

BZAN 615 - Homework 2

Due: March 8, 2024

1 Variance of the Least Squares

Show that

$$\mathbb{E} \left[\left(x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right)^2 \right] = \sigma^2 \mathbb{E} \left[x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 \right]$$

where ϵ is the error term that is independent of \mathbf{X} and x_0 and follows multivariate normal distribution with independent components and common variance of σ^2 . \mathbf{X} is n by p data matrix, and x_0 is the independent new p dimensional observation.

Proof. We have

$$\begin{aligned} \mathbb{E} \left[\left(x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right)^2 \right] &= \mathbb{E} \left[\left(x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right)^T \left(x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right) \right] \\ &= \mathbb{E} \left[\epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right] \\ &= \mathbb{E} \left[\text{tr} \left(\epsilon^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \right) \right] \\ &= \mathbb{E} \left[\text{tr} \left(\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \epsilon^T \right) \right] \\ (\text{Independence}) &= \mathbb{E} \left[\text{tr} \left(\mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\sigma^2 \mathbf{I}) \right) \right] \\ &= \sigma^2 \mathbb{E} \left[\text{tr} \left((\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \right) \right] \\ &= \sigma^2 \mathbb{E} \left[\text{tr} \left((\mathbf{X}^T \mathbf{X})^{-1} x_0 x_0^T \right) \right] \\ (\text{tr}(BA^T) = \text{tr}(A^T B)) &= \sigma^2 \mathbb{E} \left[\text{tr} \left(x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 \right) \right] \\ &= \sigma^2 \mathbb{E} \left[x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 \right]. \end{aligned}$$

□

2 Ridge Regression

Recall that the Ridge regression estimator is given by

$$\hat{\beta}^{\text{ridge}} = \arg \min_{\beta} RSS(\beta) + \lambda \|\beta\|_2^2.$$

Then, show that the explicit solution of this equation is given by

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

Proof. Denoting $f(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \beta^T \beta$, we have

$$\begin{aligned} \arg \min_{\beta} f(\beta) &= \arg \min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda \beta^T \beta \\ &= \arg \min_{\beta} \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\beta - \beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X}\beta + \lambda \beta^T \beta \\ &= \arg \min_{\beta} \mathbf{Y}^T \mathbf{Y} - 2\mathbf{Y}^T \mathbf{X}\beta + \beta^T \mathbf{X}^T \mathbf{X}\beta + \lambda \beta^T \beta. \end{aligned}$$

Then we have

$$\begin{aligned} 0 &= \frac{\partial f(\beta)}{\partial \beta} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X}\beta + 2\lambda \beta \\ &\Rightarrow 0 = -\mathbf{X}^T \mathbf{Y} + \mathbf{X}^T \mathbf{X}\beta + \lambda \beta \\ &\Rightarrow \mathbf{X}^T \mathbf{X}\beta + \lambda \beta = \mathbf{X}^T \mathbf{Y} \\ &\Rightarrow (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})\beta = \mathbf{X}^T \mathbf{Y} \\ &\Rightarrow \beta = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}, \end{aligned}$$

where the last step is valid since $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$ is invertible if $\lambda > 0$. \square

3 Lasso Coefficient Profile

When \mathbf{X} has orthonormal columns (i.e. $\mathbf{X}^T \mathbf{X} = \mathbf{I}_p$), complete the proof that $\hat{\beta}_j^{\text{lasso}} = \text{sign}(\hat{\beta}^{lr}) \left(|\hat{\beta}^{lr}| - \lambda \right)_+$ where $x_+ = x$ if $x > 0$ and $x_+ = 0$ if $x \leq 0$

Proof. We have $\mathbf{X}^T \mathbf{X} = \mathbf{I}_p$, then we have $\hat{\beta}^{lr} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y}$. Then we have

then the explicit solution of the Lasso estimator is given by

$$\begin{aligned}
\hat{\beta}^{\text{lasso}} &= \arg \min_{\beta} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \\
&= \arg \min_{\beta} \frac{1}{2} (\mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X}\beta - \beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X}\beta) + \lambda \sum_{j=1}^p |\beta_j| \\
&= \arg \min_{\beta} \frac{1}{2} \left(-(\hat{\beta}^{lr})^T \beta - \beta^T \hat{\beta}^{lr} + \beta^T \beta \right) + \lambda \sum_{j=1}^p |\beta_j| \\
&= \arg \min_{\beta} -\hat{\beta}^{lr} \cdot \beta + \frac{1}{2} \beta^T \beta + \lambda \sum_{j=1}^p |\beta_j| \\
&= \arg \min_{\beta} \sum_{j=1}^p \left(-\hat{\beta}_j^{lr} \beta_j + \frac{1}{2} \beta_j^2 + \lambda |\beta_j| \right).
\end{aligned}$$

Since the loss function is separable, it suffices to minimize each component of β separately. Consider the minimization of β_j for $j = 1, \dots, p$. We have

$$f'(\beta_j) \equiv \frac{\partial}{\partial \beta_j} \left(-\hat{\beta}_j^{lr} \beta_j + \frac{1}{2} \beta_j^2 + \lambda |\beta_j| \right) = \begin{cases} -\hat{\beta}_j^{lr} + \beta_j + \lambda, & \text{if } \beta_j > 0 \\ -\hat{\beta}_j^{lr} + \beta_j - \lambda, & \text{if } \beta_j < 0 \end{cases}$$

Then we have

$$f'(\beta_j) \leq 0 \Leftrightarrow \begin{cases} \beta_j \leq \hat{\beta}_j^{lr} - \lambda, & \text{if } \beta_j > 0 \\ \beta_j \leq \hat{\beta}_j^{lr} + \lambda, & \text{if } \beta_j < 0 \end{cases}$$

And if $\hat{\beta}_j^{lr} \in [-\lambda, \lambda]$, then we have

$$f'(\beta_j) = \begin{cases} > 0, & \text{if } \beta_j > 0 \\ = 0, & \text{if } \beta_j = 0 \\ < 0, & \text{if } \beta_j < 0 \end{cases}$$

Thus, combining the above condition, we have the minimizer of $f(\beta_j)$

$$\arg \min_{\beta} f(\beta_j) = \begin{cases} \hat{\beta}_j^{lr} - \lambda, & \text{if } \hat{\beta}_j^{lr} > \lambda \\ 0, & \text{if } \hat{\beta}_j^{lr} \in [-\lambda, \lambda] \\ \hat{\beta}_j^{lr} + \lambda, & \text{if } \hat{\beta}_j^{lr} < -\lambda. \end{cases}$$

This can be written as

$$\arg \min_{\beta_j} f(\beta_j) = \text{sign}(\hat{\beta}_j^{lr}) \left(|\hat{\beta}_j^{lr}| - \lambda \right)_+,$$

for $j = 1, \dots, p$.

□

4 Review: Eigenvalues of $\mathbf{X}^T\mathbf{X}$

Show that all the eigenvalues of a matrix $\mathbf{X}^T\mathbf{X}$ are non-negative.

Proof. Suppose \mathbf{X} is a $n \times p$ matrix with rank $q \leq p$. Consider a SVD of \mathbf{X} , i.e., $\mathbf{X} = UDV^T$, where U is a $n \times p$ matrix, D is a $p \times p$ diagonal matrix, and V is a $p \times p$ matrix. Then we have

$$\begin{aligned}\mathbf{X}^T\mathbf{X} &= VD^TU^TUDV^T \\ &= VD^2V^T \\ &= V\Lambda V^T,\end{aligned}$$

where Λ is a $p \times p$ diagonal matrix with $\Lambda_{ii} = D_{ii}^2$. Then for any eigenvector v of $\mathbf{X}^T\mathbf{X}$, we have

$$\begin{aligned}\mathbf{X}^T\mathbf{X}v &= \lambda v \\ V\Lambda V^T v &= \lambda v \\ \Lambda V^T v &= \lambda V^T v \\ \Rightarrow \lambda &= \Lambda_{ii} \geq 0.\end{aligned}$$

That is, all the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are non-negative. If \mathbf{X} is of full rank, i.e., $q = p$, then all the eigenvalues of $\mathbf{X}^T\mathbf{X}$ are positive. Otherwise, the rank of \mathbf{X} is $q < p$, and the last $p - q$ eigenvalues of $\mathbf{X}^T\mathbf{X}$ are zero.

□