Penalized Least Squares Methods

Yanrong Yang

RSFAS/CBE, Australian National University

2nd August 2022

Contents of this week

- High-dimensional Linear Regression Model
- Penalized Least Squares Methods
  - Hard thresholding penalization
  - Soft thresholding penalization
  - Ridge Regression
  - The Lasso, the adaptive lasso
  - Bridge Estimation
  - The elastic net
  - SCAD
- Details of Elastic Net Approach

High-dimensional Linear Regression Model (HDLRM)

Consider the linear regression model

$$y_i = x_{1i}\theta_1 + x_{2i}\theta_2 + \ldots + x_{pi}\theta_p + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where

- $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ includes all n observations for the response variable.
- $\mathbf{X} = (x_{ji})_{p\times n}$ is a $p \times n$ matrix grouping all observations for covariates.
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^\top$ is an $n \times 1$ error component.

The vector form of model (1) is

$$\mathbf{y} = \mathbf{X}^\top\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{2}$$

For simplicity, we assume orthonormal condition on $\mathbf{X}$, i.e.

$$\mathbf{X}\mathbf{X}^\top = \mathbf{I}_p. \tag{3}$$

Aims of Regression Modelling

Purpose

- Prediction Accuracy on future data - it is difficult to defend a model that predicts poorly. The prediction for the response y at the point $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})^\top$ is

$$\widehat{y}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\theta}}. \tag{4}$$

- Easy Interpretation - scientists prefer a simpler model because it puts more light on the relationship between the response and covariates.

$$\widehat{\boldsymbol{\theta}} = \left(\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_p\right). \tag{5}$$

For instance, some estimated coefficient $\widehat{\theta}_2 = 0$ indicates no linear relationship between the response y_i and the covariant x_{2i}.

Evaluation of Regression Modelling

Methods

- Assess the prediction accuracy - Mean Squared Error (MSE)

$$\mathbb{E}\left(\widehat{y}_i - y_i\right)^2 = \mathbb{E}\left(\mathbf{x}_i^\top \widehat{\boldsymbol{\theta}} - \mathbf{x}_i^\top \boldsymbol{\theta} - \varepsilon_i\right)^2$$

$$= \underbrace{\mathbf{x}_i^\top \cdot \mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \cdot \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^\top\right] \cdot \mathbf{x}_i}_{\textit{reducible error}} + \underbrace{var(\varepsilon_i)}_{\textit{irreducible error}} .$$

The Reducible error is decomposed into

$$\textit{Reducible error} = \mathbf{x}_i^\top \cdot Var\left(\widehat{\boldsymbol{\theta}}\right) \cdot \mathbf{x}_i + \mathbf{x}_i^\top \cdot Bias\left(\widehat{\boldsymbol{\theta}}\right) Bias\left(\widehat{\boldsymbol{\theta}}\right)^\top \cdot \mathbf{x}_i$$

- Variable Selection: subset selection method, penalized method

Illustration of variance and bias for OLS $\widehat{\boldsymbol{\theta}}$

- variance matrix

$$Var\left(\widehat{\boldsymbol{\theta}}\right) = \begin{pmatrix} var(\widehat{\theta}_1) & cov(\widehat{\theta}_1, \widehat{\theta}_2) & \dots & cov(\widehat{\theta}_1, \widehat{\theta}_p) \\ cov(\widehat{\theta}_2, \widehat{\theta}_1) & var(\widehat{\theta}_2) & \dots & cov(\widehat{\theta}_2, \widehat{\theta}_p) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\widehat{\theta}_p, \widehat{\theta}_1) & cov(\widehat{\theta}_p, \widehat{\theta}_2) & \dots & var(\widehat{\theta}_p) \end{pmatrix},$$

- bias vector

$$Bias(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} Bias(\widehat{\theta}_1) \\ Bias(\widehat{\theta}_2) \\ \vdots \\ Bias(\widehat{\theta}_p) \end{pmatrix} = \mathbf{0}_p.$$

Reducible Error of OLS prediction $\widehat{y}_i$ under Large p

Reducible Error consists of two terms:

- Look at the first term: denote $Var(\widehat{\boldsymbol{\theta}}) = (\Sigma_{jk})_{p\times p}$.

$$\mathbf{x}_i^\top \cdot Var(\widehat{\boldsymbol{\theta}}) \cdot \mathbf{x}_i = \sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\Sigma_{jk}. \tag{7}$$

  - As p is large, this term is large with large chance.
  - One option to reduce this term is to decrease the values $\{\Sigma_{jk} : j, k = 1, 2, \ldots, p\}$.
  - To decrease the value of $\Sigma_{jk} = cov(\widehat{\theta}_j, \widehat{\theta}_k)$, shrinking the OLS estimators $\widehat{\theta}_j$ and $\widehat{\theta}_k$ is one feasible method.
  - We will use the penalized least squares method to attain this aim.

- The second term $\mathbf{x}_i^\top \cdot Bias\left(\widehat{\boldsymbol{\theta}}\right) Bias\left(\widehat{\boldsymbol{\theta}}\right)^\top \cdot \mathbf{x}_i = 0$ as $Bias(\boldsymbol{\theta}) = \mathbf{0}_p$.

Variable Selection for HDLRM

Two methods

- Subset Selection: the major disadvantage of this method is instability.
- Penalized Method: Taking the lasso as an example, various penalization functions are able to do variable selection.

Because the penalized least squares method could shrink/regularize OLS estimation and conduct variable selection simultaneously, we study this method this week.

Challenge of HDLRM

OLS meets Curse of Dimensionality in the high-dimensional case ($p \to \infty$ when $n \to \infty$).

- Curse in prediction accuracy:
  - caused by increase of variance $\mathbf{x}_i^\top \cdot Var(\widehat{\theta}) \cdot \mathbf{x}_i$ as p increases.
  - so the essential solution is to reduce the variance $\mathbf{x}_i^\top \cdot Var(\widehat{\theta}) \cdot \mathbf{x}_i$.
  - One feasible approach is to shrink the OLS estimator $\widehat{\theta}$ and attain smaller variance.
- High demand in variable selection: large p indicates more possibility of involving irrelevant covariates.
  - Select the irrelevant covariates by setting the corresponding coefficient estimation as zero.
  - Possible methods include subset selection and penalization methods.

As an example, a typical microarray data set has many thousands of predictors (genes) and often fewer than 100 sample observations.

Penalized Least Squares Estimation (PLSE)

- Recall the OLS estimator $\widehat{\boldsymbol{\theta}} = \mathbf{X}\mathbf{y}$ is a minimizer of the objective function

$$\left|\left|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\theta}\right|\right|^2 \tag{8}$$

- A penalized least squares estimator $\widetilde{\boldsymbol{\theta}}$ is defined as the minimizer of the objective function

$$\frac{1}{2}\left|\left|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\theta}\right|\right|^2 + \lambda \cdot \sum_{j=1}^{p} g_j(|\theta_j|). \tag{9}$$

- Here the functions $\{g_j(\cdot) : j = 1, 2, \ldots, p\}$ are penalty functions, which could be the same or different across the parameters $\theta_j, j = 1, 2, \ldots, p$.

Interpretation of PLSE

The equivalent optimization problem of PLSE ($\widetilde{\boldsymbol{\theta}}$) is

$$\min_{\theta_1,\ldots,\theta_p} \frac{1}{2}\left|\left|\mathbf{y}-\mathbf{X}^\top\boldsymbol{\theta}\right|\right|^2,$$
$$s.t. \quad \sum_{j=1}^{p} g_j(|\theta_j|) \leq \lambda.$$

- In terms of this optimization problem, we see that the OLS $\widehat{\boldsymbol{\theta}}$ are shrunk via the condition $\sum_{j=1}^{p} g_j(|\theta_j|) \leq \lambda$, and then attain the final estimator $\widetilde{\boldsymbol{\theta}}$.
- If $g_j(\theta) = |\theta|$ which is the lasso penality function, this restriction condition is $\sum_{j=1}^{p} |\theta_j| \leq \lambda$.

Another Way to Look at OLS and PLSE

Let $\mathbf{z} = \mathbf{X}\mathbf{y}$ and $\widehat{\mathbf{y}} = \mathbf{X}^\top\mathbf{X}\mathbf{y}$. The objective function for the penalized least squares estimation could be decomposed

$$\frac{1}{2}\left|\left|\mathbf{y} - \mathbf{X}^\top\boldsymbol{\theta}\right|\right|^2 + \lambda \cdot \sum_{j=1}^{p} g_j(|\theta_j|) \tag{10}$$

$$= \quad \frac{1}{2}||\mathbf{y} - \widehat{\mathbf{y}}||^2 + \frac{1}{2}\sum_{j=1}^{p}(z_j - \theta_j)^2 + \lambda\sum_{j=1}^{p} g_j(|\theta_j|). \tag{11}$$

So minimizing (10) is equivalent to minimizing the following objective function

$$\frac{1}{2}\sum_{j=1}^{p}(z_j - \theta_j)^2 + \lambda \cdot \sum_{j=1}^{p} g_j(|\theta_j|) \tag{12}$$

- The first term of (12) implies PLSE is expected to be close to the OLS z_j.
- The second term of (12) restricts $\sum_{j=1}^{p} g_j(\theta_j)$ not very large.

Summary of Common Penalty Functions

We list some common used penalty functions $\lambda \cdot g(\cdot)$.

hard thresholding	$\lambda^2 - (\lvert\theta\rvert - \lambda)^2 \cdot I(\lvert\theta\rvert < \lambda)$
soft thresholding / the Lasso	$\lambda \cdot \lvert\theta\rvert$
Ridge Regression	$\lambda \cdot \lvert\theta\rvert^2$
Bridge Regression	$\lambda \cdot \lvert\theta\rvert^q$, $q > 0$
Adaptive Lasso	$\lambda \sum_{j=1}^{p} w_j \lvert\theta_j\rvert$
Naive Elastic Net	$\lambda_1 \lvert\theta\rvert + \lambda_2 \lvert\theta\rvert^2$
SCAD	$\lambda \cdot \lvert\theta\rvert$ if $\lvert\theta\rvert \leq \lambda$
	$\frac{2\gamma\lambda\lvert\theta\rvert - \theta^2 - \lambda^2}{2(\gamma-1)}$ if $\lambda < \lvert\theta\rvert < \gamma\lambda$
	$\frac{\lambda^2(\gamma+1)}{2}$, if $\lvert\theta\rvert \geq \gamma\lambda$.

What is a good penalty function?

A good penalty function should result in an estimator with three properties

- Unbiasedness: the resulting estimator is nearly unbiased when the true unknown parameter is large to avoid unnecessary modelling bias.
- Sparsity: the resulting estimator is a thresholding rule, which automatically sets small estimated coefficients to zero to reduce model complexity.
- Continuity: the resulting estimator is continuous in data z to avoid instability in model prediction.

Sufficient Conditions for Good Properties

For simplicity of presentation, we assume that the penalty functions for all coefficients are the same, denoted by $g(|\cdot|)$. Furthermore, we denote $\lambda \cdot g(|\cdot|)$ by $g_\lambda(|\cdot|)$. So the objective function (12) (when $p = 1$) is reduced to

$$\frac{1}{2}(z-\theta)^2 + g_\lambda(|\theta|). \tag{13}$$

Next, we will discuss sufficient conditions (on penalty functions) for the penalized estimator satisfying the three good properties.

Sufficient Condition for Unbiasedness

The condition

$$g_{\lambda}^{'}(|\theta|) = 0 \ \textit{for large} \ |\theta|$$

is a sufficient condition for unbiasedness for a large true parameter.

- Let the first order derivative of (13) with respect to θ equal to zero, i.e.

$$sgn(\theta)\left[|\theta| + g_{\lambda}^{'}(|\theta|)\right] - z = 0. \tag{14}$$

- When $g_{\lambda}^{'}(|\theta|) = 0$ for large $|\theta|$, we have from (14) that $\widetilde{\theta} = z$ which is the unbiased OLS.

Sufficient Condition for Sparsity

A sufficient condition for the resulting estimator to be a thresholding rule is that

the minimum of the function $|\theta| + g_\lambda'(|\theta|)$ *is positive.*

- When $|z| < \min_{\theta \neq 0} \left[|\theta| + g_\lambda'(|\theta|) \right]$, the derivative of (13) is positive for all positive θ's and is negative for all negative θ's.
- Thus, the penalized least squares estimator is 0 in this situation.
- Namely, $\widetilde{\theta} = 0$ for $|z| < \min_{\theta \neq 0} \left[|\theta| + g_\lambda'(|\theta|) \right]$.
- This implies that when the OLS z is located around the origin point, the penalized least squares estimator $\widetilde{\theta}$ is set to be 0.

Sufficient Condition for Continuity

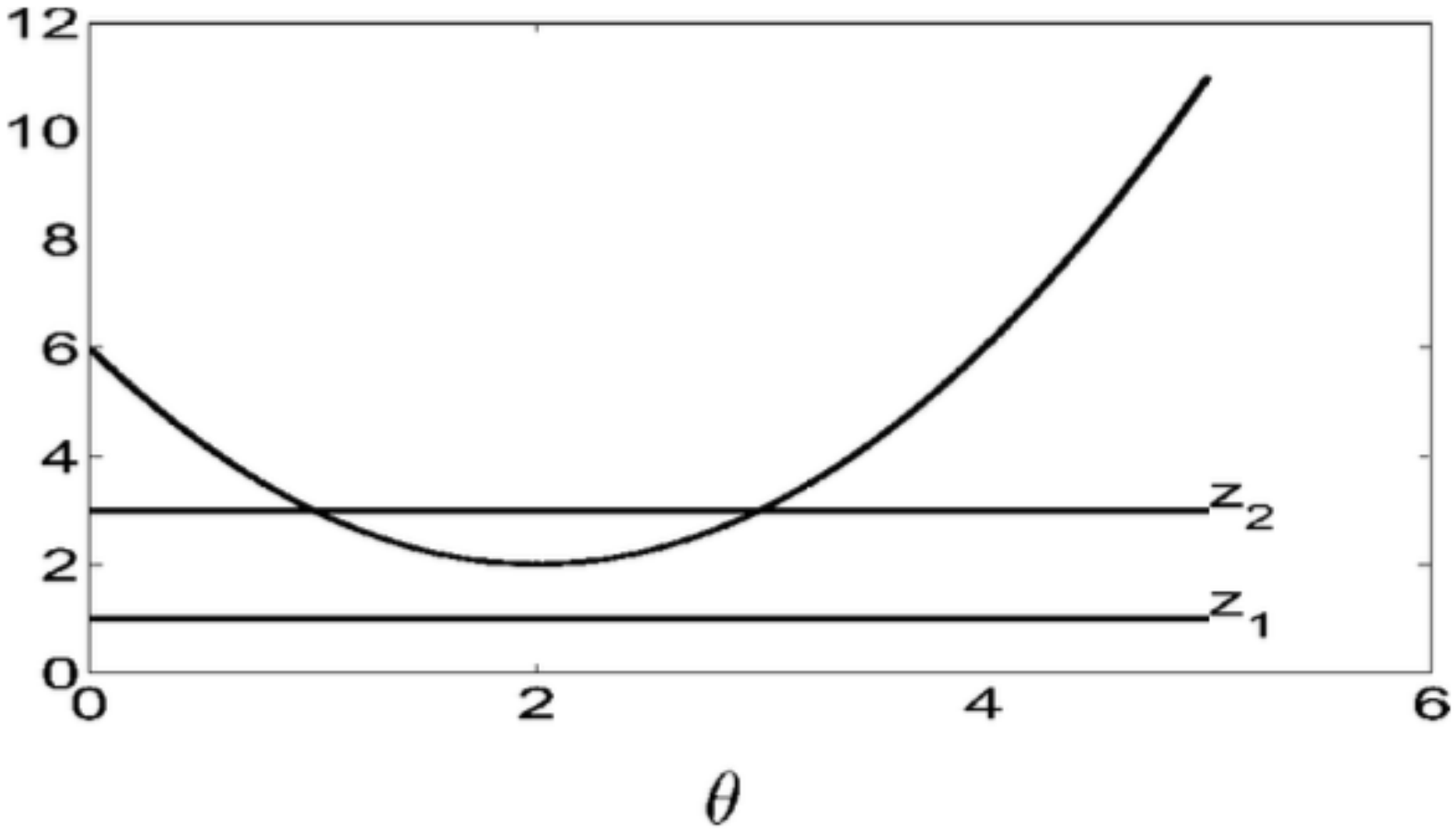
A sufficient and necessary condition for continuity is that

the minimum of the function $|\theta| + g_\lambda'(|\theta|)$ *is attained at* 0.

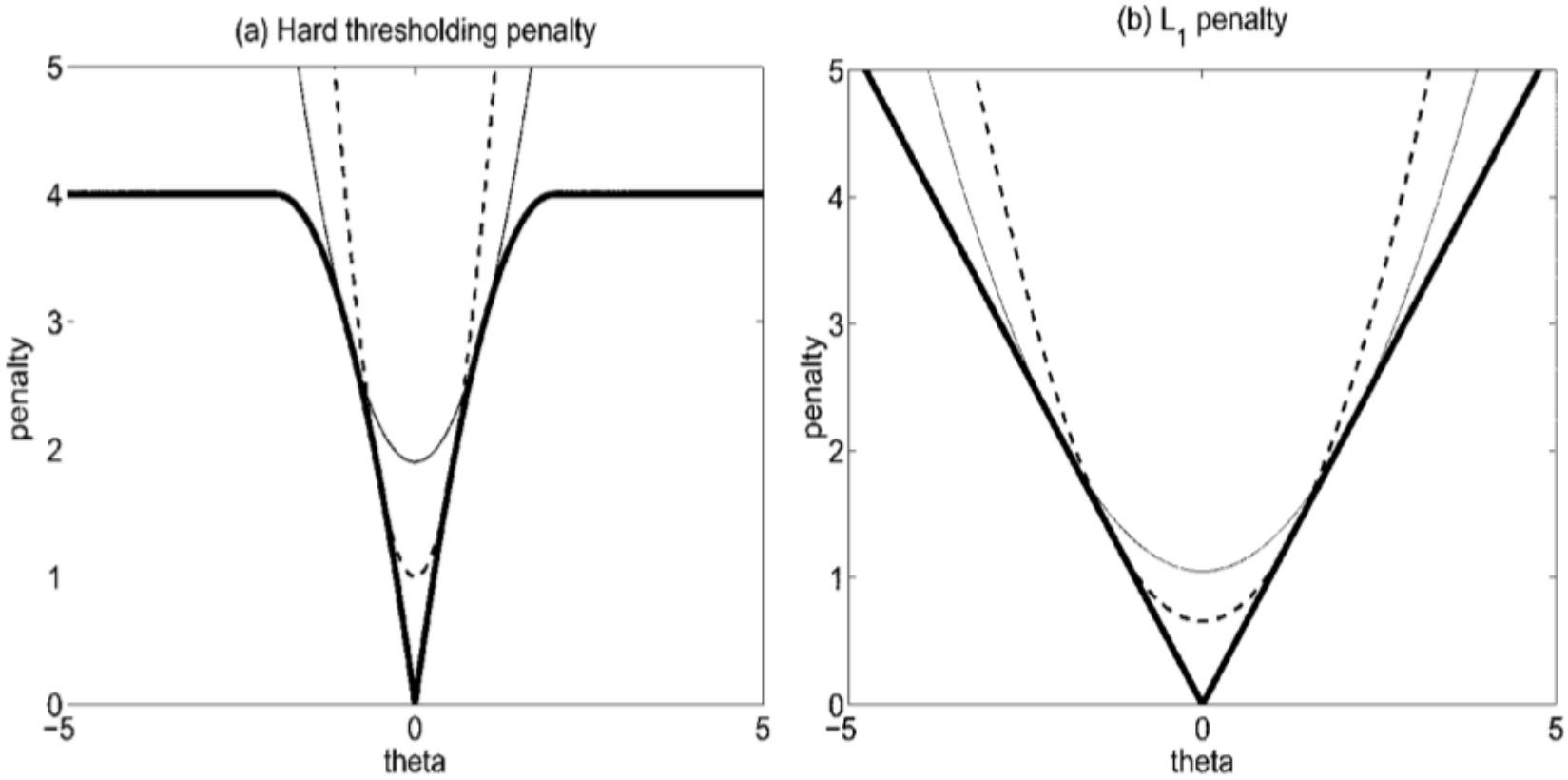
- When $|z| > \min_{\theta \neq 0} \left[|\theta| + g_\lambda'(|\theta|) \right]$, two crossings may exist as shown in the following figure.
- To avoid this phenomenon in the figure, we should let the minimum of $|\theta| + g_\lambda'(|\theta|)$ is attained at 0.

In summary, a penalty function satisfying sparsity and continuity must be singular at the origin.

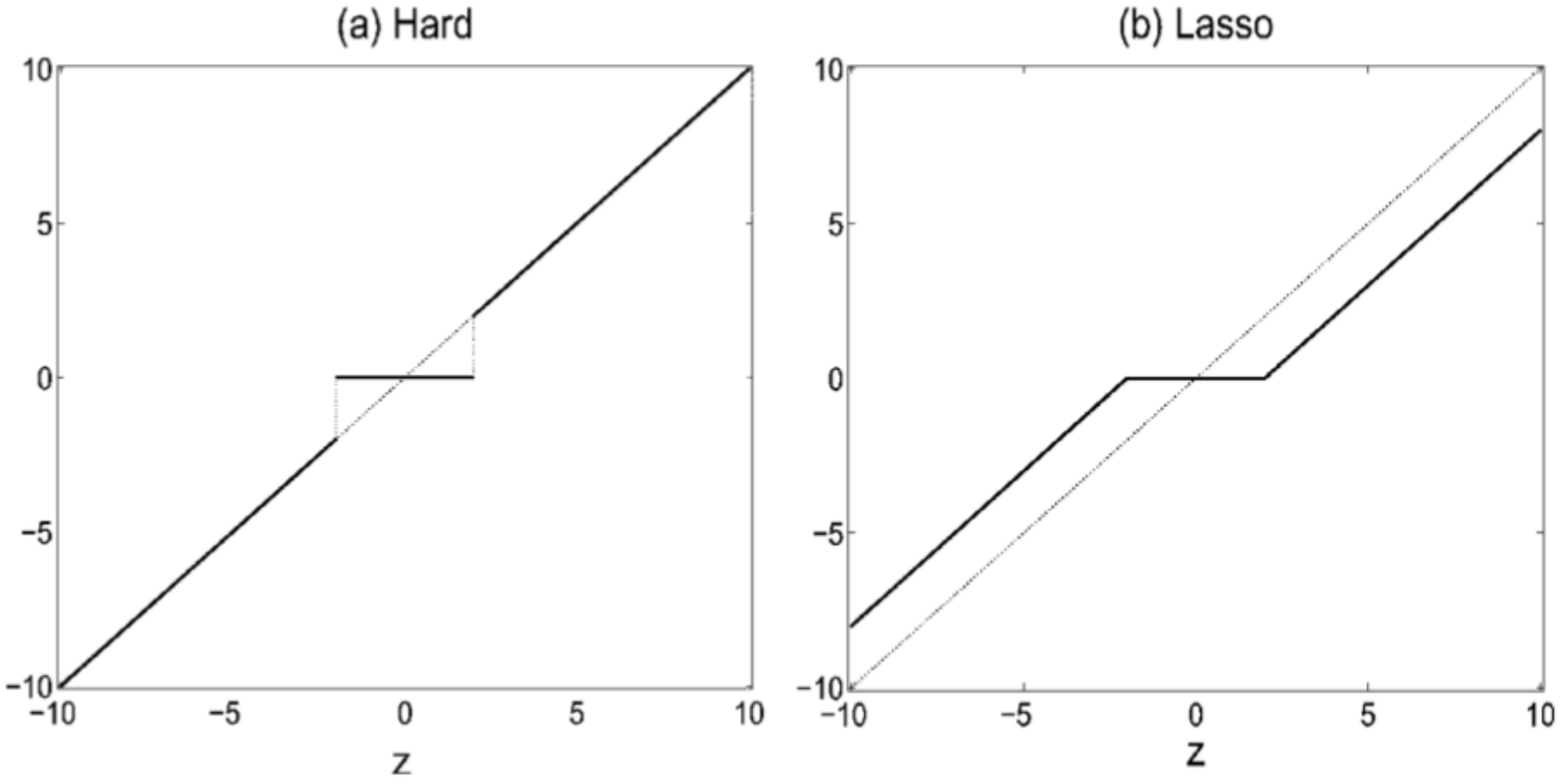
A Plot of $\theta + g_\lambda'(\theta)$ against θ



Penalty Functions and their Quadratic Approximations



PLSE's



Disadvantage of the Lasso

Next, we will look at details of a penalized least squares estimator - Elastic Net Estimation. Before details, we go through the disadvantages of the Lasso.

(a) In the $p > n$ case, the lasso selects at most n variables in the variable selection procedure. This is a limit feature for the ultra-high dimensional case.

(b) If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.

(c) For the usual $p < n$ case, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression.

Naive Elastic Net

For any fixed non-negative tuning parameters λ_1 and λ_2, we define the naive elastic net objective function

$$L(\lambda_1, \lambda_2, \boldsymbol{\theta}) = \left|\left|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\theta}\right|\right|^2 + \lambda_2 ||\boldsymbol{\theta}||^2 + \lambda_1 ||\boldsymbol{\theta}||_1, \tag{15}$$

where

$$||\boldsymbol{\theta}||^2 = \sum_{j=1}^{p} \theta_j^2, \quad ||\boldsymbol{\theta}||_1 = \sum_{j=1}^{p} |\theta_j|.$$

The naive elastic net estimator $\widetilde{\boldsymbol{\theta}}$ is the minimizer of $L(\lambda_1, \lambda_2, \boldsymbol{\theta})$, i.e.

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} L(\lambda_1, \lambda_2, \boldsymbol{\theta}). \tag{16}$$

Interpretation for Naive Elastic Net

- Let $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$. Then (16) is equivalent to the optimization problem

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left|\left|\mathbf{y} - \mathbf{X}^\top \boldsymbol{\theta}\right|\right|^2$$
$$s.t.\ \ (1-\alpha)||\boldsymbol{\theta}||_1 + \alpha||\boldsymbol{\theta}||^2 \leq t,\ \ \textit{for some } t.$$

- The elastic net penalty function $(1-\alpha)||\boldsymbol{\theta}||_1 + \alpha||\boldsymbol{\theta}||^2$, is a convex combination of the lasso and ridge penalty.
- When $\alpha = 1$, the naive elastic net becomes ridge regression; when $\alpha = 0$, it is the lasso estimation.
- For all $\alpha \in [0, 1)$, the elastic net penalty function is singular at the origin.

Comparison: Restricted Area incurred by Penalty Functions

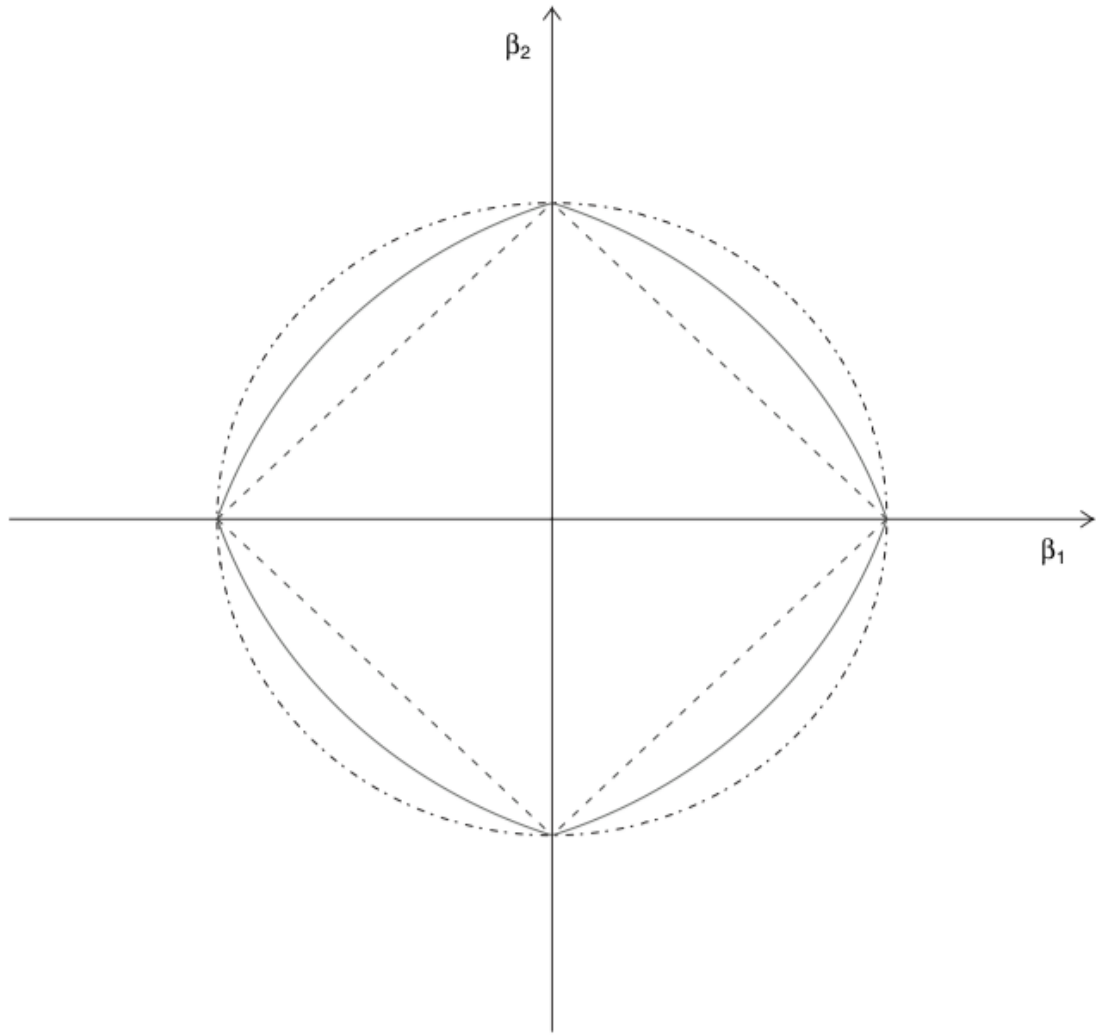


Fig. 1. Two-dimensional contour plots (level 1) (·-·-·-·-, shape of the ridge penalty; -------, contour of the lasso penalty; ———, contour of the elastic net penalty with $\alpha = 0.5$): we see that singularities at the vertices and the edges are strictly convex; the strength of convexity varies with α

Relationship with the Lasso

Theorem (Augmented Lasso)

Given data set $(\mathbf{y}, \mathbf{X})$ *and tuning parameters* (λ_1, λ_2)*, define an artificial data set* $(\mathbf{y}^*, \mathbf{X}^*)$ *by*

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X}^\top \\ \sqrt{\lambda_2}\mathbf{I}_p \end{pmatrix}, \quad \mathbf{y}^*_{(n+p)} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{p\times 1} \end{pmatrix}. \quad (17)$$

Let $\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$ *and* $\boldsymbol{\theta}^* = \sqrt{1+\lambda_2}\boldsymbol{\theta}$*. Then the naive elastic net objective function can be written as*

$$L(\gamma, \boldsymbol{\theta}^*) = ||\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}^*||^2 + \gamma \cdot ||\boldsymbol{\theta}^*||_1. \quad (18)$$

Then

$$\widetilde{\boldsymbol{\theta}} = \frac{1}{\sqrt{1+\lambda_2}}\widetilde{\boldsymbol{\theta}}^*, \quad (19)$$

where $\widetilde{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta}^*} L(\gamma, \boldsymbol{\theta}^*)$.

Insights from this Theorem

This theorem informs us that

- The naive elastic net problem is an equivalent lasso problem on augmented data.
- Note that the sample size in the augmented problem is $n + p$ and $\mathbf{X}^*$ has rank p, which means that the naive elastic net can potentially select all p predictors in all situations.
- Similar to the lasso, the naive elastic net estimator can perform an automatic variable selection.

Comparison: Ridge, Lasso and Naive Elastic Net

In the case of an orthonormal design $\mathbf{X}\mathbf{X}^\top = \mathbf{I}$, the three penalized least squares estimators (Ridge, Lasso and Naive Elastic Net) have the following expressions:

- The ridge estimator with parameter λ_2 is

$$\widetilde{\boldsymbol{\theta}}(ridge) = \frac{\widehat{\boldsymbol{\theta}}(OLS)}{1+\lambda_2}. \tag{20}$$

- The lasso estimator with the parameter λ_1 is

$$\widetilde{\theta}_i(lasso) = \left(|\widehat{\theta}_i(OLS)| - \frac{\lambda_1}{2}\right)_+ sgn(\widehat{\theta}_i(OLS)). \tag{21}$$

- The naive elastic net estimator with parameters λ_1 and λ_2 is

$$\widetilde{\theta}_i(naive\ elastic\ net) = \frac{\left(|\widehat{\theta}_i(OLS)| - \frac{\lambda_1}{2}\right)_+}{1+\lambda_2} sgn(\widehat{\theta}_i(OLS)). \tag{22}$$

OLS, Lasso, Ridge and Naive Elastic Net

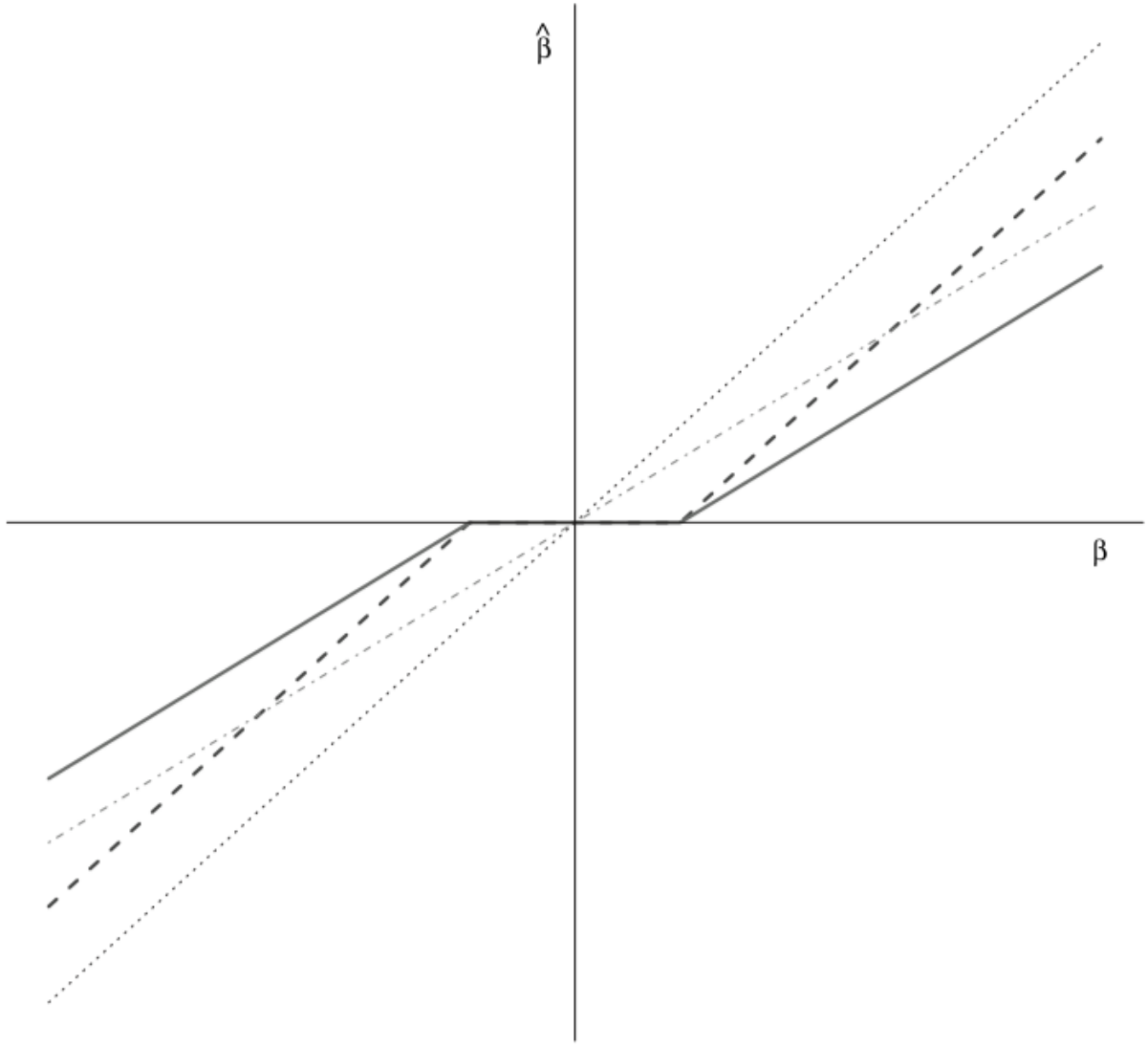


Fig. 2. Exact solutions for the lasso (-------), ridge regression (· · · · · ·) and the naïve elastic net (———) in an orthogonal design (· · · · · · ·, OLS): the shrinkage parameters are $\lambda_1 = 2$ and $\lambda_2 = 1$

The Grouping Effect

- Qualitatively speaking, a regression method exhibits the grouping effect if the regression coefficients of a group of highly correlated variables tend to be equal (up to a change of sign if negatively correlated).
- In the extreme situation where some variables are exactly identical, the regression method should assign identical coefficients to the identical variables.

Penalty Functions with Grouping Effect

Strict convexity of penalty functions guarantees the grouping effect.

Theorem (Grouping Effect)

Assume that $\mathbf{x}_i = \mathbf{x}_j$, $i, j \in \{1, 2, \ldots, p\}$.

- *If the penalty function is strictly convex, then* $\widetilde{\theta}_i = \widetilde{\theta}_j$.
- *If the penalty function is* $||\boldsymbol{\theta}||_1$, *then* $\widetilde{\theta}_i \widetilde{\theta}_j \geq 0$ *and* $\bar{\boldsymbol{\theta}}$ *is another minimizer of the penalized least squares objective function, where*

$$\bar{\theta}_k = \begin{cases} \widetilde{\theta}_k, & \text{if } k \neq i, k \neq j; \\ (\widetilde{\theta}_i + \widetilde{\theta}_j) \cdot s, & \text{if } k = i; \\ (\widetilde{\theta}_i + \widetilde{\theta}_j) \cdot (1 - s), & \text{if } k = j, \end{cases} \tag{23}$$

Grouping Effect of Naive Elastic Net

The elastic net penalty with $\lambda_2 > 0$ is strictly convex and enjoy the grouping effect as follows.

Theorem (Elastic Net Grouping Effect)

Given the data $(\mathbf{y}, \mathbf{X})$ *and parameters* (λ_1, λ_2). *Let* $\widetilde{\boldsymbol{\theta}}(\lambda_1, \lambda_2)$ *be the naive elastic net estimation. Suppose that* $\widetilde{\theta}_i(\lambda_1, \lambda_2)\widetilde{\theta}_j(\lambda_1, \lambda_2) > 0$. *Define*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{||\mathbf{y}||_1} \left| \widetilde{\theta}_i(\lambda_1, \lambda_2) - \widetilde{\theta}_j(\lambda_1, \lambda_2) \right|. \tag{24}$$

Then

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)}, \tag{25}$$

where $\rho = \mathbf{x}_i^\top \mathbf{x}_j$ *is the sample correlation.*

Insights from this Theorem

- The quantity $D_{\lambda_1,\lambda_2}(i,j)$ describes the difference between the coefficient paths of predictors $\mathbf{x}_i$ and $\mathbf{x}_j$.
- If predictors $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated, i.e. $\rho = 1$ (if $\rho = -1$, then consider $-\mathbf{x}_j$), this theorem says that the difference between the coefficient paths of predictor $\mathbf{x}_i$ and predictor $\mathbf{x}_j$ is 0.
- The upper bound in this equality provides a qualitative description for the grouping effect of the naive elastic net.

Benefits and Deficiency of the Naive Elastic Net

- Benefits: as an automatic variable selection method, the naive elastic net overcomes the limitations of the lasso in scenarios (a) and (b).
- Deficiency: empirical evidence shows that the naive elastic net does not perform satisfactorily unless it is very close to either ridge regression or the lasso. This is why we call it naive.
- The major reason is a double amount of shrinkage: for each fixed λ_2, we first find the ridge regression coefficients, and then we do the lasso-type shrinkage along the lasso coefficient solution paths.

The Elastic Net Estimation

The elastic net estimation is rescaled naive elastic net estimation.

$$\widetilde{\boldsymbol{\theta}}(\textit{elastic net}) = (1+\lambda_2)\cdot\widetilde{\boldsymbol{\theta}}(\textit{naive elastic net}). \tag{26}$$

- Recall that $\widetilde{\boldsymbol{\theta}}(\textit{naive elastic net}) = \frac{1}{\sqrt{1+\lambda_2}}\widetilde{\boldsymbol{\theta}}^*$, where

$$\widetilde{\boldsymbol{\theta}}^* = \arg\min_{\boldsymbol{\theta}^*} ||\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\theta}^*||^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}||\boldsymbol{\theta}^*||_1. \tag{27}$$

- A scaling transformation $1+\lambda_2$ in elastic net estimation preserves the variable selection property of the naive elastic net and meanwhile it is the simplest way to undo shrinkage.

Another Interpretation of Elastic Net

The following theorem interprets the elastic net as a stabilized version of the lasso.

Theorem

Given data $(\mathbf{y}, \mathbf{X})$ *and* (λ_1, λ_2)*, then the elastic net estimates* $\widetilde{\boldsymbol{\theta}}$ *are given by*

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^\top \left(\frac{\mathbf{X}\mathbf{X}^\top + \lambda_2 \mathbf{I}_p}{1+\lambda_2} \right) \boldsymbol{\theta} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \lambda_1 ||\boldsymbol{\theta}||_1. \qquad (28)$$

It is easy to see that

$$\widetilde{\boldsymbol{\theta}}(\mathit{lasso}) = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^\top \left(\mathbf{X}\mathbf{X}^\top \right) \boldsymbol{\theta} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \lambda_1 ||\boldsymbol{\theta}||_1 \qquad (29)$$

Insights from this Theorem

- The matrix

$$\frac{\mathbf{X}\mathbf{X}^\top + \lambda_2 \mathbf{I}_p}{1+\lambda_2} = (1-\gamma)\mathbf{X}\mathbf{X}^\top + \gamma \mathbf{I}_p,$$

  where $\gamma = \frac{\lambda_2}{1+\lambda_2}$.
- This matrix shrinks the sample covariance matrix $\mathbf{X}\mathbf{X}^\top$ towards the identity matrix $\mathbf{I}_p$.
- The elastic net estimation is equivalent to replacing the sample covariance matrix $\mathbf{X}\mathbf{X}^\top$ with its shrunken version in the lasso.

Connection with Univariate Soft Thresholding

The lasso is a special case of the elastic net with $\lambda_2 = 0$. The other interesting special case of the elastic net emerges when $\lambda_2 \to \infty$.

- In terms of (28), the elastic net estimator $\widetilde{\boldsymbol{\theta}} \to \widetilde{\boldsymbol{\theta}}(\infty)$ as $\lambda_2 \to \infty$, where

$$\widetilde{\boldsymbol{\theta}}(\infty) = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\theta}^\top \boldsymbol{\theta} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \lambda_1 ||\boldsymbol{\theta}||_1. \tag{30}$$

- $\widetilde{\boldsymbol{\theta}}(\infty)$ has a simple closed form

$$\widetilde{\boldsymbol{\theta}}(\infty) = \left(|\mathbf{y}^\top \mathbf{x}_i| - \frac{\lambda_1}{2}\right)_+ sgn(\mathbf{y}^\top \mathbf{x}_i), \quad i = 1, 2, \ldots, p. \tag{31}$$

- Note that $\widetilde{\boldsymbol{\theta}}(\infty)$ are the estimates by applying soft thresholding on univariate OLS regression coefficients $\{\mathbf{y}^\top \mathbf{x}_i, i = 1, 2, \ldots, p\}$.

Conclusion

Understand

- Curse of dimensionality in high-dimensional linear regression.
- Penalized least squares estimation and influences of penalty functions properties in final penalized estimation.
- The advantages of elastic net estimation over the lasso and ridge estimation.
- The relationship between elastic net estimation and the lasso, ridge estimation.