

Mathematical techniques in data science

Model consistency of lasso

The Lasso

Settings

$$Y \in \mathbb{R}^{n \times 1}, \quad X \in \mathbb{R}^{n \times (p+1)}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & | & | & \dots & | \\ \dots & x^{(1)} & x^{(2)} & \dots & x^{(p)} \\ 1 & | & | & \dots & | \end{bmatrix}$$

Linear model: settings

- Linear model

$$Y = \beta^{(0)} + \beta^{(1)}X^{(1)} + \beta^{(2)}X^{(2)} + \dots + \beta^{(p)}X^{(p)} + \epsilon$$

- Equivalent to

$$\mathbf{Y} = \mathbf{X}\beta, \quad \beta = \begin{bmatrix} \beta^{(0)} \\ \beta^{(1)} \\ \dots \\ \beta^{(p)} \end{bmatrix}$$

- Least squares regression

$$\hat{\beta}^{LS} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2$$

Lasso

- The Lasso (Least Absolute Shrinkage and Selection Operator)

$$\hat{\beta}^{lasso} = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta^{(j)}|$$

Lasso: alternative form

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

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

Lasso: alternative form

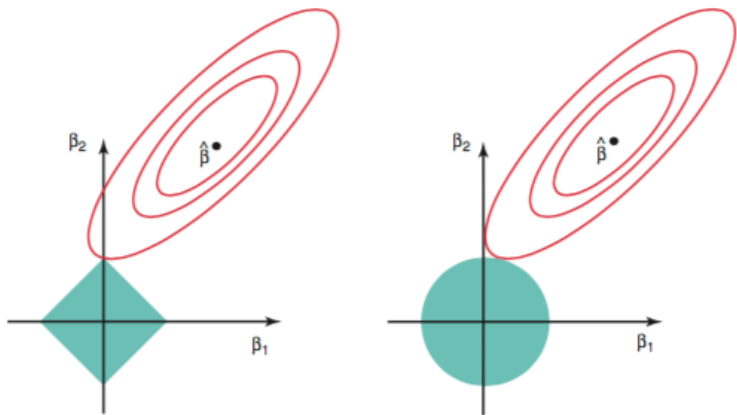
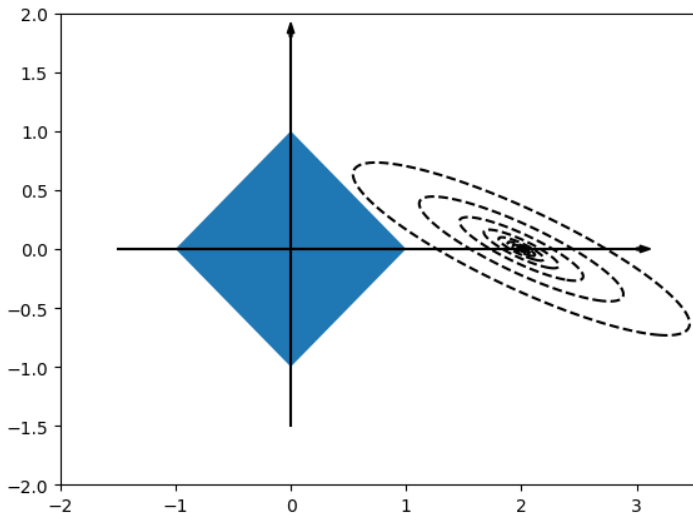
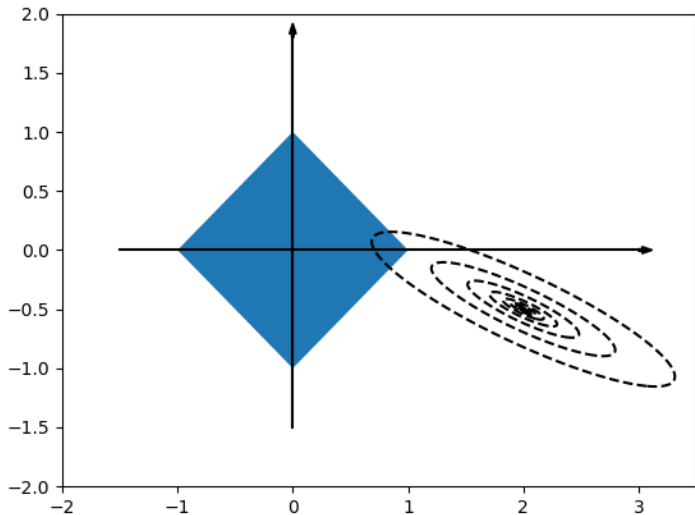


FIGURE 6.7. Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.

When the lasso fails



When the lasso fails



Lasso: model consistency

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 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- use the same notations as in the previous lectures

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

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

$$\text{sign}(\hat{\beta}^{(1)}) = \text{sign}(\beta^{*(1)}) \quad \text{and} \quad \hat{\beta}^{(2)} = 0$$

Sub-gradient

Issue: the penalty of lasso is non-differentiable

Definition

We say that a vector $s \in \mathbb{R}^p$ is a subgradient for the ℓ_1 -norm evaluated at $\beta \in \mathbb{R}^p$, written as $s \in \partial\|\beta\|$ if for $i = 1, \dots, p$ we have

$$s^{(i)} = \text{sign}(\beta^{(i)}) \quad \text{if } \beta^{(i)} \neq 0 \quad \text{and } s_i \in [-1, 1] \quad \text{otherwise.}$$

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

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

- (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*

$$\tilde{z}^{(1)} = \text{sign}(\beta^{*(1)})$$

is satisfied.

PDW

- This procedure is not a practical method for solving the ℓ_1 -regularized optimization problem, since solving the restricted problem in Step 1 requires knowledge about the sparsity of β^*
- Rather, the utility of this constructive procedure is as a proof technique: it succeeds if and only if the lasso has a optimal solution with the correct signed support.

A more detailed computation

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

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

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

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

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

This is acceptable as long as $\tilde{z}^{(1)} \in \partial|\tilde{\beta}^{(1)}|$. That is,

$$\text{sign}(\tilde{\beta}^{(1)}) = \text{sign}(\beta^{*(1)})$$

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:

- $\text{sign}(\beta^{*(1)}) = \text{sign}(\beta^{*(1)} + \Delta)$

where

$$\Delta = C_{11}^{-1}([\mathbf{X}^{(1)}]^T \epsilon - \lambda \text{sign}(\beta^{*(1)}))$$

- $|\tilde{z}^{(2)}| < 1$ 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

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

and

$$\tilde{z}^{(2)} = \frac{-1}{\lambda} C_{21} \Delta = C_{21} C_{11}^{-1} \text{sign}(\beta^{*(1)})$$

- 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

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

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

$$\frac{1}{n}C_{11} \rightarrow \text{Var}(X^{(1)}),$$

this means that when $n \rightarrow \infty$, we need

$$\frac{\lambda_n}{n} \rightarrow 0$$

Noisy setting

In principle, we want two conditions:

- $\text{sign}(\beta^{*(1)}) = \text{sign}(\beta^{*(1)} + \Delta)$

where

$$\Delta = C_{11}^{-1}([\mathbf{X}^{(1)}]^T \epsilon - \lambda \text{sign}(\beta^{*(1)}))$$

- $|\tilde{z}^{(2)}| < 1$ where

$$\tilde{z}^{(2)} = \frac{1}{\lambda} [\mathbf{X}^{(2)}]^T (\mathbf{X}^{(1)} \Delta + \epsilon)$$

- We want an upper bound on

$$[\mathbf{X}^{(1)}]^T \epsilon \quad \text{and} \quad [\mathbf{X}^{(2)}]^T \epsilon$$

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

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

with probability at least $1 - \delta$

General settings

Without loss of generality, assume $\beta^n = (\beta_1^n, \dots, \beta_q^n, \beta_{q+1}^n, \dots, \beta_p^n)^T$ where $\beta_j^n \neq 0$ for $j = 1, \dots, q$ and $\beta_j^n = 0$ for $j = q+1, \dots, p$. Let $\beta_{(1)}^n = (\beta_1^n, \dots, \beta_q^n)^T$ and $\beta_{(2)}^n = (\beta_{q+1}^n, \dots, \beta_p^n)$. 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 $C_{11}^n = \frac{1}{n} \mathbf{X}_n(1)^T \mathbf{X}_n(1)$, $C_{22}^n = \frac{1}{n} \mathbf{X}_n(2)^T \mathbf{X}_n(2)$, $C_{12}^n = \frac{1}{n} \mathbf{X}_n(1)^T \mathbf{X}_n(2)$ and $C_{21}^n = \frac{1}{n} \mathbf{X}_n(2)^T \mathbf{X}_n(1)$. 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

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

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

Weak Irrepresentable Condition.

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