Regression Shrinkage and Selection via the Lasso

Boyu Chen

2023-06-12

Introduction

The "lasso" minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint it tends to produce some coefficients that are exactly 0 and hence gives interpretable models.

The LASSO

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

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

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:

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

In matrix form:

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

LASSO estimators have no closed form unless X is orthogonal.

Orthogonal Design Case

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

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

For a certain i, the Lagrangian function is

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

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

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

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

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

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 to get the SE is by bootstrap.

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

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

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

2.5. Standard Errors - Approximate Form

We rewrite the penalty constraint for the LASSO problem as

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

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

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

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

4. Prediction error and estimation of t

Recall: t is the number of the non-zero predictors

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

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

In LASSO

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

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

where V is the population covariance matrix of X.

Generalized cross-validation

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

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

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

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

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

$$GCV(t) = \frac{1}{N} \frac{rss(t)}{(1 - \frac{\rho(t)}{N})^2} \tag{9}$$

Stein's Unbiased Risk Estimation (SURE)

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

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

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

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

For the orthogonal case, we may derive the formula as an approximately unbiased estimate of the risk:

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

Stein's Unbiased Risk Estimation (SURE)

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

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 obtian an estimate of the lasso parameter t:

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

Discussion for SURE

Although the derivation of $\hat{t}$ assumes an orthogonal design, we may still try to use it in the usual non-orthogonal setting. Since the predictors have been standardized, the optimal value of t is roughly a function of the overall signal-to-noise ratio in the data, and it should be relatively insensitive to the covariance of $\mathbf{X}$.

The Stein method enjoys a significant computational advantage over the cross-validation-based estimate of t.

6. Algorithms for Finding LASSO Solutions

We fix $t \geq 0$, the problem (12) can be expressed as a least squares problem with 2^p inequality constraints, corresponding to the 2^p different possible signs for the β_js.

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) := \arg\min \left\{ \sum_{i=1}^{N} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \right\} \text{ s.t. } \sum_j |\beta_j| \leq t \tag{12}$$

Then the condition $\sum |\beta_j| \leq t$ is equivalent to $\boldsymbol{\delta}_i^T \boldsymbol{\beta} \leq t$ for all i. where $\boldsymbol{\delta}_i = (\pm 1, \pm 1, ..., \pm 1), i = 1, 2, ..., 2^p$ be the p-tuples.

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

6. Algorithms for Finding LASSO Solutions

The algorithms starts with $E = \{i_0\}$ where $\boldsymbol{\delta}_{i0} = \text{sign}(\hat{\boldsymbol{\beta}}^o)$, $\hat{\boldsymbol{\beta}}^o$ being the overall LS estimate.

Algorithms

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

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

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

end

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

$\mathbf{1}$ is a vector of 1s of length equal to the number of rows of G_E.

7. Simulation

The author gave us four examples:

Example 1

Simulated 50 data sets consisting of 20 observations from the model

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

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ 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 $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)$ and $\sigma = 2$

7. Simulation

Example 4

Simulated 50 data sets each having 100 observations and 40 variables. We defind 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 $\boldsymbol{\beta} = (0, 0, \ldots, 0, 2, 2, \ldots, 2, 0, 0, \ldots, 0, 2, 2, \ldots, 2)$, there being 10 repeats in each block.Finally, we defined $y = \boldsymbol{\beta}^T \boldsymbol{x} + 15\epsilon$ where ϵ was standard normal.

Models Performance

Example 1

TABLE 3
Results for example 1†

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average $\hat{s}$*
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†

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average $\hat{s}$*
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†

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average $\hat{s}$*
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†

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average* $\hat{s}$
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 samll effects** – ridge does best by a good margin, followed by the lasso and then subset selection.

Thank you