Untitled

Boyu Chen

2023-06-09

The LASSO

Def. the data $(\mathbf{x}_i, y_i)$, $i = 1, 2, ..., N$ 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 $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, the LASSO estimate $(\hat{\alpha}, \hat{\beta})$:

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

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:

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

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

Orthogonal Design Case

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

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

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 $\hat{\beta}_i^o < 0$, $\beta_i \leq 0$.

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

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

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 $\tilde{\beta}^{(b)}$ 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

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

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

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

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