Mathematical techniques in data science

Shrinkage methods

Reminders

- Homework 4 on the course's webpage. Due in 2 weeks.
- Check in with groups about projects this week
- I'm giving a talk at the Math Department's colloquium this Friday (3:30pm, 104 Gore Hall).
  Topic: Feature selection for non-linear models: (phylogenetic) trees and (deep neural) networks

Settings

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

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

Linear model: settings

- Linear model

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

- Equivalent to

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

- Least squares regression

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

Trade-off: complexity vs. interpretability

Linear model

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

- Higher values of p lead to more complex model $\rightarrow$ increases prediction power/accuracy
- Higher values of p make it more difficult to interpret the model: It is often the case that some or many of the variables regression model are in fact not associated with the response

Moderns settings

Linear model

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

- it is often the case that $n \ll p$
- requires supplementary assumptions (e.g. sparsity)
- can still build good models with very few observations.

ℓ_0 regularization

- ℓ_0 regularization

$$\hat{\beta}^0 = \min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{i=1}^{p} \mathbf{1}_{\beta^{(i)} \neq 0}$$

  where $\lambda > 0$ is a parameter
- pay a fixed price λ for including a given variable into the model
- variables that do not significantly contribute to reducing the error are excluded from the model (i.e., $\beta_i = 0$)
- problem: difficult to solve (combinatorial optimization). Cannot be solved efficiently for a large number of variables.

ℓ_2 (Tikhonov) regularization

- Ridge regression/ Tikhonov regularization

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

  where $\lambda > 0$ is a parameter
- shrinks the coefficients by imposing a penalty on their size
- penalty is a smooth function.
- easy to solve (solution can be written in closed form)
- can be used to regularize a rank deficient problem ($n < p$)

ℓ_2 (Tikhonov) regularization

$$\frac{\partial \left(\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|^2 \right)}{\partial \beta} = 2\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta$$

- The critical point satisfies

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

- Note: $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is positive definite, and thus invertible
- Thus

$$\hat{\beta}^{RIDGE} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

ℓ_2 (Tikhonov) regularization

$$\hat{\beta}^{RIDGE} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

- When $\lambda > 0$, the estimator is defined even when $n < p$
- When $\lambda = 0$ and $n > p$, we recover the usual least squares solution

The Lasso

Lasso

- The Lasso (Least Absolute Shrinkage and Selection Operator)

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

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large
- the lasso performs variable selection $\rightarrow$ models are easier to interpret

Lasso: alternative form

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

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

$$\text{subject to } \sum_{j=1}^{p} |\beta^{(j)}| \leq s$$

Lasso: alternative form

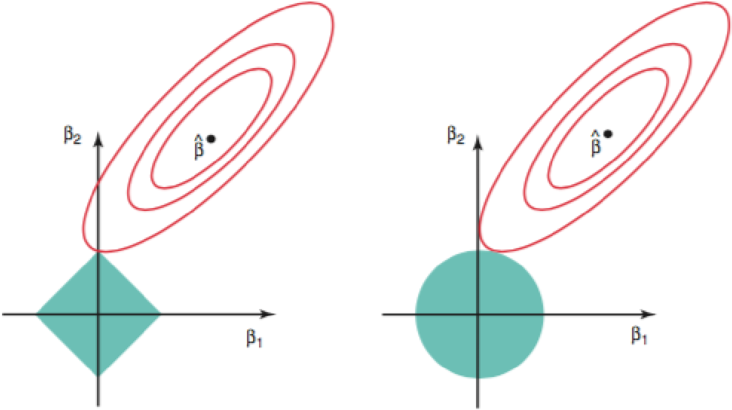


FIGURE 6.7. *Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$*, while the red ellipses are the contours of the RSS.*

Lasso

- The Lasso:

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

- More "global" approach to selecting variables compared to previously discussed greedy approaches
- Can be seen as a convex relaxation of the $\hat{\beta}^0$ problem
- No closed form solution, but can solved efficiently using convex optimization methods.
- Performs well in practice
- Very popular. Active area of research

Other shrinkage methods

- ℓ_q regularization ($q \geq 0$):

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

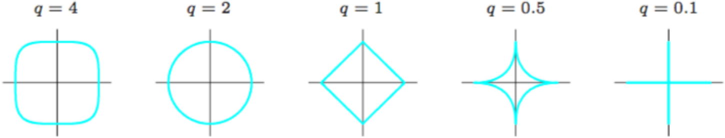


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Other shrinkage methods

- Elastic net

$$\lambda \sum_{j=1}^{p} \alpha[\beta^{(j)}]^2 + (1-\alpha)|\beta^{(j)}|$$

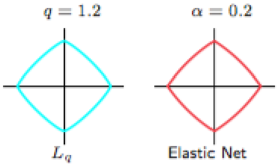


FIGURE 3.13. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for* $q = 1.2$ *(left plot), and the elastic-net penalty* $\sum_j(\alpha\beta_j^2+(1-\alpha)|\beta_j|)$ *for* $\alpha = 0.2$ *(right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the* $q = 1.2$ *penalty does not.*

Lasso: alternative form

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

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

$$\text{subject to } \sum_{j=1}^{p} |\beta^{(j)}| \leq s$$

Lasso: alternative form

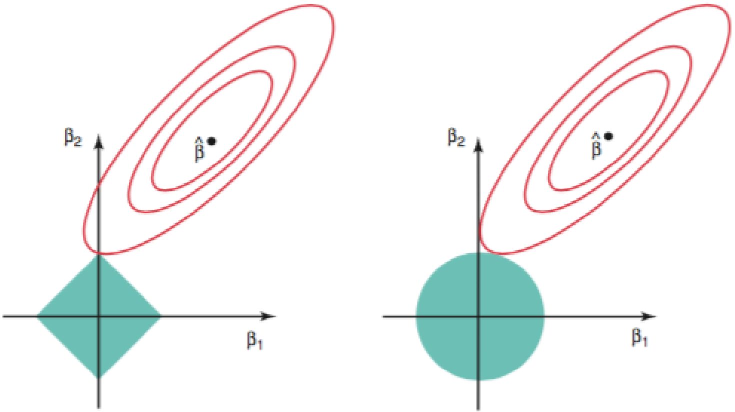
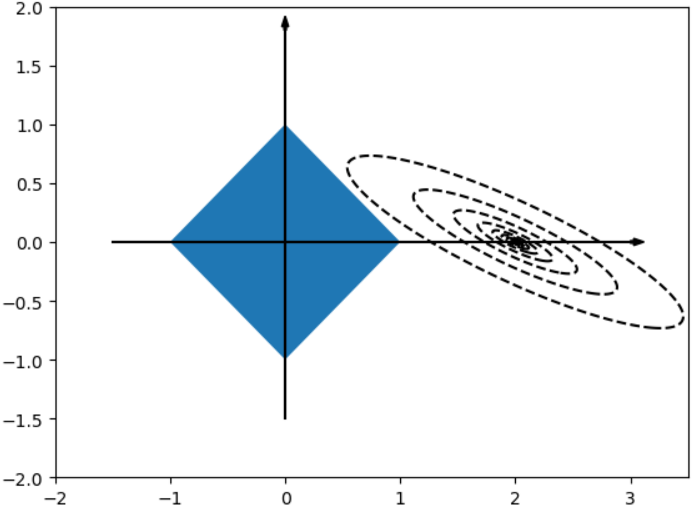
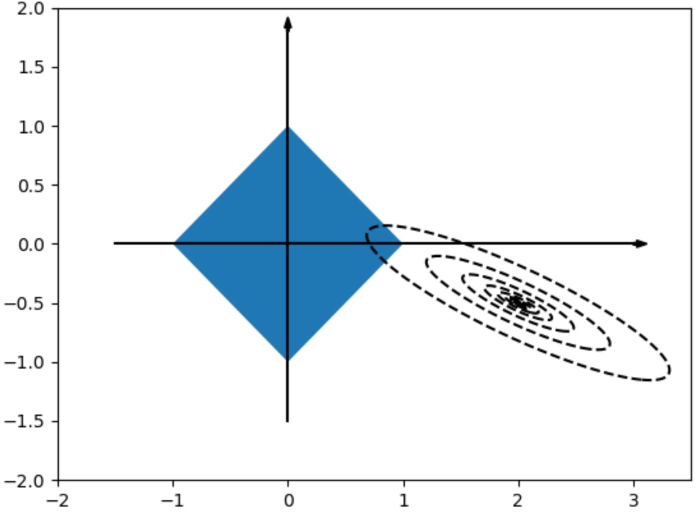


FIGURE 6.7. *Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

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

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

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

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

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 $i = 1, \ldots, p$ we have

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

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*

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

(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.

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

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

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

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

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

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

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

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

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

Weak Irrepresentable Condition.

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