Regularization and variable selection via the elastic net[1]

Hsieh Yuan-Hao

2025/02/21

[1] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the Elastic Net. Journal of the Royal Statistical Society Series B 67 301–320.

Soft Thresholding and L-1 Regularization

- Lasso in Lagrange Form:

$$RSS(\hat{\beta}, \lambda) = (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y})^{\mathsf{T}}(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda\|\hat{\boldsymbol{\beta}}\|$$
$$= \underbrace{\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p}\hat{\beta}_j x_{ij}\right)^2}_{OLS} + \underbrace{\lambda\sum_{j=1}^{p}|\hat{\beta}_j|}_{L_1-norm}$$

where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ $\boldsymbol{y} \in \mathbb{R}^{n\times 1}$ and $\lambda > 0$.

a. Consider the scalar minimization problem for each component β_i in OLS part:

$$\frac{\partial RSS^{OLS}}{\partial \hat{\beta}_j} = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{j=0}^{p}\hat{\beta}_j x_{ij}\right) = -2\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j}\left(\hat{\beta}_k x_{ik}\right) - \hat{\beta}_j x_{ij}\right)$$

$$= \underbrace{-2\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j}(\hat{\beta}_k x_{ik})\right)}_{\rho_j} + 2\hat{\beta}_j \underbrace{\sum_{i=1}^{n}(x_{ij})^2}_{z_j} \triangleq -2\rho_j + 2\hat{\beta}_j z_j$$

Derivation of the Soft Thresholding Operator

b. In L1-norm part

$$\frac{\partial RSS^{L_1 Norm}}{\partial \hat{\beta}_j} = \begin{cases} -\lambda, & \text{when } \hat{\beta}_j < 0 \\ [-\lambda, \lambda], & \text{when } \hat{\beta}_j = 0 \\ \lambda, & \text{when } \hat{\beta}_j > 0 \end{cases}$$

For $\hat{\beta}_i \neq 0$**:** The function is differentiable, and have two cases:
Case 1: $\hat{\beta}_i > 0$ Then the condition becomes

$$-2\rho_j + 2\hat{\beta}_j z_j + \lambda = 0 \Rightarrow \hat{\beta}_j^{\,*} = \frac{1}{z_j}(\rho_j - \frac{\lambda}{2})$$

For $\hat{\beta}_j^{\,*} > 0, \rho_j - \frac{\lambda}{2} > 0 \Rightarrow \rho > \frac{2}{\lambda}$

Case 2: $\hat{\beta}_j < 0$ Then the condition becomes

$$-2\rho_j + 2\hat{\beta}_j z_j - \lambda = 0 \Rightarrow \hat{\beta}_j^{\,*} = \frac{1}{z_j}(\rho_j + \frac{\lambda}{2})$$

For $\hat{\beta}_j^{\,*} < 0, \rho_j + \frac{\lambda}{2} < 0 \Rightarrow \rho_j < -\frac{2}{\lambda}$

Subgradients Analysis and Final Soft Thresholding Formula

- **For $\hat{\beta}_i = 0$:** The function is undifferentiable,
  The condition becomes $[-2\rho_j - \lambda, -2\rho_j + \lambda]$ contain $0 \Rightarrow$
  $-2\rho_j - \lambda < 0, -2\rho_j + \lambda > 0 \Rightarrow -\frac{\lambda}{2} < \rho_j < \frac{\lambda}{2}$
- **Final Soft Thresholding Operator:** Combining the cases, the closed-form solution for each $\hat{\beta}_i^*$ is:

$$\hat{\beta}_j^* = \begin{cases} \frac{1}{z_j}\left(\rho_j + \frac{\lambda}{2}\right), & \text{when } \rho_j < -\frac{\lambda}{2} \\ 0, & \text{when } -\frac{\lambda}{2} < \rho_j < \frac{\lambda}{2} \\ \frac{1}{z_j}\left(\rho_j - \frac{\lambda}{2}\right), & \text{when } \rho_j > \frac{\lambda}{2} \end{cases}$$

- This operator is widely known as the **Soft Thresholding** operator and serves as the solution to the LASSO problem for each component.

Implementation

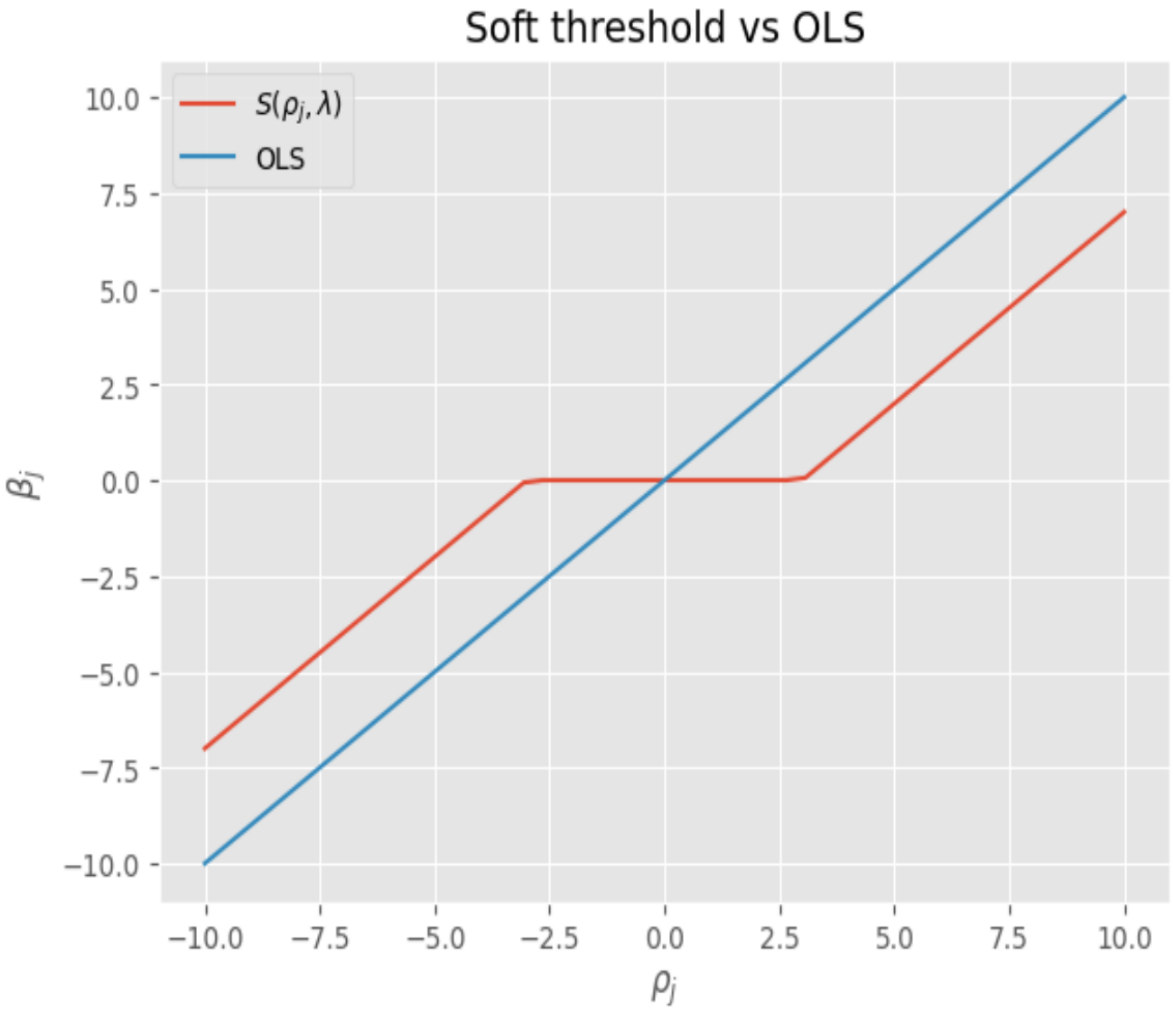

Figure: Comparison of Soft-Thresholding and OLS Solutions

Algorithm

Lasso coordinate descent update Algorithm[2]

Repeat until convergence or max number of iterations:

For $j = 1, 2, \ldots, p$

Compute $\rho_j = \sum_{i=1}^{n} x_{ij} \left(y_i - \sum_{k \neq j} \left(\hat{\beta}_k x_{ik} \right) \right)$

Compute $z_j = \sum_{i=1}^{n} (x_{ij})^2$

Set

$$\hat{\beta}_j^* = \frac{1}{z_j} S\left(\rho_j, \frac{\lambda}{2}\right) = \frac{1}{z_j} \text{sign}(\rho_j) \left(|\rho_j| - \frac{\lambda}{2} \right)^+$$

[2] Xavier Bourret Sicotte. Lasso regression: derivation of the coordinate descent update rule.

Go to elasticnet

- Coordinate descent:

$$RSS(\hat{\boldsymbol{\beta}}, \lambda_1, \lambda_2) = (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y})^T(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda_2\|\hat{\boldsymbol{\beta}}\|^2 + \lambda_1\|\hat{\boldsymbol{\beta}}\|$$

$$= \underbrace{\sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\hat{\beta}_j x_{ij})^2 + \lambda_2 \sum_{j=1}^{p}\hat{\beta}_j{}^2}_{A} + \underbrace{\lambda_1 \sum_{j=1}^{p}|\hat{\beta}_j|}_{B}$$

$$\frac{\partial RSS^A}{\partial \hat{\beta}_j} = -2\sum_{i=1}^{n} x_{ij}(y_i - \sum_{j=0}^{p}\hat{\beta}_j x_{ij}) + 2\lambda_2 \sum_{j=1}^{p}\hat{\beta}_j$$

$$= -2\sum_{i=1}^{n} x_{ij}(y_i - \sum_{k\neq j}(\hat{\beta}_k x_{ik})) + 2\hat{\beta}_j \sum_{i=1}^{n}(x_{ij})^2 + 2\lambda_2\hat{\beta}_j \approx -2\rho_j + 2\hat{\beta}_j(z_j + \lambda_2)$$

$$\frac{\partial RSS^B}{\partial \hat{\beta}_j} = \begin{cases} -\lambda_1, & \text{when } \hat{\beta}_j < 0 \\ [-\lambda_1, \lambda_1], & \text{when } \hat{\beta}_j = 0 \\ \lambda_1, & \text{when } \hat{\beta}_j > 0 \end{cases}$$

Go to elasticnet

$$\frac{\partial RSS}{\partial \hat{\beta}_j} = \begin{cases} -2\rho_j + 2\hat{\beta}_j(z_j + \lambda_2) - \lambda_1, & \text{when } \hat{\beta}_j < 0 \\ [-2\rho_j - \lambda_1, -2\rho_j + \lambda_1], & \text{when } \hat{\beta}_j = 0 \\ -2\rho_j + 2\hat{\beta}_j(z_j + \lambda_2) + \lambda_1, & \text{when } \hat{\beta}_j > 0 \end{cases}$$

- In order to be optimal solution, by KKT condition:
  1. When $\hat{\beta}_j < 0$:
     $-2\rho_j + 2\hat{\beta}_j(z_j + \lambda_2) - \lambda_1 = 0 \Rightarrow \hat{\beta}_j^* = \frac{1}{(z_j + \lambda_2)}(\rho_j + \frac{\lambda_1}{2})$
     For $\hat{\beta}_j^* < 0$,we need $\rho + \frac{\lambda_1}{2} < 0 \Rightarrow \rho_j < -\frac{\lambda_1}{2}$
  2. When $\hat{\beta}_j = 0$: $\Rightarrow \hat{\beta}_j^* = 0$
     For $\hat{\beta}_j^* = 0$, we need $[-2\rho_j - \lambda_1, -2\rho_j + \lambda_1]$ to contain 0
     $$\Rightarrow -2\rho_j - \lambda_1 \leq 0, \quad -2\rho_j + \lambda_1 \geq 0 \Rightarrow -\frac{\lambda_1}{2} \leq \rho_j \leq \frac{\lambda_1}{2}$$
  3. When $\hat{\beta}_j > 0$:$-2\rho_j + 2\hat{\beta}_j(z_j + \lambda_2) + \lambda_1 = 0$
     $\Rightarrow \hat{\beta}_j^* = \frac{1}{(z_j + \lambda_2)}\left(\rho_j - \frac{\lambda_1}{2}\right)$.
     For $\hat{\beta}_j^* < 0$, we need $\rho_j - \frac{\lambda_1}{2} > 0$:$\Rightarrow \rho_j \geq \frac{\lambda_1}{2}$

Go to elasticnet

$$\hat{\beta}_j^* = \begin{cases} \frac{1}{(z_j+\lambda_2)}\left(\rho_j + \frac{\lambda_1}{2}\right), & \text{when } \rho_j < -\frac{\lambda_1}{2} \\ 0, & \text{when } -\frac{\lambda_1}{2} \leq \rho_j \leq \frac{\lambda_1}{2} \\ \frac{1}{(z_j+\lambda_2)}\left(\rho_j - \frac{\lambda_1}{2}\right), & \text{when } \rho_j > \frac{\lambda_1}{2} \end{cases}$$

$\Rightarrow$ Soft-thresholding: $S(\rho, \lambda) = sign(\rho)(|\rho| - \lambda)^+$

- **Algorithm of Elastic net via coordinate descent:**
  Repeat until convergence or max number of iterations:
  For $j = 1, 2, ..., p$
  compute $\rho_j = \sum_{i=1}^n x_{ij}(y_j - \sum_{k \neq j}(\hat{\beta}_k x_{ik}))$
  compute $z_j = \sum_{i=1}^n (x_{ij})^2$
  Set $\hat{\beta_j}^* = \frac{1}{(z_j+\lambda_2)} S(\rho_j, \frac{\lambda_1}{2}) = \frac{1}{(z_j+\lambda_2)} sign(\rho)(|\rho_j| - \frac{\lambda_1}{2})^+$

Simulation

- The simulated data comes from the true model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1).$$

- Each simulated dataset is divided into training set, validation set, and test set. Models were fitted on the training set only, and the validation data were used to select the tuning parameters.
- The test error (the mean-squared error) was computed on the test set.

Simulation Example 1, 2 and 3

- **Simulation example 1:** 50 data sets were simulated consisting of 20/20/200 observations and 8 predictors:

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0), \quad \sigma = 3, \quad \text{and} \quad \text{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (0.5)^{|i-j|}$$

- **Simulation example 2:** Same as example 1, except $\beta_j = 0.85$ for all j.
- **Simulation example 3:** 50 data sets were simulated consisting of 100/100/400 observations and 40 predictors:

$$\boldsymbol{\beta} = \left(\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10} \right), \quad \sigma = 15,$$

$$\text{cor}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5 \quad \text{for all } i, j = 1, \ldots, 40.$$

Simulation Example 4

- **Simulation example 4:** 50 data sets were simulated consisting of 50/50/400 observations and 40 predictors:

$$\beta = \left(\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25} \right), \quad \sigma = 15.$$

$$\boldsymbol{x}_i = Z_1 + \epsilon_i^x, \quad Z_1 \sim \mathcal{N}(0,1), \quad i = 1, \ldots, 5.$$

$$\boldsymbol{x}_i = Z_2 + \epsilon_i^x, \quad Z_2 \sim \mathcal{N}(0,1), \quad i = 6, \ldots, 10.$$

$$\boldsymbol{x}_i = Z_3 + \epsilon_i^x, \quad Z_3 \sim \mathcal{N}(0,1), \quad i = 11, \ldots, 15.$$

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 16, \ldots, 40.$$

$$\epsilon_i^x \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.01), \quad i = 1, \ldots, 15.$$

Simulated Examples - Median MSE(from paper)

Method	**Ex.1**	**Ex.2**	**Ex.3**	**Ex.4**
Ridge	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic Net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
Naive Elastic Net	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

Table: Median MSE for the simulated examples and 4 methods in paper

Method	**Ex.1**	**Ex.2**	**Ex.3**	**Ex.4**
Ridge	12.71 (0.5)	11.47 (0.38)	251.31 (2.77)	307.52 (6.41)
Lasso	12.89 (0.5)	13.4 (0.24)	269.4 (3.39)	294.94 (7.02)
Elastic Net	12.41 (0.38)	13.56 (0.19)	268.63 (3.08)	269.21 (7.06)
Naive Elastic Net	12.8 (0.67)	11.57 (0.37)	260.24 (4.27)	289.49 (5.43)

Table: Median MSE for the simulated examples and 4 methods in R

the boxplot of MSE from paper

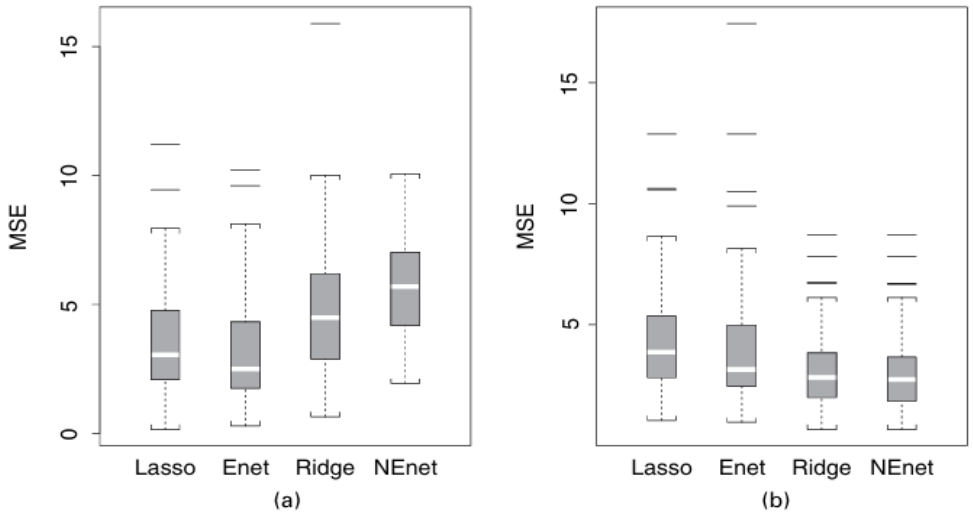

Fig. 4. Comparing the accuracy of prediction of the lasso, the elasic net (Enet), ridge regression and the naïve elastic net (NEnet) (the elastic net outperforms the lasso in all four examples): (a) example 1; (b) example 2; (c) example 3; (d) example 4

Figure: Median MSE and Bootstrap SE (B=500)

the boxplot of MSE in R

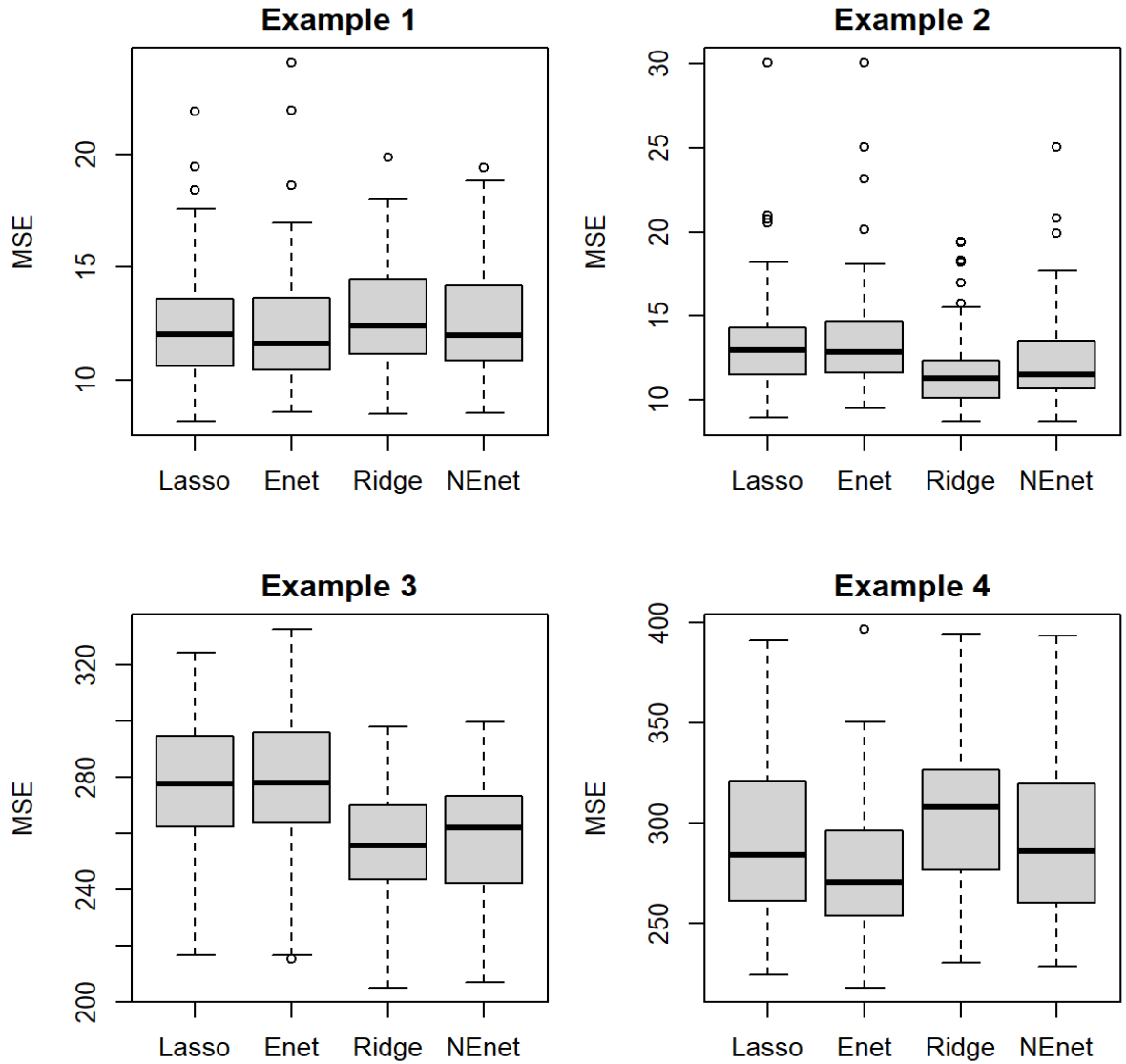

Figure: Median MSE and Bootstrap SE (B=500)

Conclusion

- **Correlation Structures**:
  1. Examples with stronger predictor correlations (e.g., Example3) favor methods like elastic net or ridge that better handle collinearity.
  2. The lasso can struggle if there are groups of highly correlated variables—often it will pick only one from the group, causing higher MSE if the others also matter.
- **Method Behavior**:
  1. **Ridge** shrinks coefficients continuously, so it handles correlation well but never sets coefficients exactly to zero.
  2. **Lasso** can zero out coefficients for variable selection but may suffer when predictors are strongly correlated.
  3. **Elastic Net** combines ridge and lasso's strengths—often outperforming lasso when there is moderate/high correlation.
  4. **Naïve EN** includes the same penalty terms but does not rescale the final coefficients—often resulting in extra shrinkage.