# Elastic net

## Reference

- Regularization and variable selection via the elastic net, Zou et al.

## Introduction

elastic net은 lasso 이후에 제안된 방법이다. lasso는 ridge 이후에 제안된 방법으로, ridge와는 달리 lasso는 추정된 변수의 일부를 0으로 만들어서 변수 선택의 효과를 가져온다. 언뜻보면 lasso가 ridge보다 좋아보이지만 꼭 그렇다고만 말할 수 없다. 왜냐하면 보통의 $n > p$일 때 변수 사이의 상관성이 높다면 lasso의 성능이 ridge보다 좋지 않다는 결과가 자주 보고되기 때문이다. 더욱이 lasso만의 문제가 있는데, 이는 $p > n$일 때, lasso는 최대 $n$개의 변수만 선택할 수 있다. 만약 $p >> n$인 경우, 예를 들어 유전자 데이터에서는 $p = 7000, n = 200$일 때도 있는데, 7000개의 변수를 최대 200개로 줄인다면 그 성능이 의심되는 부분이다. 또한 변수간 상관성이 높다면 그 상관성이 높은 그룹에서 한 변수만 선택하는 경향이 있다. 생각해보면 이는 non-sense인데, 한 그룹의 변수들이 서로 상관되어 있으므로 그 중 하나의 변수만 의미가 있을리 없기 때문이다.
이러한 이유로, lasso가 variable selection의 대안으로 나왔다지만, 사실 $p >>> n$인 경우, variable selection을 사실상 잘 하지 못하는 것으로 밝혀졌다.
elastic net은 lasso의 변수 선택 장점은 취하되, 위에서 언급한 단점들은 보완한 모델이다. 즉, $p >> n$인 경우에 최대 $n$개의 변수만 선택했던 lasso와는 달리, 더 많은 변수를 선택해주고 서로 상관성이 높은 변수 중 일부만 선택하는 것이 아니라 그 그룹을 모두 선택한다. naive elastic net을 설명하고 elastic net, 그리고 numerical solution을 얻기 위한 LARS-EN을 소개한다.

## Naive elastic net

$n$개의 데이터와 $p$개의 변수가 있다고 생각하자. $y$는 centering을 통해 평균을 0으로, 각 변수들도 column 별로 centering 및 표준화를 끝냈다고 하자. non-negative $\lambda_1, \lambda_2$에 대해서 최소화하고자 하는 quantity을 아래와 같이 정의한다.

$$L(\lambda_1, \lambda_2, \beta) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2|\boldsymbol{\beta}|^2 + \lambda_1|\boldsymbol{\beta}|_1 \tag{1}$$

naive elastic net estimator는

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}\, L(\lambda_1, \lambda_2, \boldsymbol{\beta}) \tag{2}$$

이는 $\alpha = \dfrac{\lambda_2}{\lambda_1 + \lambda_2}$인 penalized least squares로 볼 수도 있다.

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2,\ subject\ to\ (1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \le t \tag{3}$$

Lemma 1에 의하면 (1)에 대한 solution은 lasso 타입의 optimization 문제를 푸는 것과 동일하다. 따라서 elastic net의 optimization 문제를 풀 때, lasso에서 사용했던 계산의 이점을 이용할 수 있다.

*Lemma 1.* Given data set $(\mathbf{y}, \mathbf{X})$ and $(\lambda_1, \lambda_2)$, define an artificial data set $(\mathbf{y}^*, \mathbf{X}^*)$ by

$$\mathbf{X}^*_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{I}\end{pmatrix}, \qquad \mathbf{y}^*_{(n+p)} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix}.$$

Let $\gamma = \lambda_1/\sqrt{(1+\lambda_2)}$ and $\beta^* = \sqrt{(1+\lambda_2)}\beta$. Then the naïve elastic net criterion can be written as

$$L(\gamma, \beta) = L(\gamma, \beta^*) = |\mathbf{y}^* - \mathbf{X}^*\beta^*|^2 + \gamma|\beta^*|_1.$$

Let

$$\hat{\beta}^* = \arg\min_{\beta^*} L\{(\gamma, \beta^*)\};$$

then

$$\hat{\beta} = \frac{1}{\sqrt{(1+\lambda_2)}}\hat{\beta}^*.$$

Lemma 1에 의하면 elastic net 해를 구하는 것이 augmented matrix을 만들고 이에 대한 lasso 해를 구하는 것과 동일하다. augmented matrix는 $(n+p)\times p$이며 $\mathbf{X}^*$는 rank가 $p$이므로 결국 $\mathbf{X}^*$를 이용한 lasso해는 $p$개 모두를 선택할 수 있다. 이로써 elastic net이 lasso의 단점 중 하나, 변수 선택을 최대 $n$개까지 할수 있다는 단점을 보완함을 알 수 있다.

다음으로는 lasso가 상관성 높은 변수 그룹에서 한 개의 변수만 고르는 단점을 elastic net이 어떻게 보완하는지 알아보자. 초고차원 데이터, $p >> n$인 데이터는 현실에서 종종 발생한다. 대표적인 예가 유전자 데이터이다. 유전자 데이터는 샘플 수가 곧 환자의 수인데, 보통 200을 넘지 않는다. 반면에, 한 사람의 유전자 변수는 수천개에서 많으면 10000개까지 된다. 이렇게 많은 변수 각각이 의미가 있기 보다는, 서로 상관성이 아주 높기 마련이다. 즉, 우리가 모르는 그룹이 있을텐데, 그 그룹을 모델이 잘 탐지해서 의미가 있는 그룹과 그렇지 않은 그룹으로 나누어야 한다. 이때, 의미가 있는 그룹 A을 생각해보자. 그룹 A에 있는 변수들은 상관성이 높을 텐데, lasso는 이 변수들 중에 한 개만 선택하는 경향이 있다. 한 개의 변수보다는 그 변수가 속해있는 set을 선택하는 것이 더 좋음에도, 이러한 경향성 때문에 lasso의 단점이 지적되어왔던 것이다.
극단적으로 어떤 두 변수가 같다고 가정해보자. 즉, 상관계수가 1인 두 변수의 계수 추정치는 같아야 할 것이다.

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}}|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda J(\boldsymbol{\beta})$$

로 정의하고 아래 Lemma 2를 살펴보자.

*Lemma 2.* Assume that $\mathbf{x}_i = \mathbf{x}_j$, $i, j \in \{1, \ldots, p\}$.

(a) If $J(\cdot)$ is strictly convex, then $\hat{\beta}_i = \hat{\beta}_j$, $\forall\lambda > 0$.
(b) If $J(\beta) = |\beta|_1$, then $\hat{\beta}_i\hat{\beta}_j \geqslant 0$ and $\hat{\beta}^*$ is another minimizer of equation (7), 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 $s \in [0, 1]$.

Lemma 2는 $J(\boldsymbol{\beta})$가 strict convex라면 두 추정치가 같지만, $L_1$ norm은 strict convex가 아니기 때문에 unique solution도 없을 뿐더러 같은 추정치를 내지도 않는다. 하지만 elastic net은 $\lambda_2$ 항으로 인해서 strict convex하고 따라서 grouping effect가 보장되는 것이다.

## Elastic net

naive elastic net은 lasso의 두 단점을 보완하지만, 실험에서는 ridge나 lasso에 매우 가깝지 않으면 만족할만한 결과를 내지 않는다고 한다. 이러한 이유로 인해서 저자들이 'naive'라고 이름 붙인 것이다. navie elastic net은 two-stage 모델이다. $\lambda_2$를 고정하고 ridge regression 계수를 찾고, lasso 타입의 shrinkage를 한다. 즉 두 번의 shrinkage를 하는 셈인데, 이는 variance를 크게 줄이지는 못하고 오히려 불필요한 bias를 증가시킨다. 이러한 navie elastic net의 단점을 보완하기 위해 elastic net이 제안되었다.

augmented data인 $(\mathbf{y}^*, \mathbf{X}^*)$와 penalty parameter $(\lambda_1, \lambda_2)$에 대해서 naive elastic net은 아래의 lasso 형태 optimization을 푸는 것과 동일하다.

$$\hat{\boldsymbol{\beta}}^* = argmin_{\boldsymbol{\beta}^*} |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}}|\boldsymbol{\beta}^*|_1$$

elastic net 추정치 $\hat{\boldsymbol{\beta}}^E$은 아래와 같이 정의된다. (논문에서 정의된다고 나와있고 구체적인 설명은 없다.)

$$\hat{\boldsymbol{\beta}}^E = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*$$

naive elastic net 추정치는

$$\hat{\boldsymbol{\beta}}^{NE} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$$

이므로

$$\hat{\boldsymbol{\beta}}^E = (1+\lambda_2)\hat{\boldsymbol{\beta}}^{NE}$$

즉 elastic net 추정치는 naive elastic net 추정치를 rescale한 것이다. 이러한 rescaling은 변수 선택 성질을 보존하고 불필요한 shrinkage을 줄인다고 한다.

**Why $1+\lambda_2$?**

왜 하필 rescaling factor가 $1+\lambda_2$ 일까? 이는 ridge의 성질과 연관시켜서 봐야한다. $\lambda_2$는 ridge의 penalty, $|\boldsymbol{\beta}|^2$와 관련된 hyperparameter이다. 표준화된 $\mathbf{X}$에 대한 ridge regression 추정치는 $\hat{\boldsymbol{\beta}}^R = \mathbf{Ry},\ where\ \mathbf{R} = \frac{1}{1+\lambda_2}\mathbf{R}^* = \frac{1}{1+\lambda_2}\begin{pmatrix} 1 & \frac{\rho_{12}}{1+\lambda_2} & \cdot & \frac{\rho_{1p}}{1+\lambda_2} \\ & 1 & \cdot & \cdot \\ & & 1 & \frac{\rho_{p-1,p}}{1+\lambda_2} \\ & & & 1 \end{pmatrix}\mathbf{X}^T$ 인데, 여기서 $\mathbf{R}^*$은 correlation이 $1/(1+\lambda_2)$만큼 shrink되었다는 것(이를 decorrelation이라 부른다.)을 제외하면, 기존의 OLS이다. 따라서 ridge 추정치에 대한 위 분해로부터 ridge perator가 decorrelation을 하고, 또 direct scaling, 즉 전체 행렬을 $1+\lambda_2$로 나눔을 알 수 있다. 또한 이 분해는 ridge의 grouping effect가 decorrelation step의해서 발생함을 암시한다 (이렇게만 나와있고 구체적인 증명은 없다.) 그니까 elastic net에서 $1/(1+\lambda_2)$

로 rescale함으로써 ridge의 decorrelation 효과도 얻고 (grouping effect도 자연스럽게 얻는다) lasso penalty을 사용하여 variance을 control하는 것이다.