Sparsity & Shrinkage: High-Dimensional Regression

전정빈

인하대학교 통계학과

March 23, 2026

이 노트는 여러 주제를 스스로 학습하는 과정에서 정리한 것이다. 주된 내용은 박성현 et al. [2023], Hastie et al. [2009]와 같은 훌륭한 저서들에 근거하고 있으며, 아울러 Lasso와 Elastic Net의 원전 논문을 비롯한 다양한 문헌의 정리와 증명 또한 함께 참고하여 구성하였다. Subgradient와 KKT 조건을 비롯한 내용들은 Ryan Tibshirani의 Convex Optimization 강의가 큰 도움이 되었다.

1 변수 선택 문제

특정 상황에서는 회귀분석에서 모든 변수를 사용하지 않고, 일부의 변수를 선택하여 모형을 구축하게 된다. 그러한 상황을 변수 선택 문제라고 일컫는다. 본 노트에서는, 절편의 표기를 생략하기 위해 모든 데이터가 표준화되어있음을 가정한다.

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = n \quad \forall\, j \in [p].$$

Penalized Regression에서 회귀계수 추정량은, 다음의 일반화된 오차제곱합을 최소로 하는 해로 결정된다.

$$\frac{1}{2n}\|Y - Xb\|^2 + J_\lambda(b) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} b_j x_{ij})^2 + \sum_{j=1}^{p} J_\lambda(b_j).$$

- 첫번째 항의 경우, 다양한 손실함수로 확장할 수 있다. 힌지(Hinge) 손실, 로지스틱 손실, 후버(Huber) 손실 등이 그 예가 될 것이다. 이러한 손실함수의 경우, 분석의 편의를 위해 일반적으로 *볼록성*을 지니도록 선택한다.
- 두번째 항의 경우, 일반적으로 Vector Space 위의 *Norm*이 되도록 결정한다. 대표적인 예시는 다음과 같다. Norm이 좋은 성질을 가지도록 선택하는 것은 중요한 문제이고, ℓ_1-norm의 경우 후술하겠지만 decomposable하다는 좋은 성질을 가지고 있다.
  - ℓ_0-norm : $J_\lambda(b) = \lambda \sum_{i=1}^{p} \mathbb{I}(b_j \neq 0)$

- ℓ_1-norm : $J_\lambda(b) = \lambda \sum_{i=1}^{p} |b_j|$
- ℓ_2-norm : $J_\lambda(b) = \lambda \sum_{i=1}^{p} b_j^2$

2 Best Subset Selection with ℓ_0-norm

ℓ_0-norm으로 penalty를 주는 회귀 모형을 Best Subset Selection이라고 한다.

$$\min_{b \in \mathbb{R}^p} \frac{1}{2n} \|Y - Xb\|^2 \quad \text{subject to} \quad \|b\|_0 \leq s.$$

위의 문제는 크기가 s인 회귀계수의 부분 집합 중, 오차제곱합을 최소화하는 조합을 찾는 문제와 동일하다. 이 경우에는 $\binom{p}{s}$개의 모형 적합이 필요하고, 곧 all possible regression을 수행하는 것이 된다. p가 증가함에 따라 계산량이 지수적으로 증가하므로 p가 큰 고차원 회귀 문제에서는 수행하기 어렵게 된다.

3 Lasso Regression

ℓ_0-norm에서 계산 문제를 해결하기 위해, ℓ_0-norm과 가장 가까우며 목적함수를 convex하게 하는 Lasso가 고안되었다.

$$\hat{\beta}^{\text{LASSO}} = \arg\min_{b \in \mathbb{R}^p} \left[\frac{1}{2n} \|Y - Xb\|^2 + \lambda \|b\|_1 \right]$$

위 문제를 제약조건 하의 최적화 문제로 고려하면 다음과 같이 표기할 수있다.

$$\min_{b \in \mathbb{R}^p} \frac{1}{2n} \|Y - Xb\|^2 \quad \text{subject to} \quad \|b\|_1 \leq t.$$

이때, t와 λ는 서로 일대일 대응 관계에 놓인다.

3.1 Lasso 추정량의 형태

주어진 λ에 대해서 Lasso 문제의 최적해를 살펴보기 위하여, X가 서로 직교하는 경우 ($X^\top X = nI_p$)에 대해 알아보자. 우선 $\lambda = 0$인 경우, 자명하게 $\hat{\beta}^{\text{LASSO}} = (X^\top X)^{-1} X'Y = X^\top Y/n$로 주어진다. $\lambda > 0$인 경우를 조사하기 위해 목적함수 $Q_\lambda(b)$를 아래와 같이 정의하자. 또한 기호의 편의를 위해 부호 벡터 $s(b_j) = 2\mathbb{I}(b_j > 0) - 1$, $s(b) = (s(b_1), \ldots, s(b_p))^\top$으로 정의하자.

$$Q_\lambda(b) = \frac{1}{2n} \|Y - Xb\|^2 + \lambda\, s(b)'b.$$

Penalty 항에 대해서는 다음의 관계에 의함을 알 수 있다.

$$\lambda \|b\|_1 = \lambda \sum_{j=1}^{p} |b_j| = \lambda \sum_{j=1}^{p} s(b_j) b_j = \lambda\, s(b)'b.$$

Lasso 추정량에서 유의한 회귀계수들의 인덱스 집합을 $\mathcal{S} = \{j : \hat{\beta}_j^{\text{LASSO}} \neq 0\}$라고 하자. 이때 목적함수 $Q_\lambda(b)$는 convex하기 때문에 Karush-Kuhn-Tucker (KKT) 조건의 Stationarity에 의해 최적해 $\hat{\beta}^{\text{LASSO}}$에서 다음이 성립한다.

$$0 \in \partial Q_\lambda(\hat{\beta}^{\text{LASSO}}).$$

목적함수의 subgradient는 다음과 같이 주어진다.

$$\partial Q_\lambda(\hat{\beta}^{\text{LASSO}}) = \left\{ -\frac{1}{n} x_j^\top (Y - X\hat{\beta}^{\text{LASSO}}) + \lambda z \,\middle|\, z \in \mathbb{R}^p, z_k \in \partial |\hat{\beta}^{\text{LASSO}}| \right\}$$

따라서 다음의 두 경우로 해를 나누어 생각할 수 있을 것이다.

1. $\hat{\beta}_j^{\text{LASSO}} \neq 0$인 경우, 직교성 가정에 의해서 $\hat{\beta}^{\text{LSE}} = (X^\top X)^{-1} X^\top Y = X^\top Y/n$의 사실과 $\partial\, |\hat{\beta}_j^{\text{LASSO}}| = s(\hat{\beta}_j^{\text{LASSO}})$로 주어진다는 사실을 통해 다음을 확인할 수 있다.

$$\begin{aligned} & -\frac{1}{n} x_j^\top (Y - X\hat{\beta}^{\text{LASSO}}) + \lambda\, s(\hat{\beta}_j^{\text{LASSO}}) = 0 \\ & \Rightarrow \hat{\beta}^{\text{LASSO}} = \hat{\beta}^{\text{LSE}} - \lambda\, s(\hat{\beta}^{\text{LASSO}}) \end{aligned}$$

2. $\hat{\beta}_j^{\text{LASSO}} = 0$인 경우, $\partial |\hat{\beta}_j^{\text{LASSO}}| = [-1, 1]$로 주어진다는 사실을 이용하면 다음의 사실을 알 수 있다.

$$\begin{aligned} & -\frac{1}{n} x_j^\top (Y - X\hat{\beta}^{\text{LASSO}}) + \lambda z = 0 \quad \forall\, z \in [-1, 1] \\ & \Rightarrow |-\frac{1}{n} x_j^\top (Y - X\hat{\beta}^{\text{LASSO}})| \leq \lambda \\ & \Rightarrow |-\hat{\beta}_j^{\text{LSE}} + \hat{\beta}_j^{\text{LASSO}}| \leq \lambda \\ & \Rightarrow |\hat{\beta}_j^{\text{LSE}}| \leq \lambda \end{aligned}$$

그러므로, 위의 두가지 사실을 종합하면 Lasso의 최적해가 설명변수가 직교행렬인 경우에 다음과 같이 주어짐을 알 수 있다.

$$\hat{\beta}_j^{\text{LASSO}} = \text{sign}(\hat{\beta}_j^{\text{LSE}}) \max(|\hat{\beta}_j^{\text{LSE}}| - \lambda, 0)$$

3.2 Lasso 추정량의 기하학적인 이해

다음은 Lasso 문제의 유용한 변환에 대해서 소개한다. 이 변환을 통해서 문제를 직교성 가정 아래에서 더 직관적으로 이해할 수 있을 것이다. 사실 아까의 추정량을 구하는 과정도, 아래와 같이 변환하여 풀었으면 더 쉽게 풀 수 있었다. 우선 오차제곱합에 대해서 간단한 식 조작을 가해볼 것이다.

$$\begin{aligned} \|Y - Xb\|^2 &= (Y - Xb)^\top (Y - Xb) \\ &= (Y - X\hat{\beta}^{\text{LSE}} + X\hat{\beta}^{\text{LSE}} - Xb)^\top (Y - X\hat{\beta}^{\text{LSE}} + X\hat{\beta}^{\text{LSE}} - Xb) \\ &= (Y - X\hat{\beta}^{\text{LSE}})^\top (Y - X\hat{\beta}^{\text{LSE}}) + (X\hat{\beta}^{\text{LSE}} - Xb)^\top (X\hat{\beta}^{\text{LSE}} - Xb). \end{aligned}$$

교차항은 아래의 사실에 의해 사라진다.

$$(Y - \hat{\beta}^{\mathrm{LSE}})^\top (X\hat{\beta}^{\mathrm{LSE}} - Xb) = Y^\top (I - H) X(\hat{\beta}^{\mathrm{LSE}} - b) = 0.$$

따라서, Lasso 문제는 아래와 같이 다시 표현될 수 있다.

$$\min_{b \in \mathbb{R}^p} (b - \hat{\beta}^{\mathrm{LSE}})^\top (X^\top X)(b - \hat{\beta}^{\mathrm{LSE}}) \quad \text{subject to} \quad \|b\|_1 \leq t.$$

이는 $\hat{\beta}^{\mathrm{LSE}}$를 중심으로하는 타원으로 이해할 수 있다. $\|b\|_1 = 0$에서 시작해서 $\hat{\beta}^{\mathrm{LSE}}$와 가까워지는 방향으로 나아가려 하지만, 마름모의 꼭짓점에서 걸리게 된다. 이 관점을 그림으로 표현하면 아래와 같다.

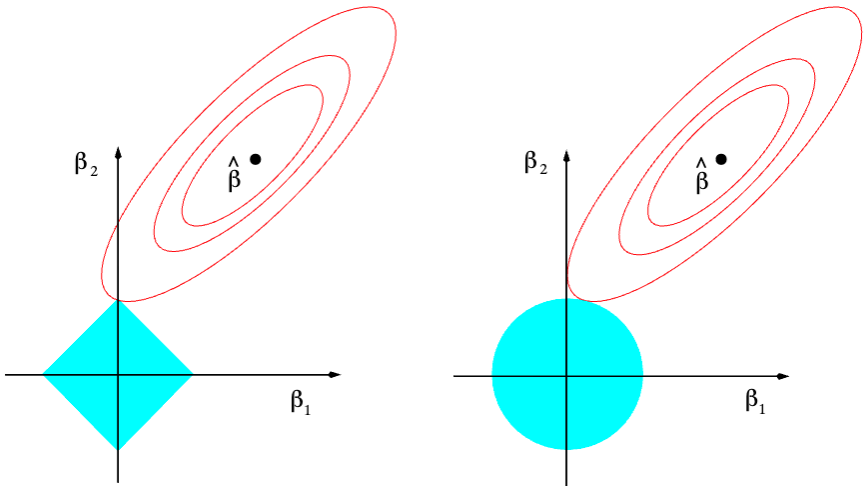

Figure 1: 왼쪽은 Lasso, 오른쪽은 Ridge 문제를 시각화한 것이다.

Remark 1. Lasso 추정량은 일반적으로 LARS 알고리즘을 통해 계산하는 것으로 알려져있다. LARS 알고리즘은 $p > n$인 고차원 상황에서도 적용이 가능하며, λ의 변화에 따른 Solution Path를 계산해주는 것으로 알려져 있다. 그 이론적 성질은 몇가지 적절한 가정 아래에서 추정된 인덱스 집합 $\mathcal{S}$가 True 인덱스 집합으로 확률수렴함이 알려져 있다. 이론이 꽤나 복잡하기 때문에 자세한 설명은 JRSS-B의 원문 등을 확인하는 것을 추천한다.

Remark 2. 앞서 언급한대로 Lasso 추정을 위해 사용되는 ℓ_1-norm은 decomposable하다는 좋은 성질이 있다. $j \in \mathcal{S}$에 대해서 $b_j \neq 0$이고, $l \notin \mathcal{S}$에 대해서 $b_l = 0$이라고 하자. 그러면 다음이 성립한다.

$$\|b\|_1 = \sum_{j \in \mathcal{S}} |b_j| + \sum_{l \in \mathcal{S}^c} |b_l| = \|b_{\mathcal{S}}\|_1 + \|b_{\mathcal{S}^\perp}\|_1.$$

위 성질은 Lasso 추정량에서 signal과 noise를 분리해서 해석하는 것을 용이하게 해주고, 관련된 증명에서도 요긴하게 사용이 된다.

3.3 Lasso 추정량의 확장

문제의 설정에 따라서 Penalty Term을 교체할 수 있을 것이다. 다음의 두 가지 확장을 소개한다.

Adaptive Lasso Lasso의 Bias를 줄이기 위하여 고안된 추정량으로 가중치 $w_j = 1/|\hat{\beta}^{\mathrm{LSE}}|^\nu, \nu > 0$에 대하여 가중된 ℓ_1-norm을 사용한다.

$$\hat{\beta}^{\mathrm{AL}} = \underset{b\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2n}\|Y - Xb\|^2 + \lambda \sum_{j=1}^{p} w_j |b_j| \right]$$

해는 직교성 가정($X^\top X = nI_p$) 아래에서 다음과 같이 주어진다. Lasso에서 가중치만 바뀐 것이기 때문에 Lasso와 매우 유사하게 나온다.

$$\hat{\beta}_j^{\mathrm{AL}} = \mathrm{sign}(\hat{\beta}_j^{\mathrm{LSE}}) \max(|\hat{\beta}_j^{\mathrm{LSE}}| - \lambda w_j, 0)$$

앞서 보였던 것과 같이 목적식은 직교성 가정을 이용하면 아래와 같이 표현할 수 있다.

$$\begin{aligned}\|Y - Xb\| &= \|Y - X\hat{\beta}^{\mathrm{LSE}}\|^2 + \|Xb - X\hat{\beta}^{\mathrm{LSE}}\|^2 \\ &= n\|b - \hat{\beta}^{\mathrm{LSE}}\|^2 + \|Y - X\hat{\beta}^{\mathrm{LSE}}\|^2\end{aligned}$$

따라서,

$$Q_\lambda(b) = \frac{1}{2}\|b - \hat{\beta}^{\mathrm{LSE}}\|^2 + \lambda \sum_{j=1}^{p} w_j |b_j|.$$

위 식의 j-th component 별 subdifferential을 구하면,

$$\partial_{b_j} Q_\lambda(b) = \left\{ (b_j - \hat{\beta}_j^{\mathrm{LSE}}) + \lambda w_j z \;\middle|\; z_j \in \partial |z_j| \right\}$$

따라서, 앞서 보인것과 마찬가지로 KKT condition의 stationary 조건에 의해 subdifferential이 0을 원소로 가져야함을 이용하면 최적해를 어렵지 않게 보일 수 있다.

Elastic Net Lasso는 상관관계가 있는 변수 중 흔히 하나만 선택하는 성질을 가지고 있다. 간단한 예시로 $p = 2$일때, $x_1 = x_2$인 경우 $Y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$의 회귀모형을 따른다고 할때, Lasso 추정량은 $\hat{\beta}_1^{\mathrm{LASSO}} = 0$ 또는 $\hat{\beta}_2^{\mathrm{LASSO}} = 0$이 된다. 반면에, Elastic Net은 ℓ_1, ℓ_2-norm의 convex 조합을 사용하여, 상관관계가 있는 유의미한 변수를 모두 선택한다. (Ridge의 성질을 계승하는 듯 하다.)

$$\hat{\beta}^{\mathrm{EL}} = \underset{b\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2n}\|Y - Xb\|^2 + \lambda \sum_{j=1}^{p} ((1-\alpha)|b_j| + \alpha\, b_j^2) \right]$$

해는 직교성 가정 아래에서 다음과 같이 주어진다.

$$\hat{\beta}_j^{\mathrm{EL}} = \frac{1}{1+2\lambda\alpha}\mathrm{sign}(\hat{\beta}_j^{\mathrm{LSE}})\max(|\hat{\beta}_j^{\mathrm{LSE}}| - \lambda(1-\alpha), 0).$$

마찬가지로 목적식을 변형하여 j-th component의 subdifferential을 구하면 다음과 같이 주어진다.

$$\partial_{b_j} Q_\lambda(b) = \left\{ (b_j - \hat{\beta}_j^{\mathrm{LSE}}) + \lambda(1-\alpha)z + 2\lambda\alpha b_j \,\middle|\, z \in \partial|b_j| \right\}$$

따라서, $b_j \neq 0$인 경우, $b_j = \hat{\beta}_j^{\mathrm{LSE}} - \lambda(1-\alpha)\mathrm{sign}(b_j) - 2\lambda\alpha b_j$로 주어지고, $b_j = 0$인 경우, $|\hat{\beta}_j^{\mathrm{LSE}}| \leq \lambda(1-\alpha)$로 주어지게 되어 어렵지 않게 위의 최적해를 구할 수 있다. 따라서 해의 형태는 Lasso와 같이 soft threshold를 하며, 동시에 Ridge와 같이 shirinkage를 하게 됨을 알 수 있다.

다음은 Elastic Net 논문 원전 [Zou and Hastie, 2005]의 이론적 결과를 소개한다. 이를 직접 증명해보고 공부해보는 것은 도움이 될 듯하여 기록을 남긴다.

Lemma 1 (Lemma 1 of [Zou and Hastie, 2005])**.** *$\lambda_1 = 2n\lambda(1-\alpha)$, $\lambda_2 = 2n\lambda\,\alpha$라고 하자. 자료 (Y, X)가 주어졌을때, 증강된 자료 (Y^*, X^*)를 아래와 같이 정의하자.*

$$X^*_{(n+p)\times p} = \frac{1}{\sqrt{n+\lambda_2}}\begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix}, \qquad Y^*_{n+p} = \begin{pmatrix} Y \\ 0 \end{pmatrix}.$$

그러면 $\gamma = \lambda_1/\sqrt{1+\lambda_2}$에 대해 최적해 $\hat{\beta}^{\mathrm{EL}}$은 증강된 자료에서의 Lasso 문제로 주어진다.

$$\hat{\beta}^\star = \underset{b\in\mathbb{R}^p}{\arg\min}\left[\|Y^* - X^* b\|^2 + \gamma|b\|_1\right], \qquad \hat{\beta}^{\mathrm{EL}} = \frac{1}{\sqrt{n+\lambda_2}}\hat{\beta}^\star.$$

Proof. 우선, 목적식을 기술하는 것에서부터 시작하자.

$$\begin{aligned}\|Y - Xb\|^2 + \lambda_1\|b\|_1 + \lambda_2\|b\|^2 &= Y^\top Y - 2Y^\top Xb + b^\top X^\top Xb + \lambda_2 b^\top b + \lambda_1\|b\|_1 \\ &= Y^\top Y - 2Y^\top Xb + b^\top(X^\top X + \lambda_2 I_p)b + \lambda_1\|b\|_1.\end{aligned}$$

이제 $(n+p)\times p$ 크기의 증강행렬 $\tilde{X}$를 고려하자:

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda_2} I_p. \end{pmatrix}$$

그러면 어렵지 않게, $\tilde{X}^\top\tilde{X} = X^\top X + \lambda_2 I_p$임을 확인할 수 있다. 한편 $\tilde{X}$는 표준화되어있지않다. 우리가 맨 앞에서 만든 가정을 만족하지 않는데,

$$\tilde{x_j}^\top\tilde{x_j} = x_j^\top x_j + \lambda_2 = n + \lambda_2.$$

따라서, 이제 우리는 표준화된 버전의 $\tilde{X}$를 고려한다:

$$X^* = \frac{1}{\sqrt{n+\lambda_2}} \begin{pmatrix} X \\ \sqrt{\lambda_2} I_p. \end{pmatrix}$$

위의 정의들을 이용해서 다시 목적식을 표현해보면,

$$\begin{aligned}
&\|Y - Xb\|^2 + \lambda_1 \|b\|_1 + \lambda_2 \|b\|^2 \\
&= Y^\top Y - 2Y^\top Xb + b^\top (X^\top X + \lambda_2 I_p) b + \lambda_1 \|b\|_1 \\
&= Y^{*\top} Y^* - 2\sqrt{n+\lambda_2}\, Y^{*\top} X^* b + (n+\lambda_2) b^\top (X^{*\top} X^*) b + \frac{\sqrt{n+\lambda_2}}{\sqrt{n+\lambda_2}} \lambda_1 \|b\|_1 \\
&= Y^{*\top} Y^* + 2Y^{*\top} X (\sqrt{n+\lambda_2}\, b) + (\sqrt{n+\lambda_2}\, b)^\top X^{*\top} X^* (\sqrt{n+\lambda_2}\, b) + \frac{\lambda_1}{\sqrt{n+\lambda_2}} \|\sqrt{n+\lambda_2} b\|_1
\end{aligned}$$

이제, $b^* = \sqrt{n+\lambda_2} b$, $\gamma = \lambda_1 / \sqrt{n+\lambda_2}$로 정의하면, 다음의 표현을 얻을 수 있다.

$$\begin{aligned}
&Y^{*\top} Y^* - 2Y^{*\top} X b^* + b^{*\top} X^{*\top} X^* b^* + \gamma \|b^*\|_1 \\
&= \|Y^* - X^* b^*\|^2 + \gamma \|b^*\|_1.
\end{aligned}$$

위 문제는 Lasso 문제로 귀결되므로, 이로써 증명이 완성된다. □

위 Lemma가 전해주는 사실은, Elastic net의 solution은 증강된 자료에서의 Lasso의 solution과 같다는 것이다. 사실 위 Lemma를 통해서 자연스럽게 증강된 자료에서의 최소제곱추정량이 원 자료에서의 Ridge Estimator와 정확히 일치함도 보일 수 있다. 여러모로 유용한 Lemma이다.

Theorem 2 (Elastic Net에서 추정계수와 상관성의 관계, Theorem 1 of [Zou and Hastie, 2005])**.** *공변량이 x_1, x_2 뿐인 회귀모형에서, 두 공변량 $x_1^\top x_1 = x_2^\top x_2 = n$, $x_1^\top x_2 = n r_{12}$의 관계를 가진다고 하자. 또한 $M := \max_i |Y_i|$라고 하자. 그러면 다음이 성립한다.*

$$|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{M}{2\lambda\alpha} \sqrt{2(1-r_{12})}.$$

Proof. $\hat{\beta}_1^{\mathrm{EL}} \cdot \hat{\beta}_2^{\mathrm{EL}} > 0$이라고 하자. 그러면 $\mathrm{sign}(\hat{\beta}_1^{\mathrm{EL}}) = \mathrm{sign}(\hat{\beta}_2^{\mathrm{EL}})$이므로 목적함수 최적해의 subgradient 조건에 의해 다음이 성립한다.

$$-\frac{1}{n} x_j^\top (Y - X\hat{\beta}^{\mathrm{EL}}) + \lambda(1-\alpha)\,\mathrm{sign}(\hat{\beta}_j^{\mathrm{EL}}) + 2\lambda\alpha\hat{\beta}_j^{\mathrm{EL}} = 0, \qquad j = 1, 2.$$

이때, $j = 1, 2$일 때의 두 식을 서로 빼주면 부호항이 소거되어 다음을 얻을 수 있다.

$$\begin{aligned}
&-\frac{1}{n}(x_1 - x_2)^\top (Y - X\hat{\beta}^{\mathrm{EL}}) + 2\lambda\alpha(\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}) = 0 \\
&\qquad\qquad \Rightarrow \lambda\alpha(\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}) = \frac{1}{2n}(x_1 - x_2)^\top (Y - X\hat{\beta}^{\mathrm{EL}}).
\end{aligned}$$

삼각부등식에 의해 다음이 성립한다.

$$\lambda\alpha|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{1}{2n}|(x_1 - x_2)^\top Y| + \frac{1}{2n}|(x_1 - x_2)^\top X\hat{\beta}^{\mathrm{EL}}|.$$

한편, 아까의 가정에 의해 $\hat{\beta}_1^{\mathrm{EL}}, \hat{\beta}_2^{\mathrm{EL}}$가 같은 부호이므로,

$$\begin{aligned}(x_1 - x_2)^\top X\hat{\beta}^{\mathrm{EL}} &= (x_1 - x_2)^\top(x_1\hat{\beta}_1^{\mathrm{EL}} + x_2\hat{\beta}_2^{\mathrm{EL}}) \\ &= (x_1^\top x_1 - x_2^\top x_1)\hat{\beta}_1^{\mathrm{EL}} + (x_1^\top x_2 - x_2^\top x_2)\hat{\beta}_2^{\mathrm{EL}} \\ &= n(1 - r_{12})\hat{\beta}_1^{\mathrm{EL}} - n(1 - r_{12})\hat{\beta}_2^{\mathrm{EL}} \\ &= n(1 - r_{12})(\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}).\end{aligned}$$

그러므로 이를 다시 대입하여 식을 정리하면,

$$\left(\lambda\alpha + \frac{1 - r_{12}}{2}\right)|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{1}{2n}|(x_1 - x_2)^\top Y|.$$

특히, 더 간단화하면,

$$\lambda\alpha|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{1}{2n}|(x_1 - x_2)^\top Y|.$$

이제 Cauchy–Schwarz 부등식과 $|Y_i| \le M$의 조건을 사용하면,

$$\begin{aligned}|(x_1 - x_2)^\top Y| &\le \|x_1 - x_2\|\,\|Y\| \\ &\le \sqrt{\|x_1\|^2 + \|x_2\|^2 - 2x_1^\top x_2}\,\sqrt{n}M \\ &= \sqrt{2n(1 - r_{12})}\,\sqrt{n}M \\ &= nM\sqrt{2(1 - r_{12})}.\end{aligned}$$

따라서

$$\lambda\alpha|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{M}{2}\sqrt{2(1 - r_{12})},$$

즉

$$|\hat{\beta}_1^{\mathrm{EL}} - \hat{\beta}_2^{\mathrm{EL}}| \le \frac{M}{2\lambda\alpha}\sqrt{2(1 - r_{12})}.$$

□

위의 Theorem에 의해 몇 가지 재미있는 사실을 알 수 있다. Ridge를 포함하여 $\alpha > 0$이고 $r_{12} = 1$인 경우에 $\hat{\beta}_1^{\mathrm{EL}} = \hat{\beta}_2^{\mathrm{EL}}$가 된다. 한편 $\alpha = 1$인 Lasso의 경우에는 Bound가 매우 커져, 두 회귀계수가 동일하지 않을 수 있게 된다. 이는 곧 한쪽 변수가 0으로 갈 수 있음을 시사한다. 더 정확히는 $x_{i1} = x_{i2}(:= x_i)$ 인

경우에 Lasso는 다음의 문제를 푸는 것과 같다.

$$\begin{aligned}
&\min_{b_1,b_2\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i b_1 - x_i b_2)^2 \quad \text{subject to} \quad |b_1| + |b_2| \le t \\
&\iff \min_{b_1,b_2\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i (b_1 + b_2))^2 \quad \text{subject to} \quad |b_1 + b_2| \le |b_1| + |b_2| \le t \\
&\iff \min_{\theta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - x_i \theta)^2 \quad \text{subject to} \quad |\theta| \le t
\end{aligned}$$

따라서, 최적해는 단일 변수의 Lasso 해와 같이 나오게 된다.

$$\hat{\theta}^{\text{LASSO}} := \hat{\beta}_1^{\text{LASSO}} + \hat{\beta}_2^{\text{LASSO}} = \text{sign}\left(\frac{1}{n}X^\top Y\right)\max\left(\left|\frac{1}{n}X^\top Y\right| - \lambda, 0\right)$$

두 계수의 합은 결정했지만, 어떻게 분배할지는 결정되지 않아 해가 무수히 많은 상태가 된다.

4 다중공선성의 문제

이번에는 선형회귀모형의 공변량이 다음과 같이 표준화되어있다고 가정하자.

$$\sum_{i=1}^{n} x_{ij} = 0, \quad \sum_{i=1}^{n} x_{ij}^2 = 1 \quad \forall\, j \in [p].$$

행렬 $X = (x_{[j]}, X_{-j})$에 대해서 알아보자. Woodbury의 block matrix inversion formula에 의해 다음을 알고 있다.

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ -A_{22}^{-1}A_{21} & I \end{bmatrix} \begin{bmatrix} A_{1|2}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{bmatrix} \begin{bmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{1|2}^{-1} & \star \\ \star & \star \end{bmatrix}$$

따라서, 위의 사실을 이용하면 $(X^\top X)^{-1}$의 (j,j)번째 원소가 다음과 같이 주어짐을 알 수 있다.

$$c_{jj} = (x_{[j]}^\top x_{[j]} - x_{[j]}^\top X_{-j}(X_{-j}^\top X_{-j})^{-1}X_{-j}^\top x_{[j]})^{-1}$$

그러면, $x[j]$를 반응변수로 하고 X_{-j}를 설명변수로 하는 절편이 없는 선형회귀모형 $x_{[j]} = X_{-j}\gamma$을 적합했을때, 결정계수 R_j^2는 아래와 같이 주어진다.

$$R_j^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\|X_{-j}\,\hat{\gamma}\|^2}{\|x_{[j]}\|^2} = \frac{x_{[j]}^\top X_{-j}(X_{-j}^\top X_{-j})^{-1}X_{-j}^\top x_{[j]}}{x_{[j]}^\top x_{[j]}} = x_{[j]}^\top X_{-j}(X_{-j}^\top X_{-j})^{-1}X_{-j}^\top x_{[j]}.$$

$x_{[j]}$와 X_{-j}의 각 열들이 서로 직교한다면 $R_j^2 = 0$일 것이고, $x_{[j]} \in \text{col}(X_{-j})$이면 $x_{[j]}^\top X_{-j}(X_{-j}^\top X_{-j})^{-1}X_{-j}^\top x_{[j]} = x_{[j]}^\top x_{[j]} = 1$이 되어 $R_j^2 = 1$일 것이다. (Hat matrix의 성질 혹은 정사영에 정의에 따라서 성립한다.)

한편, $c_{jj} = (1 - R_j^2)^{-1}$의 관계가 있고, $R_j^2 \approx 1$이면 c_{jj}가 기하급수적으로 커지게 되어 분산추정에 어려움을 겪을 것이다. 이때, c_{jj}를 VIF_j라고 하고, VIF_j가 큰 상황을 다중공선성이 있다고 한다.
이러한 다중공선성의 문제를 해결하기 위해 Ridge, Principal Component Regression 등의 방법이 사용된다.

5 Ridge Regression

Penalty를 ℓ_2-norm으로 하는 penalized regression을 Ridge라고 한다. 본 절에서는 고유값을 표기할때 λ의 기호를 사용하기 위해서, 혼동을 막기 위해 조절모수를 표기할때 k를 사용하도록 한다.

$$\min_{b \in \mathbb{R}^p} \frac{1}{2}\|Y - Xb\|^2 \quad \text{subject to} \quad \|b\|^2 \leq t.$$

라그랑주 승수법을 이용하여 목적식을 설계하여 다시 나타내면 다음과 같다.

$$\hat{\beta}_k = \arg\min_{b \in \mathbb{R}^p} \left[\frac{1}{2}\|Y - Xb\|^2 + \frac{1}{2}k\|b\|^2 \right]$$

위 목적식을 최소화하는 최적해는 아래와 같이 주어진다.

$$\hat{\beta}_k = (X^\top X + kI)^{-1} X^\top Y.$$

최소제곱추정량 $\hat{\beta}^{\mathrm{LSE}} = (X^\top X)^{-1} X^\top Y$의 MSE를 조사해보자. λ_j를 $X^\top X$의 고유값이라고 하면,

$$\mathsf{MSE}(\hat{\beta}^{\mathrm{LSE}}) = \|\mathsf{bias}(\hat{\beta}^{\mathrm{LSE}})\|^2 + \mathrm{Tr}(\mathsf{Var}(\hat{\beta}^{\mathrm{LSE}})) = \mathrm{Tr}(\sigma^2 (X^\top X)^{-1}) = \sigma^2 \sum_{i=1}^{p} \lambda_j^{-1}.$$

따라서 설명변수 간에 완전에 가까운 다중공선성이 있다면, $\lambda_j \approx 0$인 값이 존재하게 되고 이는 곧 MSE를 크게 만드므로 MSE 관점에서 좋은 추정량이라고 할 수 없다. 이제, Ridge 추정량의 MSE를 조사해보자. 이를 위해, 먼저 Bias와 Variance를 구해보자.

$$\begin{aligned}
\mathsf{bias}(\hat{\beta}_k) = \mathbb{E}(\hat{\beta}_k) - \beta &= (X^\top X + kI)^{-1} X^\top X \beta - \beta \\
&= (X^\top X + kI)^{-1} (X^\top X + kI - kI)\beta - \beta \\
&= -k(X^\top X + kI)^{-1} \beta
\end{aligned}$$

$$\begin{aligned}
\mathrm{Tr}(\mathsf{Var}(\hat{\beta}_k)) &= \sigma^2 \mathrm{Tr}((X^\top X + kI)^{-1} X^\top X (X^\top X + kI)^{-1}) \\
&= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}
\end{aligned}$$

위의 두 결과를 종합하면, $\alpha = P\beta$ for some orthogonal matrix $P^\top P = PP^\top = I$에 대해 아래와 같이 나타낼 수 있다.

$$\mathsf{MSE}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p} \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}$$

5.1 고유값 분해를 통한 Ridge 추정량의 이해

고유값 분해를 통해 Ridge 추정량을 표기해보도록 하자. $X^\top X = \Gamma \Lambda \Gamma^\top$을 $X^\top X$의 고유값 분해라고 하자. 이때, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $\Gamma = (\gamma_1, \ldots, \gamma_p)$이다. 고유값 분해의 성질에 의해 $\Gamma^\top \Gamma = \Gamma \Gamma^\top = I$가 성립한다. 또한,

$$(X^\top X - \lambda_j I_p)\gamma_j = 0, \quad \forall j \in [p].$$

선형모형 $Y = X\beta + \epsilon$을 고유값 분해를 이용해 다시 표현하면,

$$Y = X\beta + \epsilon = X\Gamma\Gamma^\top \beta + \epsilon = Z\alpha + \epsilon.$$

위 모형에서 $Z^\top Z = \Gamma^\top X^\top X \Gamma = \Gamma^\top \Gamma \Lambda \Gamma^\top \Gamma = \Lambda$이므로, α의 최소제곱추정량은 아래와 같이 주어진다.

$$\hat{\alpha} = (Z^\top Z)^{-1} Z^\top Y = \Lambda^{-1} Z^\top Y.$$

한편, Ridge 추정량은 $\hat{\alpha}_k = (Z^\top Z + kI)^{-1} Z^\top Y = (\Lambda + kI)^{-1} Z^\top Y = (\Lambda + kI)^{-1} \lambda \hat{\alpha}$로 주어진다. 즉,

$$\hat{\alpha}_{k,j} = \frac{\lambda_j}{\lambda_j + k} \hat{\alpha}_j, \quad \forall j \in [p]$$

이므로, Ridge 추정량은 최소제곱추정량보다 작은값을 가짐을 알 수 있다.
한편 MSE를 계산하면 기존의 Ridge 추정량의 MSE와 동일하게 주어짐을 알 수 있다.

$$\mathsf{MSE}(\hat{\alpha}_k) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^{p} \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}$$

Theorem 3. *최소제곱추정량 $\hat{\alpha}$의 MSE보다 Ridge 추정량 $\hat{\alpha}_k$의 MSE가 더 작게 되는 $k > 0$이 존재한다.*

$$\exists k > 0 \quad s.t. \quad \mathsf{MSE}(\hat{\alpha}_k) \leq \mathsf{MSE}(\hat{\alpha}).$$

Proof. 우선, Ridge 추정량의 MSE를 k로 미분해보자.

$$\frac{\partial \mathsf{MSE}(\hat{\alpha}_k)}{\partial k} = -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2} - 2k^2 \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3}.$$

$k = 0$일때, $\mathsf{MSE}(\hat{\alpha}_k) = \mathsf{MSE}(\hat{\alpha})$이고, $\mathsf{MSE}(\hat{\alpha}_k)$는 k에 대한 연속함수이다. 또한,

$$\lim_{k \to 0+} \frac{\partial \mathsf{MSE}(\hat{\alpha}_k)}{\partial k} = -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} < 0$$

이므로, $k = 0$ 근방에서 k에 대한 감소함수이고, $\mathsf{MSE}(\hat{\alpha}_k)$가 $\mathsf{MSE}(\hat{\alpha})$보다 작게 되는 k의 값이 존재한다. □

구체적으로 MSE를 작게하는 k는 직교성 가정 $X^\top X = I$ 아래에서 다음과 같이 구할 수 있다. 직교성 가정 아래에서는 고유값 $\lambda_j = 1$로 주어지므로, MSE를 미분한 식이 다음과 같이 주어진다.

$$\begin{aligned}
\frac{\partial \mathsf{MSE}(\hat{\alpha}_k)}{\partial k} &= -2\sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^3} + 2k \sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j + k)^2} - 2k^2 \sum_{j=1}^{p} \frac{\lambda_j \alpha_j^2}{(\lambda_j + k)^3} \\
&= -2\sigma^2 \frac{p}{(1+k)^3} + +2k \frac{\alpha^\top \alpha}{(1+k)^2} - 2k^2 \frac{\alpha^\top \alpha}{(1+k)^3}
\end{aligned}$$

양변에 $(1+k)^3$을 곱하고 식을 정리하면,

$$-2p\sigma^2 + 2k\alpha^\top \alpha(1+k) - 2k^2 \alpha^\top \alpha = -2p\sigma^2 + 2k\alpha^\top \alpha.$$

따라서,

$$k = \frac{p\sigma^2}{\alpha^\top \alpha}$$

에서 MSE가 최소가 된다. 당연히 k를 각 변수마다 모두 다르게 할 수 있는 확장을 자연스럽게 고려할 수 있고, $K = (k_1, \ldots, k_p)^\top$으로 조절모수 벡터를 고려할 수 있을 것이다. 이 경우에도 직교성 가정 아래에 최적의 k가 아래와 같이 주어진다.

$$k_j = \frac{\sigma^2}{\alpha_j}.$$

5.2 특이값 분해를 통한 Ridge 추정량의 이해

특이값 분해를 통해 Ridge 추정량을 이해할 수도 있다. 먼저 이를 이해하기 위해서는 특이값 분해에 대해 이해해야한다.

특이값 분해 $\mathsf{rank}(X) = r$인 행렬 $X \in \mathbb{R}^{n \times p}$를 고려해보자. 행렬 X의 특이값 분해는 다음과 같다.

$$\begin{aligned}
X_{n \times p} &= U_{n \times n} D_{n \times p} P^\top_{p \times p} \\
&= \sum_{j=1}^{r} d_j u_r p_r^\top \\
&= U_{n \times r} D_{r \times r} P^\top_{r \times p}.
\end{aligned}$$

특이값 분해을 통해 주어진 두 U, P는 직교이다. $U^\top U = UU^\top = I_n$, $P^\top P = PP^\top = I_p$. 또한 $n > p$ 인 경우에는 $D_{n\times p} = \begin{bmatrix} D_{p\times p} \\ O_{(n-p)\times p} \end{bmatrix}$로 주어지고 $p > n$인 고차원의 경우에는 $D_{n\times p} = [D_{n\times n} \quad O_{n\times(n-p)}]$ 으로 주어진다.
또한 특이값 분해에는 Four fundamental subspace와 관련하여 아래의 기하학적인 성질이 있다.

- 임의의 벡터 ξ에 대해, $X\xi = U_r D_r P_r^\top \xi \in \mathsf{span}(U_r)$이므로, $\mathsf{col}(X) \subseteq \mathsf{span}(U_r)$. 한편, U_r의 각 열이 서로 직교하고 그 개수가 r개 이므로, $\mathsf{rank}(U_r) = r = \mathsf{rank}(X)$이다. 따라서 $\mathsf{col}(X) = \mathsf{span}(U_r)$ 이고, U_r의 각 열은 $\mathsf{col}(X)$의 basis이다.
- $X^\top = P_r D_r U_r^\top$이고, 비슷한 논증을 통해 $\mathsf{col}(X^\top) = \mathsf{span}(P_r)$을 얻을 수 있다.

Moore-Penrose 의사역행렬 다음의 네가지 조건을 만족시키는 행렬 $A^\dagger$을 Moore-Penrose 의사역행렬이라고 한다.

1. $AA^\dagger A = A$
2. $A^\dagger AA^\dagger = A^\dagger$
3. $(A^\dagger A)^\top = A^\dagger A$
4. $(AA^\dagger)^\top = AA^\dagger$

위 네가지 조건을 만족하는 행렬 $A^\dagger$는 아래와 같이 행렬 A의 특이값 분해에 의해 유일하게 결정된다.

$$A^\dagger = UD^\dagger P^\top = U \begin{bmatrix} D^{-1} & O \end{bmatrix} P^\top \quad \text{or} \quad \begin{bmatrix} D^{-1}1 \\ O \end{bmatrix}.$$

이제 Ridge 추정량에서 행렬 X의 특이값 분해를 $X = UDP^\top$이라고 하자. 그러면 $X^\top X = PD^\top U^\top UDP^\top = P(D^\top D)P^\top$이고, 이는 곧 $X^\top X$의 고유값 분해와 같다. 또한 $\Lambda = D^\top D$임도 알 수 있다. 최소제곱추정량을 특이값 분해를 통해 나타내면,

$$\begin{aligned} \hat\beta^{\mathrm{LSE}} &= (X^\top X)^{-1} X^\top Y \\ &= (P^\top (D^\top D)P)^{-1} PD^\top U^\top Y \\ &= P(D^\top D)^{-1} D^\top U^\top Y. \end{aligned}$$

그리고 Ridge 추정량을 특이값 분해를 통해 나타내면,

$$\begin{aligned} \hat\beta_k &= (X^\top X + kI_p)^{-1} X^\top Y \\ &= (P^\top D^\top DP + kP^\top P)^{-1} PD^\top U^\top Y \\ &= P(D^\top D + kI)^{-1} D^\top U^\top Y. \end{aligned}$$

D의 대각원소를 d_j라고 할때, $d_j/d_j^2 > d_j/(d_j+k)^2$이므로, Ridge 추정량이 최소제곱추정량을 축소하는 추정량이 됨을 알 수 있다.

고차원의 경우 $p > n$인 고차원의 경우, $(X^\top X + kI_p)^{-1} = \sum_{j=1}^{p}(d_j^2+k)p_j p_j^\top$으로 주어짐을 알 수 있다. 이때, $j > n$인 경우, $d_j = 0$이지만 $k > 0$로 인해서 역행렬이 존재함을 알 수 있다. 즉, Ridge 추정량은 고차원에서도 계산이 가능하다.

$$\lim_{k\to 0}\frac{d_j}{d_j^2+k} = \begin{cases} d_j^{-1} & d_j \neq 0 \\ 0 & d_j = 0 \end{cases} = d_j^{-1}\mathbb{I}(d_j \neq 0).$$

따라서,

$$\lim_{k\to 0}(X^\top X + kI_p)^{-1} = \sum_{j=1}^{p} d_j^{-1}\mathbb{I}(d_j \neq 0).$$

이고, Ridge Estimator가 다음과 같이 주어짐을 알 수 있다.

$$\lim_{k\to 0}\hat{\beta}_k = (X^\top X)^\dagger X^\top Y.$$

반대로 $k \to \infty$인 경우, 자명하게 회귀계수를 모두 0으로 축소함을 알 수 있다.
고차원의 경우에서 MSE를 구해보자.

$$\begin{aligned}
\mathsf{bias}(\hat{\beta}_k) &= (X^\top X + kI_p)^{-1}X^\top X\beta - \beta \\
&= (X^\top X + kI_p)^{-1}(X^\top X + kI_p - kI_p)\beta - \beta \\
&= -k(X^\top X + kI_p)^{-1}\beta \\
&= -kP(D^\top D + kI_p)^{-1}P^\top\beta.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Tr}(\mathsf{Var}(\hat{\beta}_k)) &= \sigma^2\mathrm{Tr}((X^\top X + kI_p)^{-1}X^\top X(X^\top X + kI_p)^{-1}) \\
&= \sigma^2\mathrm{Tr}(P(D^\top D + kI_p)^{-1}P^\top PD^\top DP^\top P(D^\top D + kI_p)^{-1}P^\top) \\
&= \sigma^2\mathrm{Tr}(P(D^\top D + kI_p)^{-1}D^\top D(D^\top D + kI_p)^{-1}P^\top) \\
&= \sigma^2\mathrm{Tr}((D^\top D + kI_p)^{-1}D^\top D(D^\top D + kI_p)^{-1}) \\
&= \sigma^2\sum_{j=1}^{n}\frac{d_j^2}{(d_j^2+k)^2}
\end{aligned}$$

위에서 기존의 p와 달리 n까지 더하는 이유는, $n+1,\ldots,p$까지는 $d_j = 0$이기 때문이다. 따라서 MSE는

아래와 같이 주어진다.

$$\mathsf{MSE}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{n} \frac{d_j^2}{(d_j^2+k)^2} + k^2 \sum_{j=1}^{n} \frac{p_j^\top \beta}{(d_j^2+k)^2}.$$

6 Principal Component Regression

Principal Component Regression (PCR)은 Ridge와 달리 가장 설명력이 높은 주성분 g개만 설명변수로 선택하여 회귀분석하는 것을 지칭한다.

1. $P_{p\times p} = [P_g \quad P_{-g}]$
2. $P_g^\top P = \begin{bmatrix} p_1^\top \\ \vdots \\ p_g^\top \end{bmatrix} \begin{bmatrix} p_1 & \cdots & p_g & \cdots & p_p \end{bmatrix} = \begin{bmatrix} I_{g\times g} & O_{g\times(g-p)} \end{bmatrix} = I_{g\times p}$
3. $I_{g\times p} = \begin{bmatrix} I_{g\times g} & O_{g\times(g-p)} \end{bmatrix} \begin{bmatrix} D_{g\times n}^\top \\ D_{(p-g)\times n}^\top \end{bmatrix} = D_{g\times n}^\top = D_g^\top$

이제 선형회귀모형 $Y = X\beta + \epsilon = XPP^\top\beta + \epsilon = Z\alpha + \epsilon$에서 가장 설명력이 높은 g개의 주성분만 선택한 모형 $Y = Z_g\alpha_g + \epsilon$을 고려해보자. 그러면, α_g의 최소제곱추정량은 아래와 같이 주어진다.

$$\begin{aligned}
\hat{\alpha}_g &= (Z_g^\top Z_g)^{-1} Z_g^\top Y \\
&= (P_g^\top X^\top X P_g)^{-1} P_g^\top X^\top Y \\
&= (P_g^\top P D^\top D P^\top P_g)^{-1} P_g^\top P D^\top U^\top Y \\
&= (I_{g\times p} D^\top D I_{p\times g})^{-1} I_{g\times p} D^\top U^\top Y \\
&= (D_g^\top D_g)^{-1} D_g^\top U^\top Y.
\end{aligned}$$

따라서,

$$\hat{\beta}_g = P_g \hat{\alpha}_g = P_g (D_g^\top D_g)^{-1} D_g^\top U^\top Y.$$

이제 PCR의 MSE를 조사해보자. $p > n$의 경우에, thin SVD에 의해 $X = U_{n\times n} D_{n\times n} P_{n\times p} = \sum_{j=1}^{n} d_j u_j p_j^\top$으로 분해할 수 있다. 그러면,

$$X^\top X = PD^2P^\top = \begin{bmatrix} P_g & P_s \end{bmatrix} \begin{bmatrix} \Lambda_g & 0 \\ 0 & \Lambda_s \end{bmatrix} \begin{bmatrix} P_g^\top \\ P_s^\top \end{bmatrix}$$

이고, 최소제곱추정량은 다시 $\hat{\beta}_g = P_g \Lambda_g^{-1} P_g^\top X^\top Y$로 표현할 수 있다.

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_g) &= P_g \Lambda_g^{-1} P_g^\top X^\top X \beta \\
&= (P_g \Lambda_g^{-1} P_g^\top)(P \Lambda P^\top) \beta \\
&= \left(\sum_{j=1}^{g} \frac{1}{\lambda_j} p_j p_j^\top \right) \left(\sum_{k=1}^{m} \lambda_k p_k p_k^\top \right) \beta \\
&= \left(\sum_{j=1}^{g} p_j p_j^\top \right) \beta \\
&= \beta - \sum_{j=g+1}^{m} (p_j^\top \beta)\, p_j
\end{aligned}$$

위의 계산에서는 P가 직교하므로 $\sum_{j=1}^{p} p_j p_j^\top = I$라는 사실을 사용하였다.

$$\begin{aligned}
\mathsf{Var}(\hat{\beta}_g) &= \sigma^2 P_g \Lambda_g^{-1} P_g^\top X^\top X P_g \Lambda_g^{-1} P_g^\top \\
&= \sigma^2 \left(\sum_{j=1}^{g} \frac{1}{\lambda_j} p_j p_j^\top \right) \left(\sum_{k=1}^{m} \lambda_k p_k p_k^\top \right) \left(\sum_{j=1}^{g} \frac{1}{\lambda_j} p_j p_j^\top \right) \\
&= \sigma^2 \left(\sum_{j=1}^{g} \frac{1}{\lambda_j} p_j p_j^\top \right)
\end{aligned}$$

$$\begin{aligned}
\mathsf{MSE}(\hat{\beta}_g) &= \mathrm{Tr}\left(\mathsf{Var}(\hat{\beta}_g)\right) + \|\mathsf{bias}(\hat{\beta}_g)\|^2 \\
&= \sigma^2 \sum_{j=1}^{g} \lambda_j^{-1} + \left\| \sum_{j=g+1}^{m} (p_j^\top \beta)\, p_j \right\|^2 \\
&= \sigma^2 \sum_{j=1}^{g} \lambda_j^{-1} + \sum_{j=g+1}^{m} (p_j^\top \beta)^2 \quad (P \text{ orthogonal} \Rightarrow \text{cross terms vanish})
\end{aligned}$$

References

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

박성현, 이성임, and 임요한. *고급회귀분석*. 민영사, 2023.