Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 11: LASSO for High-dimensional Sparse Linear Models

Instructor: Ma, Jun

Renmin University of China

December 1, 2021

High-dimensional sparse models

- In the model

$$Y_i = \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + U_i,$$

  the number of potential regressors k can be of comparable order to the sample size n. In some applications, k can be larger than n.
- Statistical analysis of high-dimensional models abandons the assumption that $n \uparrow \infty$ but k is fixed. Instead we assume that $k \uparrow \infty$.
- In a sparse model, only a few regressors have non-zero coefficients.
- Such statistical analysis requires advanced mathematical tools. We present one of the most basic results.

- The list of non-zero coefficients is $\mathcal{A} = \left\{ j : \beta_j \neq 0 \right\}$.
- The $\mathscr{L}^0$ norm: $\|\beta\|_0 = |\mathcal{A}|$, where $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$.
- The simplest sparse model assumption is that $\|\beta\|_0$ is a fixed number, although $n, k \uparrow \infty$.
- Note that in the following statistical analysis, we do not treat LASSO as an algorithm for high-performance out-of-sample prediction. Our objective is to see what selection rule for the penalty parameter λ results in high-quality estimation of the parameters β.

Performance of LASSO

- Assume the model is homoskedastic: $\mathrm{E}\left[U_i^2 \mid \mathbf{X}\right] = \sigma^2$.
- We consider the following measure of distance between b and β:

$$\frac{1}{n}\left\|\mathbf{X}\left(b-\beta\right)\right\|^2 = \left(b-\beta\right)^\top \left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\left(b-\beta\right),$$

  which is like a weighted $\mathscr{L}^2$ norm.
- If you know the identities of the zero coefficients ($\mathcal{A}$), the oracle estimator can be computed:

$$\widehat{\beta}_{\textsf{oracle}} = \underset{b\in\mathbb{R}^k, b_j=0, j\in\mathcal{A}^c}{\operatorname{argmin}} \left\|\mathbf{Y}-\mathbf{X}b\right\|^2,$$

  where "$b_j = 0, j \in \mathcal{A}^c$" ($\mathcal{A}^c = \left\{j : \beta_j = 0\right\}$) is a constraint such that all out-of-$\mathcal{A}$ coordinates of b are constrained to be zero.

- It is easy to see that

$$\mathrm{E}\left[\frac{1}{n}\left\|\mathbf{X}\left(\widehat{\beta}_{\mathsf{oracle}}-\beta\right)\right\|^2\right]=\sigma^2\frac{\|\beta\|_0}{n}.$$

  $n^{-1}\left\|\mathbf{X}\left(\widehat{\beta}_{\mathsf{oracle}}-\beta\right)\right\|^2$ behaves like a stochastic sequence of order n^{-1} and $\left\|\widehat{\beta}_{\mathsf{oracle}}-\beta\right\|$ is of order $n^{-1/2}$.
- The LASSO:

$$\widehat{\beta}_\lambda=\underset{b\in\mathbb{R}^k}{\operatorname{argmin}}\frac{1}{n}\|\mathbf{Y}-\mathbf{X}b\|^2+\lambda\|b\|_1.$$

- We can show that if λ is properly chosen so that large enough penalty is imposed, $n^{-1}\left\|\mathbf{X}\left(\widehat{\beta}_\lambda-\beta\right)\right\|^2$ behaves like a stochastic sequence of order $\log(k)/n$ and $\left\|\widehat{\beta}_\lambda-\beta\right\|_1$ is like $\sqrt{\log(k)/n}$.
- Price of not knowing $\mathcal{A}$ is a $\log(k)$ loss in convergence speed.
- No other procedure achieves faster convergence speed without requiring knowledge of $\mathcal{A}$.

Consistency of LASSO and rate of λ

- We sketch an even weaker result: consistency of LASSO and the required rate for λ.
- Remember the matrix form of the model $\mathbf{Y} = \mathbf{X}\beta + \mathbf{U}$. We can show

$$\frac{1}{n}\left\|\mathbf{X}\left(\widehat{\beta}_\lambda - \beta\right)\right\|^2 + \lambda\left\|\widehat{\beta}_\lambda\right\|_1 \leq 2\frac{\mathbf{U}^\top\mathbf{X}}{n}\left(\widehat{\beta}_\lambda - \beta\right) + \lambda\|\beta\|_1 .$$

- Then,

$$\left|\frac{\mathbf{U}^\top\mathbf{X}}{n}\left(\widehat{\beta}_\lambda - \beta\right)\right| \leq 2\cdot\left(\max_{1\leq j\leq k}\left|\frac{1}{n}\sum_{i=1}^{n} U_i X_{i,j}\right|\right)\left\|\widehat{\beta}_\lambda - \beta\right\|_1 .$$

- If λ dominates the noise $\lambda > 2\left(\max_{1\leq j\leq k}\left|\frac{1}{n}\sum_{i=1}^{n} U_i X_{i,j}\right|\right)$ with high probability, then

$$\frac{1}{n}\left\|\mathbf{X}\left(\widehat{\beta}_\lambda - \beta\right)\right\|^2 \leq 2\lambda\|\beta\|_1 ,$$

  with high probability.

- If $\lambda \downarrow 0$ as $n \uparrow \infty$ and at the same time $\lambda > 2\left(\max_{1 \leq j \leq k} \left|\frac{1}{n}\sum_{i=1}^{n} U_i X_{i,j}\right|\right)$ with high probability, we have consistency $n^{-1}\left\|\mathbf{X}\left(\widehat{\beta}_\lambda - \beta\right)\right\|^2 \to_p 0$.
- Assume that the regressors are normalized so that $n^{-1}\sum_{i=1}^{n} X_{i,j}^2 = 1$. By CLT, $n^{-1/2}\sum_{i=1}^{n} U_i X_{i,j} \overset{a}{\sim} \mathrm{N}\left(0, \sigma^2\right)$. So if n is large enough, $n^{-1/2}\sum_{i=1}^{n} U_i X_{i,j}$ behaves like an $\mathrm{N}\left(0, \sigma^2\right)$ random variable.
- In general, if $n^{-1}\sum_{i=1}^{n} X_{i,j}^2 \neq 1$, we use weighted LASSO: the penalty term is $\lambda \sum_{j=1}^{k} w_j \left|b_j\right|$ with $w_j = \sqrt{n^{-1}\sum_{i=1}^{n} X_{i,j}^2}$.

- $\xi_1, \xi_2, ..., \xi_k$ are $\mathrm{N}\left(0, \sigma^2\right)$ random variables, then $\mathrm{E}\left[\max_{1 \le i \le k} |\xi_i|\right] \le \sqrt{2\sigma^2 \log(2k)}$. The maximum of k normal random variables with zero mean and variance σ^2 diverges to ∞ at the speed $\sqrt{\log(k)}$.
- Therefore, $\max_{1 \le j \le k} \left| n^{-1/2} \sum_{i=1}^{n} U_i X_{i,j} \right|$ is stochastically bounded by $\sqrt{2\sigma^2 \log(2k)}$, or

$$\frac{\max\limits_{1 \le j \le k} \left| n^{-1/2} \sum_{i=1}^{n} U_i X_{i,j} \right|}{\sqrt{2\sigma^2 \log(2k)}} = O_p(1).$$

- When the number of regressors is large, the penalty parameter λ needs to be adjusted by including $\sqrt{\log(k)}$ and σ^2.
- We choose the penalty parameter λ to be slightly dominating the noise component $2\left(\max_{1 \le j \le k} \left| \frac{1}{n} \sum_{i=1}^{n} U_i X_{i,j} \right|\right)$, since large λ results in a heavily constrained model and higher bias.

- We can choose the penalty parameter as

$$\lambda = 2\sigma\sqrt{\frac{2\log(kn)}{n}}.$$

- Then,

$$\begin{aligned}
&\Pr\left[2\left(\max_{1\le j\le k}\left|\frac{1}{n}\sum_{i=1}^{n}U_iX_{i,j}\right|\right) < \lambda\right] \\
&\quad = \Pr\left[\max_{1\le j\le k}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_iX_{i,j}\right| < \sqrt{2\sigma^2\log(kn)}\right] \\
&\quad = \Pr\left[\frac{\max_{1\le j\le k}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_iX_{i,j}\right|}{\sqrt{2\sigma^2\log(2k)}} < \frac{\sqrt{\log(kn)}}{\sqrt{\log(2k)}}\right] \to 1.
\end{aligned}$$

- With the same choice of λ, $n^{-1}\left\|\mathbf{X}\left(\widehat{\beta}_\lambda - \beta\right)\right\|^2$ converges to zero at the speed $\log(k)/n$.

Square root LASSO

- The penalty parameter λ needs to be adjusted for the variance σ^2 of the error term.
- Estimation of σ^2 can be difficult if $k > n$.
- Belloni, Chernozhukov and Wang (2011) proposed a modified LASSO procedure that removes the dependence on σ^2.
- The LASSO problem can be written as

$$\begin{aligned} \widehat{\beta}_\lambda &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n} RSS(b) + \lambda \|b\|_1 \\ RSS(b) &= \|\mathbf{Y} - \mathbf{X}b\|^2 . \end{aligned}$$

- It is easy to check:

$$\begin{pmatrix} \frac{1}{2}\frac{\partial n^{-1} RSS(\beta)}{\partial b_1} \\ \vdots \\ \frac{1}{2}\frac{\partial n^{-1} RSS(\beta)}{\partial b_k} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n U_i X_{i,1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n U_i X_{i,k} \end{pmatrix}.$$

- We should choose λ to dominate $\max_{1\leq j\leq k}\left|\partial n^{-1}RSS\left(\beta\right)/\partial b_j\right|$, the order of which depends on σ^2.
- Consider the square root LASSO:

$$\widehat{\beta}_{\lambda}^{\mathsf{SR}} = \underset{b\in\mathbb{R}^k}{\operatorname{argmin}}\sqrt{\frac{1}{n}RSS\left(b\right)} + \lambda\left\|b\right\|_1 .$$

- Then,

$$\begin{pmatrix} \frac{\partial\sqrt{n^{-1}RSS(\beta)}}{\partial b_1} \\ \vdots \\ \frac{\partial\sqrt{n^{-1}RSS(\beta)}}{\partial b_k} \end{pmatrix} = \begin{pmatrix} \frac{1}{2\sqrt{n^{-1}RSS(\beta)}}\frac{\partial n^{-1}RSS(\beta)}{\partial b_1} \\ \vdots \\ \frac{1}{2\sqrt{n^{-1}RSS(\beta)}}\frac{\partial n^{-1}RSS(\beta)}{\partial b_k} \end{pmatrix} = \begin{pmatrix} \frac{n^{-1}\sum_{i=1}^n U_i X_{i,1}}{\sqrt{n^{-1}\sum_{i=1}^n U_i^2}} \\ \vdots \\ \frac{n^{-1}\sum_{i=1}^n U_i X_{i,k}}{\sqrt{n^{-1}\sum_{i=1}^n U_i^2}} \end{pmatrix}$$

and $n^{-1}\sum_{i=1}^n U_i^2 \rightarrow_p \sigma^2$.

- Now
$$\frac{n^{-1}\sum_{i=1}^{n}U_iX_{i,j}}{\sqrt{n^{-1}\sum_{i=1}^{n}U_i^2}} \approx \frac{1}{n}\sum_{i=1}^{n}\frac{U_i}{\sigma}X_{i,j}$$
and $n^{-1/2}\sum_{i=1}^{n}(U_i/\sigma)\,X_{i,j} \to_d \mathrm{N}(0,1)$.
- For the square root LASSO, we can choose the penalty term as
$$\lambda = \sqrt{\frac{2\log(kn)}{n}},$$
which dominates $\max_{1\leq j\leq k}\left|\partial\sqrt{n^{-1}RSS(\beta)}/\partial b_j\right|$ and is independent from σ^2.