Mathematical techniques in data science

Lecture 14: Model consistency of the lasso estimator

March 20th, 2019

Schedule

Week	Chapter
1	Chapter 2: Intro to statistical learning
3	Chapter 4: Classification
4	Chapter 9: Support vector machine and kernels
5, 6	Chapter 3: Linear regression
7	Chapter 8: Tree-based methods + Random forest
8	
9	Neural network
12	PCA $\rightarrow$ Manifold learning
11	Clustering: K-means $\rightarrow$ Spectral Clustering
10	Bootstrap + Bayesian methods + UQ
13	Reinforcement learning/Online learning/Active learning
14	Project presentation

Model selection consistency lasso

- Note: Model consistency of lasso
- Further readings:
  - Zhao and Yu (2006)
  - Wainright (2009)
  - Sparsity, the lasso, and friends (Ryan Tibshirani)

Settings

- We start with the simple linear regression problem

$$Y = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Sparsity: assume that the data is generated using the "true" vector of parameters $\beta^* = (\beta_1^*, 0)$.
- We assume that $E[X^{(1)}] = E[X^{(2)}] = 0$.

Matrix form

- we observe a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- use the same notations as in the previous lectures

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} \\ \ldots & \ldots \\ x_n^{(1)} & x_n^{(2)} \end{bmatrix}$$

Goal

The lasso estimator solves the optimization problem

$$\hat{\beta} = \min_{\beta} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta_1| + |\beta_2|).$$

We want to investigate the conditions under which we can verify that

$$sign(\hat{\beta}_1) = sign(\beta_1^*) \quad \text{and} \quad \hat{\beta}_2 = 0$$

Properties of lasso solutions

Theorem

(a) *A vector* $\hat{\beta}$ *solve the lasso program if and only if there exists a* $\hat{z} \in \partial \|\hat{\beta}\|$ *such that*

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) - \lambda \hat{z} = 0 \tag{0.1}$$

(b) *Suppose that the subgradient vector satisfies the strict dual feasibility condition*

$$|\hat{z}_2| < 1$$

then **any** *lasso solution* $\tilde{\beta}$ *satisfies* $\tilde{\beta}_2 = 0$.

(c) *Under the condition of part (b), if* $\mathbf{X}^{(1)} \neq 0$, *then* $\hat{\beta}$ *is the unique lasso solution.*

The primal-dual witness method.

The primal-dual witness (PDW) method consists of constructing a pair of $(\tilde{\beta}, \tilde{z})$ according to the following steps:

- First, we obtain $\tilde{\beta}_1$ by solving the restricted lasso problem

$$\tilde{\beta}_1 = \min_{\beta=(\beta_1,0)} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda(|\beta_1|).$$

  Choose a subgradient $\tilde{z}_1 \in \mathbb{R}$ for the ℓ_1-norm evaluated at $\tilde{\beta}_1$
- Second, we solve for a vector $\tilde{z}_2$ satisfying equation (0.1), and check whether or not the dual feasibility condition $|\tilde{z}_2| < 1$ is satisfied
- Third, we check whether the *sign consistency condition*

$$\tilde{z}_1 = sign(\beta_1^*)$$

  is satisfied.

A more detailed computation

We note that the matrix form of equation (0.1) can be written as

$$[\mathbf{X}^{(1)}]^T(\mathbf{Y} - \mathbf{X}^{(1)}\beta_1 - \mathbf{X}^{(2)}\beta_2) - \lambda\hat{z}_1 = 0$$

$$[\mathbf{X}^{(2)}]^T(\mathbf{Y} - \mathbf{X}^{(1)}\beta_1 - \mathbf{X}^{(2)}\beta_2) - \lambda\hat{z}_2 = 0$$

To simplify the notation, we denote

$$C_{ij} = [\mathbf{X}^{(i)}]^T[\mathbf{X}^{(j)}]$$

Step 1

- we find $\tilde{\beta}_1$ and $\tilde{z}_1$ that satisfies

$$[\mathbf{X}^{(1)}]^T(\mathbf{Y} - \mathbf{X}^{(1)}\tilde{\beta}_1) - \lambda\tilde{z}_1 = 0$$

- Moreover, to make sure that the sign consistency in Step 3 is satisfied, we impose that

$$\tilde{z}_1 = sign(\beta_1^*) \quad \text{and} \quad \tilde{\beta}_1 = C_{11}^{-1}([\mathbf{X}^{(1)}]^T\mathbf{Y} - \lambda sign(\beta_1^*)).$$

  This is acceptable as long as $\tilde{z}_1 \in \partial|\tilde{\beta}_1|$. That is,

$$sign(\tilde{\beta}_1) = sign(\beta_1^*)$$

Step 2

- Step 2:
$$[\mathbf{X}^{(2)}]^T(\mathbf{Y}-\mathbf{X}^{(1)}\tilde{\beta}_1)-\lambda\tilde{z}_2=0$$
- Choose
$$\tilde{z}_2=\frac{1}{\lambda}[\mathbf{X}^{(2)}]^T(\mathbf{Y}-\mathbf{X}^{(1)}\tilde{\beta}_1).$$

  We want $|\tilde{z}_2|<1$.

Conditions

In principle, we want two conditions:

- $sign(\beta_1^*) = sign(\beta_1^* + \Delta)$
  where
  $$\Delta = C_{11}^{-1}([\mathbf{X}^{(1)}]^T \epsilon - \lambda sign(\beta_1^*)))$$
- $|\tilde{z}_2| < 1$ where
  $$\tilde{z}_2 = \frac{1}{\lambda}[\mathbf{X}^{(2)}]^T(\mathbf{X}^{(1)}\Delta + \epsilon)$$

Zero-noise setting

- we assume that the observations are collected with no noise ($\epsilon = 0$).
- Then

$$\Delta = -C_{11}^{-1} \lambda sign(\beta_1^*)$$

and

$$\tilde{z}_2 = \frac{-1}{\lambda} C_{21} \Delta = C_{21} C_{11}^{-1} sign(\beta_1^*)$$

- Conditions
  - Mutual incoherence: $|C_{21} C_{11}^{-1}| < 1$.
  - Minimum signal: $|\beta_1^*| > \lambda C_{11}^{-1}$

Co-linearity

- Mutual incoherence: $|C_{21}C_{11}^{-1}| < 1$.
- Recall that

$$C_{12} = [\mathbf{X}^{(1)}]^T[\mathbf{X}^{(2)}] = \sum_i x_i^{(1)} x_i^{(2)}$$

- When n is large

$$\frac{1}{n}C_{12} \to E\left([X^{(1)}]^T[X^{(2)}]\right) = Cov(X^{(1)}, X^{(2)})$$

since $E[X^{(1)}] = E[X^{(2)}] = 0$.

Conditions

- Mutual incoherence: $|C_{21}C_{11}^{-1}| < 1$.
  The condition roughly means that the covariance between the variables $X^{(1)}$ and $X^{(2)}$ are less than the variance of $X^{(1)}$
- Minimum signal: $|\beta_1^*| > \lambda C_{11}^{-1}$
  Since
  $$\frac{1}{n}C_{11} \to Var(X^{(1)}),$$
  this means that when $n \to \infty$, we needs
  $$\frac{\lambda_n}{n} \to 0$$

Lasso: alternative form

Alternative form of lasso (using the Lagrangian and min-max argument)

$$\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$
$$\text{subject to } \sum_{j=1}^{p} |\beta^{(j)}| \leq s$$

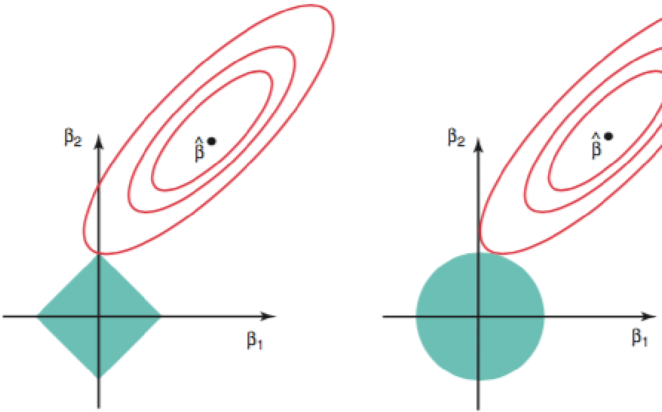
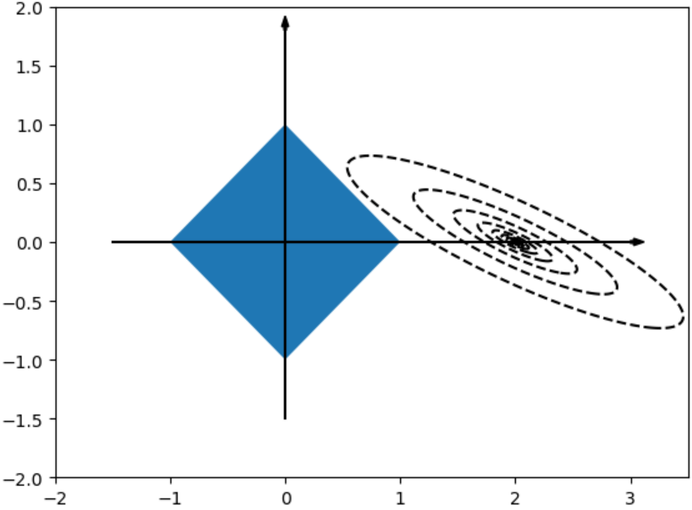
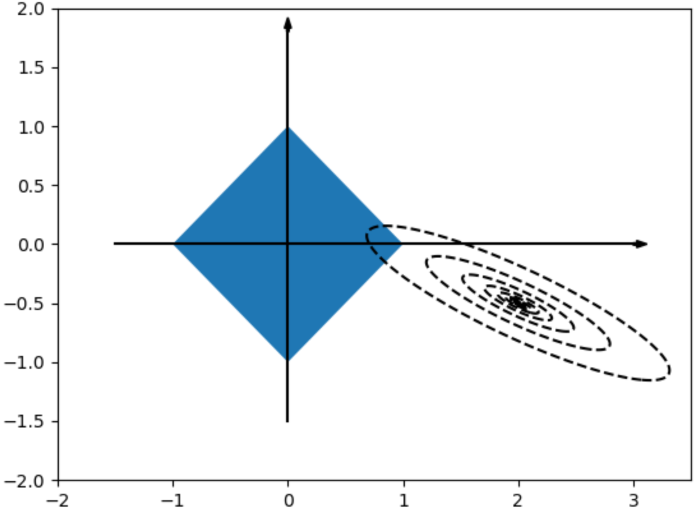


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

When the lasso fails



When the lasso fails



Noisy setting

In principle, we want two conditions:

- $sign(\beta_1^*) = sign(\beta_1^* + \Delta)$
  where
  $$\Delta = C_{11}^{-1}([\mathbf{X}^{(1)}]^T \epsilon - \lambda sign(\beta_1^*))$$
- $|\tilde{z}_2| < 1$ where
  $$\tilde{z}_2 = \frac{1}{\lambda}[\mathbf{X}^{(2)}]^T (\mathbf{X}^{(1)} \Delta + \epsilon)$$
- We want an upper bound on
  $$[\mathbf{X}^{(1)}]^T \epsilon \quad \text{and} [\mathbf{X}^{(2)}]^T \epsilon$$

Properties of Gaussian random variables

In principle, we want two conditions:

- $[\mathbf{X}^{(1)}]^T \epsilon$ is a Gaussian random variable with mean 0 and standard deviation $\sigma \|\mathbf{X}^{(1)}\|_2$
- Thus, there exists a universal constant C such that

$$|[\mathbf{X}^{(1)}]^T \epsilon| \leq C\sigma \sqrt{n Var(X^{(1)}) \log\left(\frac{1}{\delta}\right)}$$

with probability at least $1 - \delta$

General settings

Without loss of generality, assume $\beta^n = (\beta_1^n, ..., \beta_q^n, \beta_{q+1}^n, ... \beta_p^n)^T$ where $\beta_j^n \neq 0$ for $j = 1, .., q$ and $\beta_j^n = 0$ for $j = q+1, ..., p$. Let $\beta_{(1)}^n = (\beta_1^n, ..., \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, ..., \beta_p^n)$. Now write $\mathbf{X_n}(1)$ and $\mathbf{X_n}(2)$ as the first q and last $p-q$ columns of $\mathbf{X_n}$ respectively and let $C^n = \frac{1}{n}\mathbf{X_n}^T\mathbf{X_n}$. By setting $C_{11}^n = \frac{1}{n}\mathbf{X_n}(1)'\mathbf{X_n}(1)$, $C_{22}^n = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(2)$, $C_{12}^n = \frac{1}{n}\mathbf{X_n}(\mathbf{1})'\mathbf{X_n}(\mathbf{2})$ and $C_{21}^n = \frac{1}{n}\mathbf{X_n}(2)'\mathbf{X_n}(1)$. C^n can then be expressed in a block-wise form as follows:

$$C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}.$$

Assuming C_{11}^n is invertible, we define the following Irrepresentable Conditions

Strong Irrepresentable Condition. There exists a positive constant vector η

$$|C_{21}^n (C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)}^n)| \leq \mathbf{1} - \eta,$$

where $\mathbf{1}$ is a $p-q$ by 1 vector of 1's and the inequality holds element-wise.

Weak Irrepresentable Condition.

$$|C_{21}^n (C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)}^n)| < \mathbf{1},$$