The grouping effect

- We consider the generic penalization method

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} |\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|^2 + \lambda J(\boldsymbol{\beta}) \tag{5}$$

  where $J(\cdot)$ is positive valued for $\boldsymbol{\beta} \neq 0$

- *Lemma1* Assume that $\boldsymbol{x}_i = \boldsymbol{x}_j, i, j \in \{1, ..., p\}$.
  - (a) If $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, $\forall \lambda > 0$.
  - (b) If $J(\boldsymbol{\beta}) = |\boldsymbol{\beta}|_1$, then $\hat{\beta}_i \hat{\beta}_j \geq 0$ and $\hat{\boldsymbol{\beta}}^*$ is is another minimizer of equation (5), where

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot s & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & \text{if } k = j. \end{cases}$$

  for any $s \in [0, 1]$

Proof of the lemma1

- Part(a)

  Fix $\lambda > 0$. If $\hat{\beta}_i \neq \hat{\beta}_j$, let us consider $\hat{\beta}^*$ as follows:

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k & \text{if } k \neq i \text{ and } k \neq j, \\ \frac{1}{2}(\hat{\beta}_i + \hat{\beta}_j) & \text{if } k = i \text{ and } k = j \end{cases}$$

  Because $\boldsymbol{x}_i = \boldsymbol{x}_j$, it is obvious that $\boldsymbol{X}\hat{\boldsymbol{\beta}}^* = \boldsymbol{X}\hat{\beta}$;thus $|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*|^2 = |\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}|^2$.However, $J(\cdot)$ is strictly convex, so we have $J(\hat{\boldsymbol{\beta}}^*) < J(\hat{\boldsymbol{\beta}})$. Therefore,$\hat{\boldsymbol{\beta}}$ cannot be the minimizer of equation (5), which is a contradiction. So we must have $\hat{\beta}_i = \hat{\beta}_j$.

Lemma 1 (Part b)

Statement: Assume $\boldsymbol{x}_i = \boldsymbol{x}_j$, $i, j \in \{1, \ldots, p\}$. If $J(\boldsymbol{\beta}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||_1$, then:

- $\hat{\beta}_i \cdot \hat{\beta}_j > 0$,
- $\hat{\boldsymbol{\beta}}^*$ is another minimizer of $J(\boldsymbol{\beta})$, where:

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot s, & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s), & \text{if } k = j, \end{cases}$$

for any $s \in [0, 1]$.

Proof Setup (Part b of Lemma 1)

- Assume $\hat{\beta}_i \cdot \hat{\beta}_j < 0$, i.e., $\hat{\beta}_i$ and $\hat{\beta}_j$ have opposite signs.
- Define $\hat{\boldsymbol{\beta}}^*$ as:

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot s, & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s), & \text{if } k = j. \end{cases}$$

- We aim to compare the objective values $J(\hat{\boldsymbol{\beta}})$ and $J(\hat{\boldsymbol{\beta}}^*)$.

Objective Function Comparison

1. **Residual Term:**
   - Since $\boldsymbol{x}_i = \boldsymbol{x}_j$, the contribution of $\hat{\beta}_i$ and $\hat{\beta}_j$ to the residual term is unchanged:

$$||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*||^2 = ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2.$$

2. **L_1-Norm Penalty Term:**
   - The L_1-norm for $\hat{\boldsymbol{\beta}}^*$ is:

$$||\hat{\boldsymbol{\beta}}^*||_1 = \sum_{k \neq i,j} |\hat{\beta}_k| + |\hat{\beta}_i + \hat{\beta}_j|.$$

   - Since $\hat{\beta}_i \cdot \hat{\beta}_j < 0$, we have:

$$|\hat{\beta}_i + \hat{\beta}_j| < |\hat{\beta}_i| + |\hat{\beta}_j|.$$

   - Hence:

$$||\hat{\boldsymbol{\beta}}^*||_1 < ||\hat{\boldsymbol{\beta}}||_1.$$

Conclusion of Proof (Part b)

- The total objective function for $\hat{\boldsymbol{\beta}}^*$ is:

$$J(\hat{\boldsymbol{\beta}}^*) = ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*||^2 + \lambda||\hat{\boldsymbol{\beta}}^*||_1.$$

- Since:

$$||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*||^2 = ||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2,$$

and:

$$||\hat{\boldsymbol{\beta}}^*||_1 < ||\hat{\boldsymbol{\beta}}||_1,$$

it follows that:

$$J(\hat{\boldsymbol{\beta}}^*) < J(\hat{\boldsymbol{\beta}}).$$

- This contradicts the assumption that $\hat{\boldsymbol{\beta}}$ minimizes $J(\boldsymbol{\beta})$.
- Therefore, $\hat{\beta}_i \cdot \hat{\beta}_j \geq 0$. Moreover, if $\hat{\beta}_i \cdot \hat{\beta}_j > 0$, the modified solution $\hat{\boldsymbol{\beta}}^*$ is also a minimizer.

Conclusion of Lemma 1

Lemma 1 Summary:

- Assume $\boldsymbol{x}_i = \boldsymbol{x}_j$, where $i, j \in \{1, \ldots, p\}$.
- The conclusions are:
  1. If $J(\cdot)$ is strictly convex, then:

$$\hat{\beta}_i = \hat{\beta}_j, \quad \forall \lambda > 0.$$

  2. If $J(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_1$, then:
     - $\hat{\beta}_i \cdot \hat{\beta}_j > 0$, meaning $\hat{\beta}_i$ and $\hat{\beta}_j$ have the same sign.
     - Any convex combination of $\hat{\beta}_i$ and $\hat{\beta}_j$, defined as:

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & \text{if } k \neq i \text{ and } k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot s, & \text{if } k = i, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s), & \text{if } k = j, \end{cases}$$

for $s \in [0, 1]$, is also a minimizer of $J(\boldsymbol{\beta})$.

Theorem

- Lemma shows a clear distinction between strictly convex penalty functions and the lasso penalty. Strict convexity guarantees the grouping effect in the extreme situation with identical predictors. In contrast the lasso does not even have a unique solution.
- *Theorem1* Given data $(\boldsymbol{y}, \boldsymbol{X})$ and parameters (λ_1, λ_2), the response $\boldsymbol{y}$ is centered and the predictors $\boldsymbol{X}$ are stendardized. Let $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ be the naıve elastic net estimate. Suppose that $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ defeine

$$D_{\lambda_1, \lambda_2}(i, j) = \frac{1}{|\boldsymbol{y}|_1}|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|$$

then

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2}\sqrt{\{2(1-\rho)\}}$$

, where $\rho = \boldsymbol{x}_i^T \boldsymbol{x}_j$, the sample correlation.

Proof of the Theorem1

If $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$, then both $\hat{\beta}_i(\lambda_1, \lambda_2)$ and $\hat{\beta}_j(\lambda_1, \lambda_2)$ are non-zero, and we have $sign\{\hat{\beta}_i(\lambda_1, \lambda_2)\} = sign\{\hat{\beta}_j(\lambda_1, \lambda_2)\}$.Because of equation (3), $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ satisfies

$$\frac{\partial L(\lambda_1, \lambda_2, \boldsymbol{\beta})}{\partial \beta_k}\bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)} = 0 \quad \text{if } \hat{\beta}_k(\lambda_1, \lambda_2) \neq 0.$$

Hence we have

$$-2\boldsymbol{x}_i^\top\{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)\} + \lambda_1 \operatorname{sgn}\{\hat{\beta}_i(\lambda_1, \lambda_2)\} + 2\lambda_2\hat{\beta}_i(\lambda_1, \lambda_2) = 0, \tag{6}$$

$$-2\boldsymbol{x}_j^\top\{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)\} + \lambda_1 \operatorname{sgn}\{\hat{\beta}_j(\lambda_1, \lambda_2)\} + 2\lambda_2\hat{\beta}_j(\lambda_1, \lambda_2) = 0. \tag{7}$$

Subtracting equation (6) from equation (7) gives

$$(\boldsymbol{x}_j^\top - \boldsymbol{x}_i^\top)\{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)\} + \lambda_2\{\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)\} = 0,$$

which is equivalent to

$$\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2) = \frac{1}{\lambda_2}(\boldsymbol{x}_i^\top - \boldsymbol{x}_j^\top)\hat{\mathrm{r}}(\lambda_1, \lambda_2), \tag{8}$$

where $\hat{\mathrm{r}}(\lambda_1, \lambda_2) = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ is the residual vector,

Proof of the Theorem1

Since $\boldsymbol{X}$ is standardized,

$$|\boldsymbol{x}_i - \boldsymbol{x}_j|^2 = 2(1-\rho), \quad \rho = \boldsymbol{x}_i^\top \boldsymbol{x}_j.$$

By equation (3), we must have

$$L\{\lambda_1, \lambda_2, \hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)\} \leq L(\lambda_1, \lambda_2, \boldsymbol{\beta} = 0),$$

i.e.,

$$|\hat{\mathrm{r}}(\lambda_1, \lambda_2)|^2 + \lambda_2 |\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)|^2 + \lambda_1 |\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)|_1 \leq |\boldsymbol{y}|^2.$$

So $|\hat{\mathrm{r}}(\lambda_1, \lambda_2)| \leq |\mathrm{y}|$. Then equation (8) implies that

$$D_{\lambda_1, \lambda_2}(i,j) \leq \frac{1}{\lambda_2} \frac{|\hat{\mathrm{r}}(\lambda_1, \lambda_2)|}{|\mathrm{y}|} |\boldsymbol{x}_i - \boldsymbol{x}_j| \leq \frac{1}{\lambda_2} \sqrt{2(1-\rho)}.$$

Conclusion of Theorem 1

Theorem 1 Summary:

- Given standardized predictors $\boldsymbol{X}$ and centered response $\boldsymbol{y}$, let $\hat{\beta}(\lambda_1, \lambda_2)$ be the naive elastic net estimate.
- Suppose $\hat{\beta}_i(\lambda_1, \lambda_2) \cdot \hat{\beta}_j(\lambda_1, \lambda_2) > 0$, for predictors i and j.
- Define:

$$D_{\lambda_1, \lambda_2}(i, j) = \frac{1}{||\boldsymbol{y}||_1} \cdot |\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|.$$

- Then:

$$D_{\lambda_1, \lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)},$$

where $\rho = \boldsymbol{x}_i^\top \boldsymbol{x}_j$ is the sample correlation between predictors i and j.

Conclusion:

- The difference between the coefficients of highly correlated predictors is bounded by $D_{\lambda_1, \lambda_2}(i, j)$.
- This demonstrates the **grouping effect** of the naive elastic net, ensuring similar coefficients for highly correlated predictors.

Why Lasso Does Not Exhibit the Grouping Effect

1. Independent Treatment of Predictors:
   The L_1 penalty treats each predictor independently, without considering correlations between them. As a result, lasso arbitrarily selects one predictor from a group of highly correlated variables and shrinks the rest to zero.
2. Geometric Explanation:
   - Lasso's constraint region is diamond-shaped, which encourages sparsity by forcing some coefficients to zero.
   - For highly correlated predictors, the optimal solution often lies on a vertex of the diamond, where only one coefficient is nonzero. This breaks the grouping effect.
3. Example from Tibshirani (1996):
   - Consider a simple linear model with two predictors ($p = 2$)

$$|\hat{\beta}_1 - \hat{\beta}_2| = |cos(\theta)|$$

     where, θ: Angle between $\boldsymbol{y}$ and the difference vector $\boldsymbol{x}_1 - \boldsymbol{x}_2$
   - As $\rho = corr(\boldsymbol{x}_1, \boldsymbol{x}_2) \to 1$, $cos(\theta)$ does not necessarily vanish. This means the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ may still differ significantly, even for highly correlated predictors.