

Untitled

Boyu Chen

2023-06-10

The LASSO

Def. the data (\mathbf{x}_i, y_i) , $i = 1, 2, \dots, N$ where \mathbf{x}_{ij} are standardized, s.t. $\frac{1}{N} \sum_i X_{ij} = 0$, $\frac{1}{N} \sum_i X_{ij}^2 = 1$ Let $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)^T$, the LASSO estimate $(\hat{\alpha}, \hat{\beta})$:

$$(\hat{\alpha}, \hat{\beta}) := \arg \min \left\{ \sum_{i=1}^N (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right\} \text{ s.t. } \sum_j |\beta_j| \leq t \quad (1)$$

where $t \geq 0$ is tuning parameters.

$\hat{\alpha} = \bar{y}$ for all t , WLOG, set $\bar{y} = 0$ hence we can omit α .

The LASSO

The problem becomes:

$$\hat{\beta} := \arg \min \left\{ \sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2 \right\} \text{ s.t. } \sum_j |\beta_j| \leq t \quad (2)$$

$$\begin{aligned} \hat{\beta} &:= \arg \min_{\beta} \left((Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\| \right) \\ &= \arg \min_{\beta} \left(-Y^T X\beta + \frac{1}{2} \beta^T X^T X \beta + \gamma \|\beta\| \right) \end{aligned}$$

Orthogonal Design Case

X is orthogonal, i.e., $X^T X = I$, and then $\hat{\beta}^o = X^T Y$.

$$\begin{aligned} & \min_{\beta} \quad -Y^T X \beta + \beta^T X^T X \beta + \gamma \|\beta\| \\ \Rightarrow & \min_{\beta} \quad -\hat{\beta}^o \beta + \frac{1}{2} \beta^T \beta + \gamma \|\beta\| \\ \Rightarrow & \min_{\beta} \quad \sum_{i=1}^p -\hat{\beta}^o \beta + \frac{1}{2} \beta_i^2 + \gamma |\beta_i| \end{aligned}$$

For a certain i , the Lagrangian function is

$$\mathcal{L}_i = -\hat{\beta}^o \beta + \frac{1}{2} \beta_i^2 + \gamma |\beta_i| \quad (3)$$

If $\hat{\beta}_i^o > 0$, then we must have $\beta_i \geq 0$, since if $\beta_i < 0$, \mathcal{L}_i cannot be minimized. Likewise, if $\hat{\beta}_i^o < 0$, $\beta_i \leq 0$.

Derivation

Case 1: $\hat{\beta}_i^o > 0$

Since $\beta_i \geq 0$,

$$\mathcal{L}_i = -\beta_i^o \beta_i + \frac{1}{2} \beta_i^2 + \gamma \beta_i$$

Taking the first-order condition, we get

$$\frac{\partial \mathcal{L}_i}{\partial \beta_i} = -\hat{\beta}_i^o + \beta_i + \gamma = 0$$

This gives us

$$\begin{aligned}\hat{\beta}_i^{\text{lasso}} &= \begin{cases} \hat{\beta}_i^o - \gamma & \text{if } \hat{\beta}_i^o - \gamma \geq 0, \\ 0 & \text{otherwise.} \end{cases} \\ &= (\hat{\beta}_i^o - \gamma)^+ \\ &= \text{sgn}(\hat{\beta}_i^o)(|\hat{\beta}_i^o| - \gamma)^+\end{aligned}$$

Derivation

Case 2: $\hat{\beta}_i^o < 0$

Since $\beta_i \leq 0$,

$$\mathcal{L}_i = -\beta_i^o \beta_i + \frac{1}{2} \beta_i^2 - \gamma \beta_i$$

Taking the first-order condition, we get

$$\frac{\partial \mathcal{L}_i}{\partial \beta_i} = -\hat{\beta}_i^o + \beta_i - \gamma = 0$$

This gives us

$$\begin{aligned}\hat{\beta}_i^{\text{lasso}} &= \begin{cases} \hat{\beta}_i^o + \gamma & \text{if } \hat{\beta}_i^o + \gamma \leq 0, \\ 0 & \text{otherwise.} \end{cases} \\ &= (-\hat{\beta}_i^o - \gamma)^+ \\ &= \text{sgn}(\hat{\beta}_i^o)(|\hat{\beta}_i^o| - \gamma)^+\end{aligned}$$

2.5. Standard Errors - Bootstrap

In general, LASSO estimator is a non-linear and non-differentiable function. It's difficult to obtain an accurate estimate of its SE.

One way to get the SE is by bootstrap.

Let $Z_i = (x_i, y_i)$, $i = 1, \dots, n$. The steps for calculating the LASSO bootstrap standard error are as follows. First, pick a large number B , and for $b = 1, \dots, B$:

- ▶ Draw a bootstrap sample $(\tilde{Z}_1^{(b)}, \dots, \tilde{Z}_n^{(b)})$ from (Z_1, \dots, Z_n) .
- ▶ Perform LASSO and get the estimated coefficients $\tilde{\beta}^{(b)}$ on $(\tilde{Z}_1^{(b)}, \dots, \tilde{Z}_n^{(b)})$.
- ▶ Then we estimate the standard error of $\tilde{\beta}^{(b)}$ as follows:

$$SE(\tilde{\beta}^{(b)}) = \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\tilde{\beta}^{(b)} - \frac{1}{B} \sum_{r=1}^B \tilde{\beta}^{(r)} \right) \left(\tilde{\beta}^{(b)} - \frac{1}{B} \sum_{r=1}^B \tilde{\beta}^{(r)} \right)^T}$$

2.5. Standard Errors - Approximate Form

We rewrite the penalty constraint for the LASSO problem as

$$\hat{\beta} := \arg \min \left\{ \sum_{i=1}^N \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ s.t. } \frac{\sum_j |\beta_j|^2}{|\beta_j|} \leq t \quad (4)$$

Hence, at the lasso estimate, we may approximate the solution by a ridge regression of the form $\beta^* = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{W}^{-1})^{-1} \mathbf{X}^T \mathbf{Y}$.

where $\mathbf{W} = \text{diag}(|\hat{\beta}_j^{lasso}|)$, and \mathbf{W}^{-1} denotes the generalized inverse of \mathbf{W} . The covariance matrix of the estimates may then be approximated by

$$(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{W}^{-1})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{W}^{-1})^{-1} \hat{\sigma}^2 \quad (5)$$

4. Prediction error and estimation of t

There are three methods: 1. cross-validation 2. generalized cross-validation 3. Stein's unbiased risk estimation (SURE)

Cross-validation

Suppose $Y = \eta(X) + \epsilon$ where $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$. The mean-squared error of estimate $\hat{\eta}(X)$ is defined by

$$ME = E(\hat{\eta}(X) - \eta(X))^2$$

and the prediction error is

$$PE = E(Y - \hat{\eta}(X))^2 = ME + \sigma^2 \quad (6)$$

In LASSO

$\eta(X) = X\beta$ is a linear model, the ME has a simple form:

$$ME = (\hat{\beta} - \beta)^T V (\hat{\beta} - \beta)$$

where V is the population covariance matrix of X .

Generalized cross-validation

We approximate the lasso solution by a ridge regression of the form

$$\beta^* = (X^T X + \lambda W^{-1})^{-1} X^T Y \quad (7)$$

Therefore the number of effective parameters in the constrained fit β^* may be approximated by

$$\rho(t) = \text{tr} \left(X(X^T X + \lambda W^{-1})^{-1} X^T \right) \quad (8)$$

Letting $\text{rss}(t)$ be the residual sum of squares for the constrained fit with constraint t , we construct the generalized cross-validation style statistic

$$\text{GCV}(t) = \frac{1}{N} \frac{\text{rss}(t)}{\left(1 - \frac{\rho(t)}{N}\right)^2} \quad (9)$$

Stein's Unbiased Risk Estimation (SURE)

Let $\hat{\boldsymbol{\mu}}$ be the estimator of $\boldsymbol{\mu}$. write $\hat{\boldsymbol{\mu}} = \mathbf{z} + g(\mathbf{z})$, where g is an almost differentiable function from $\mathbb{R}^p \rightarrow \mathbb{R}^p$.

$$\mathbb{E}_{\boldsymbol{\mu}} ||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||^2 = p + \mathbb{E}_{\boldsymbol{\mu}} \left(||g(\mathbf{z})||^2 + 2 \sum_{i=1}^p \frac{dg_i}{dz_i} \right) \quad (10)$$

Denotes the estimated standard error of $\hat{\beta}_j^o$ by

$$SE(\hat{\beta}_j^o) = \hat{\tau} := \frac{\hat{\sigma}}{\sqrt{n}}, \hat{\sigma}^2 = \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N - p}$$

We may derive the formula as an approximately unbiased estimate of the risk:

$$R(\hat{\beta}(\gamma)) \approx \hat{\tau}^2 \left\{ p - 2\# \left(j : \frac{|\hat{\beta}_j^o|}{\hat{\tau}} < \gamma \right) + \sum_{j=1}^p \max \left(\left| \frac{\hat{\beta}_j^o}{\hat{\tau}} \right|, \gamma \right)^2 \right\} \quad (11)$$

Stein's Unbiased Risk Estimation (SURE)

$$R(\hat{\beta}(\gamma)) \approx \hat{\tau}^2 \left\{ p - 2\# \left(j : \frac{|\hat{\beta}_j^o|}{\hat{\tau}} < \gamma \right) + \sum_{j=1}^p \max \left(\left| \frac{\hat{\beta}_j^o}{\hat{\tau}} \right|, \gamma \right)^2 \right\}$$

where $\hat{\beta}_j(\gamma) = \text{sgn}(\hat{\beta}_j^o) \left(\left| \frac{\hat{\beta}_j^o}{\hat{\tau}} \right| - \gamma \right)^+$

Hence an estimate of γ can be obtained as the minimizer of $R(\hat{\beta}(\gamma))$

$$\hat{\gamma} = \arg \min_{\gamma \geq 0} R(\hat{\beta}(\gamma))$$

From this we obtain an estimate of the lasso parameter t :

$$\hat{t} = \sum_{j=1}^p (|\hat{\beta}_j^o| - \hat{\gamma})^+$$

6. Algorithms for Finding LASSO Solutions

We fix $t \geq 0$,

$$(\hat{\alpha}, \hat{\beta}) := \arg \min \left\{ \sum_{i=1}^N (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right\} \text{ s.t. } \sum_j |\beta_j| \leq t$$

$$\hat{\beta} = \arg \min \left\{ \sum_{i=1}^N \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ s.t. } \delta_i^T \beta \leq t$$

where $\delta_i = (\pm 1, \pm 1, \dots, \pm 1)$, $i = 1, 2, \dots, m$ be the p -tuples.

For a given β , let equality set $E = \{i : \delta_i^T \beta = t\}$ and slack set $S = \{i : \delta_i^T \beta < t\}$

6. Algorithms for Finding LASSO Solutions

The algorithm starts with $E = \{i_0\}$ where $\delta_{i_0} = \text{sign}(\hat{\beta}^o)$, $\hat{\beta}^o$ being the overall LS estimate.

Find $\hat{\beta}$ to minimize $\sum_{i=1}^N (y_i - \sum_j \beta_j x_{ij})^2$ s.t. $\delta_{i_0}^T \beta \leq t$

While $\sum |\hat{\beta}_j| > t$:

 add i to the set E where $\delta_i = \text{sign}(\hat{\beta})$

 Find $\hat{\beta}$ to minimize $g(\beta)$ s.t. $G_E \beta \leq t \mathbf{1}$

end

Where G_E is the matrix whose rows are δ_i for $i \in E$

7. Simulation

The author gave us four examples:

Example 1

Simulated 50 data sets consisting of 20 observations from the model

$$y = \beta^T \mathbf{x} + \sigma \epsilon$$

where $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\epsilon \sim N(0, 1)$ The correlation between x_i and x_j was $\rho^{|i-j|}$ with $\rho = 0.5$ and set $\sigma = 3$

Example 2

Same model setting as example 1, but with $\beta_j = 0.85, \forall j$ and $\sigma = 3$

Example 3

Same model setting as example 1, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)$ and $\sigma = 2$

7. Simulation

Example 4

Simulated 50 data sets each having 100 observations and 40 variables. We define predictor $x_{ij} = z_{ij} + z_i$ where z_{ij} and z_i are independent standard normal variates. This induced a pairwise correlation of 0.5 among the predictors. The coefficient vector was $\beta = (0, 0, \dots, 0, 2, 2, \dots, 2, 0, 0, \dots, 0, 2, 2, \dots, 2)$, there being 10 repeats in each block. Finally, we defined $y = \beta^T \mathbf{x} + 15\epsilon$ where ϵ was standard normal.

Models Performance

Example 1

TABLE 3
Results for example 1†


<i>Method</i>	<i>Median mean-squared error</i>	<i>Average no. of 0 coefficients</i>	<i>Average \hat{s}</i>
Least squares	2.79 (0.12)	0.0	—
Lasso (cross-validation)	2.43 (0.14)	3.3	0.63 (0.01)
Lasso (Stein)	2.07 (0.10)	2.6	0.69 (0.02)
Lasso (generalized cross-validation) 👍	1.93 (0.09)	2.4	0.73 (0.01)
Garotte	2.29 (0.16)	3.9	—
Best subset selection	2.44 (0.16)	4.8	—
Ridge regression	3.21 (0.12)	0.0	—

†Standard errors are given in parentheses.

Models Performance

Example 2

TABLE 6
Results for example 2†


<i>Method</i>	<i>Median mean-squared error</i>	<i>Average no. of 0 coefficients</i>	<i>Average \hat{s}</i>
Least squares	6.50 (0.64)	0.0	—
Lasso (cross-validation)	5.30 (0.45)	3.0	0.50 (0.03)
Lasso (Stein)	5.85 (0.36)	2.7	0.55 (0.03)
Lasso (generalized cross-validation)	4.87 (0.35)	2.3	0.69 (0.23)
Garotte	7.40 (0.48)	4.3	—
Subset selection	9.05 (0.78)	5.2	—
Ridge regression	 2.30 (0.22)	0.0	—

†Standard errors are given in parentheses.

Models Performance

Example 3

TABLE 7
Results for example 3†


<i>Method</i>	<i>Median mean-squared error</i>	<i>Average no. of 0 coefficients</i>	<i>Average \hat{s}</i>
Least squares	2.89 (0.04)	0.0	—
Lasso (cross-validation)	0.89 (0.01)	3.0	0.50 (0.03)
Lasso (Stein)	1.26 (0.02)	2.6	0.70 (0.01)
Lasso (generalized cross-validation)	1.02 (0.02)	3.9	0.63 (0.04)
Garotte	0.52 (0.01)	5.5	—
Subset selection	 0.64 (0.02)	6.3	—
Ridge regression	3.53 (0.05)	0.0	—

†Standard errors are given in parentheses.

Models Performance

Example 4

TABLE 8
Results for example 4†

<i>Method</i>	<i>Median mean-squared error</i>	<i>Average no. of 0 coefficients</i>	<i>Average \hat{s}</i>
Least squares	137.3 (7.3)	0.0	—
Lasso (Stein)	80.2 (4.9)	14.4	0.55 (0.02)
Lasso (generalized cross-validation)	64.9 (2.3)	13.6	0.60 (0.88)
Garotte	94.8 (3.2)	22.9	—
Ridge regression	 57.4 (1.4)	0.0	—

†Standard errors are given in parentheses.

Discussion

The author examined the relative merits of the methods in three different scenarios:

1. **small number of large effects** – subset selection does best here the lasso not quite as well and ridge does quite poorly.
2. **small to moderate number of moderate-sized effects** – the lasso does best, followed by ridge and then subset selection.
3. **large number of small effects** – ridge does best by a good margin, followed by the lasso and then subset selection.

Thank you