpha\right) \overset{a}{\sim} \mathrm{N}\left(c\rho, \mathrm{E}\left[D_i^2 U_i^2\right] / \left(\mathrm{E}\left[D_i^2\right]\right)^2\right)$$

  and 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 $\sqrt{n}\left(\widehat{\alpha}\left(\mathcal{A}\right) - \alpha\right)$.
- When $\rho$ is large, the post LASSO estimator can be substantially biased.