Section 3: Ridge, LASSO, and Post Double Selection

Ruofan Ma

Gov2018 2024 Spring

February 14, 2024

Motivation

- Why do we need Ridge/LASSO regression?
  - Penalize model overfits (bias-variance tradeoff)
  - Facilitate model selection
- What is double post selection, and why do we want it?
  - In big-data settings (i.e., $p >> n$), we need to explicitly consider model selection to select the most relevant controls
  - This is also one of the motivations for LASSO regression
  - However, things can go very wrong if selection is done improperly

Roadmap for today

- High-altitude review of Ridge and LASSO
  - Why do we need them? What do they do?
  - In what way are they similar to each other? In what way are they different?
- Double post selection (Belloni, Chernozhukov, Hansen)
  - What's wrong with single selection?
  - What's different about double selection?
- Some coding exercise (`Rmd` file available on course website)

Regularization as optimization

- Recall from last week's lecture that

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} := \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} := \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}$$

- Notice that $\hat{\boldsymbol{\beta}}^{\text{ridge}}$ has an analytical solution

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} := \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

- Whereas there is no closed-form solution for LASSO estimators, which are usually obtained through optimization methods like gradient descent. Why? (Hint: Think about the geometry of LASSO)

Bias-Variance Tradeoff

- For both ridge and LASSO regression, we deliberately induce bias to trade in a more robust (lower variance, less prone to overfit) estimator. Using the ridge estimator as an example, when $\lambda > 0$:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}} \mid \mathbf{X}\right] - \boldsymbol{\beta} = \left[\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}\right)^{-1} - \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} > 0$$

$$\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] - \mathbb{V}\left[\hat{\boldsymbol{\beta}}^{\text{ridge}} \mid \mathbf{X}\right] = \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}\right)^{-1} \left\{2\lambda\mathbb{I} + \lambda^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right\}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}\right)^{-1} > 0$$

- What about MSE?
  - Theorem (Theobald 1974): There always exists a value of λ such that the ridge estimator has lower MSE than the OLS estimator.

L_p Regularization

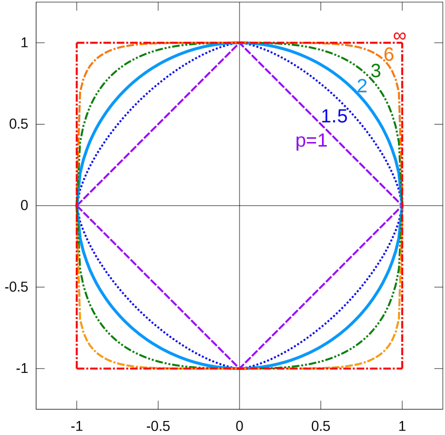

Figure: Behavior of $||\cdot||_p$ penalization term (on the unit circle)

How do things change as p changes?

- Recall that L_p assigns increasing importance to the more extreme entries in the matrix/vector as p increases:

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p\right)^{1/p}$$

$$\hat{\boldsymbol{\beta}}^{\ell_p} := \arg\min_{\boldsymbol{\beta}} \left\{\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\boldsymbol{\beta}\|_p^p\right\}$$

- $p \to \infty$: Norm measures largest absolute entry, $\|\boldsymbol{\beta}\|_\infty = \max_j \|\beta_j\|$
- $p > 2$: Norm focuses on large entries
- $p = 2$: Large entries are expensive; encourages similar-size entries
- $p = 1$: Encourages sparsity
- $p \to 0$: Simply records whether an entry is non-zero, $\|\boldsymbol{\beta}\|_0 = \sum_j \mathbb{I}\{\beta_j \neq 0\}$

Single selection and why it is evil

- (Slides acknowledgment: Victor Chernozhukov)
- Suppose we want to do model selection on the following simple endogenous model:

$$y_i = d_i\alpha + x_i\beta + \varepsilon_i, \quad d_i = x_i\gamma + v_i$$

- A common practice is to do the following post single selection procedure:
  1. Include x_i only if it is a significant predictor of y_i (as judged by, for example, Lasso). Drop it otherwise.
  2. Refit the model after selection. Report standard confidence intervals.

Remark (BCH 2010)

This can bias our causal estimation if $|\beta|$ is close to zero but not equal to zero. Formally if:

$$|\beta| \propto 1/\sqrt{n}$$

Intuition: Omitted Variable Bias

- What went wrong?
    - Distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$ is not what you might think!
    - If we drop x_i, and only regress y_i on d_i, then:

$$\sqrt{n}(\hat{\alpha} - \alpha) \dot{\sim} \underbrace{\text{good term}}_{\text{asympt. normal}} + \underbrace{\sqrt{n}\left(\mathbf{D}^\top \mathbf{D}/n\right)^{-1}\left(\mathbf{X}^\top \mathbf{X}/n\right)(\gamma\beta)}_{\text{OVB}}$$

- Therefore, to get rid of OVB, we want $\sqrt{n}\gamma\beta \to 0$ even as n increases at a certain rate
    - single selection can drop x_i only if $\beta = O(\sqrt{1/n})$. But this condition gives $\sqrt{n}\gamma\sqrt{1/n} \not\to 0$
    - double selection can drop x_i if (i) both $\beta = O(\sqrt{1/n})$ and $\gamma = O(\sqrt{1/n})$; or (ii) $\sqrt{n}\gamma\beta = O(1/\sqrt{n}) \to 0$, which is what we want!
    - Intuition: β needs to be much smaller to be dropped and not create bias than single selection would think

What is different about post double selection?

Algorithm: Pose Double Selection

1. Include x_i if it is a significant predictor of y_i as judged by LASSO
2. Include x_i if it is a significant predictor of d_i as judged by LASSO. [For example, IV models must include x_i if it is a significant predictor of z_i]
3. Refit the model after selection, use standard confidence intervals.

Theorem (BCH 2010, 2013)

Double selection works in low-dimensional settings and in high-dimensional approximately sparse settings.

- **tl;dr**: Under some conditions, double post selection works

More intuition behind post double selection

- Selection among controls x_i that predict either d_i or y_i is what creates the robustness; it finds controls whose omission would lead to a **large** OVB and includes them in the regression.
- The procedure is a model selection version of Frisch-Waugh-Lovell partialling-out procedure for estimating linear regression. (**cf.** Cinelli and Hazlet 2020)
- Double selection is robust to moderate selection mistakes in the two selection steps.
- Now, codes.

Coding Exercise

- Two options:
  - Part 1: Application of LASSO
    - How it works in R, how does it compare to OLS, how to evaluate model performance
  - Part 2: Application of Post Double Selection
    - How to simulate data based on prior knowledge of DGF, how double selection perform (vis-à-vis single selection), Monte Carlo simulation, bias-variance tradeoff.