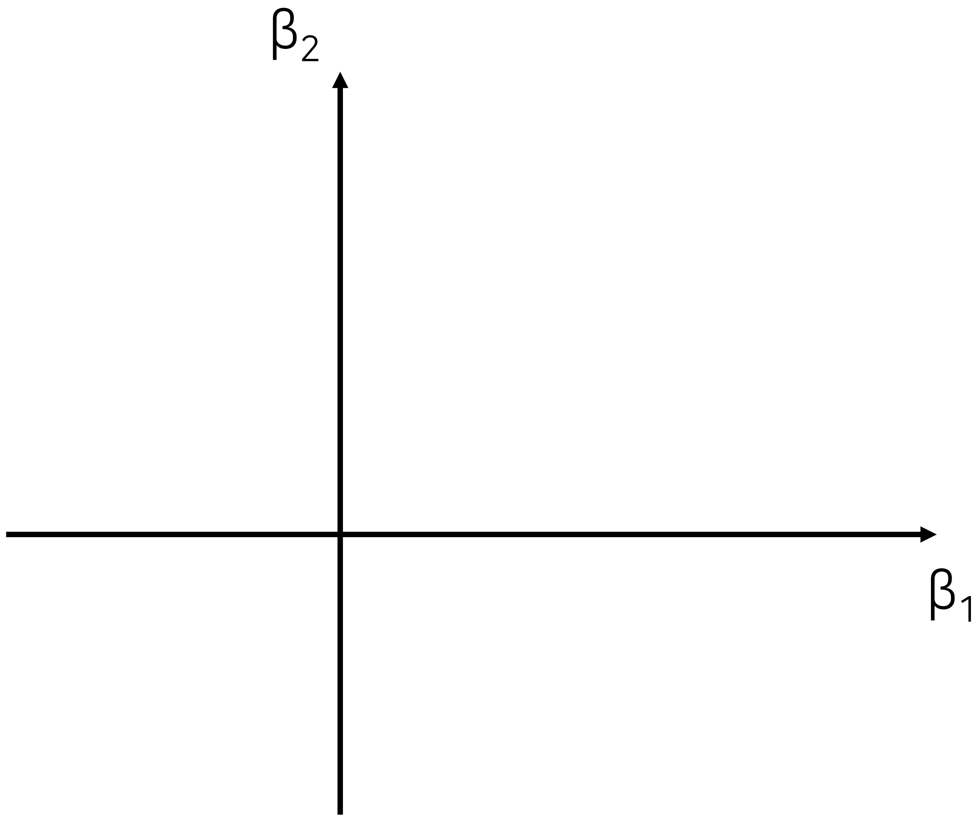
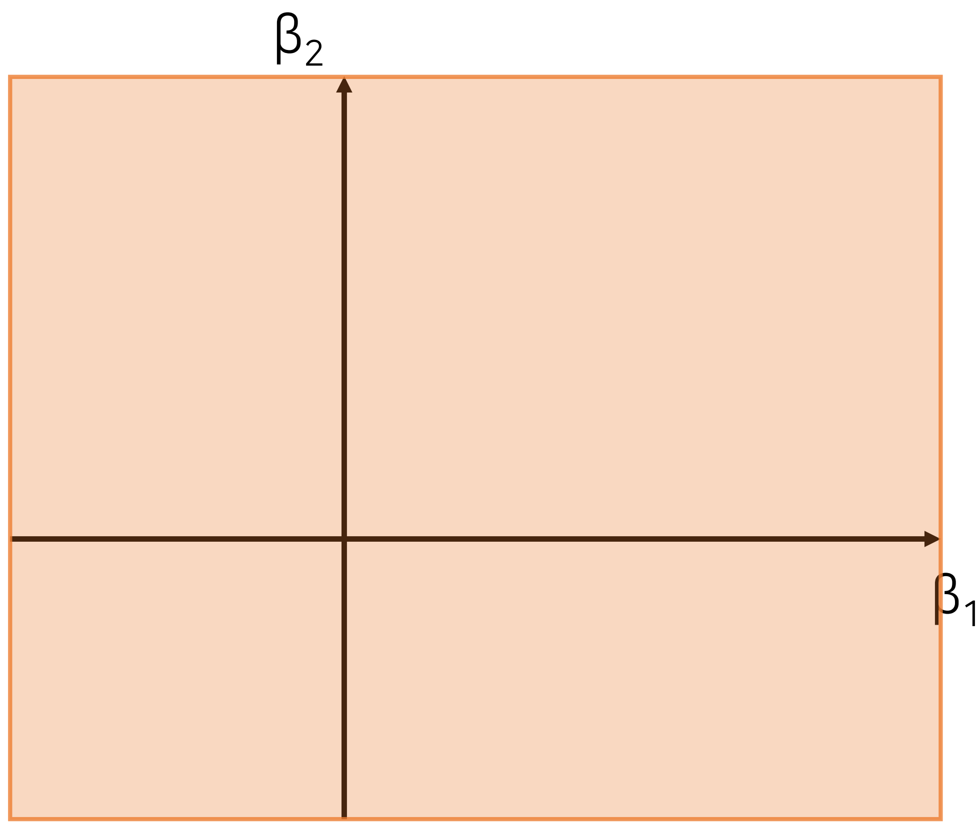
Ridge Regression method

이해를 돕기 위해 feature가 2개인 경우를 생각해 보자.



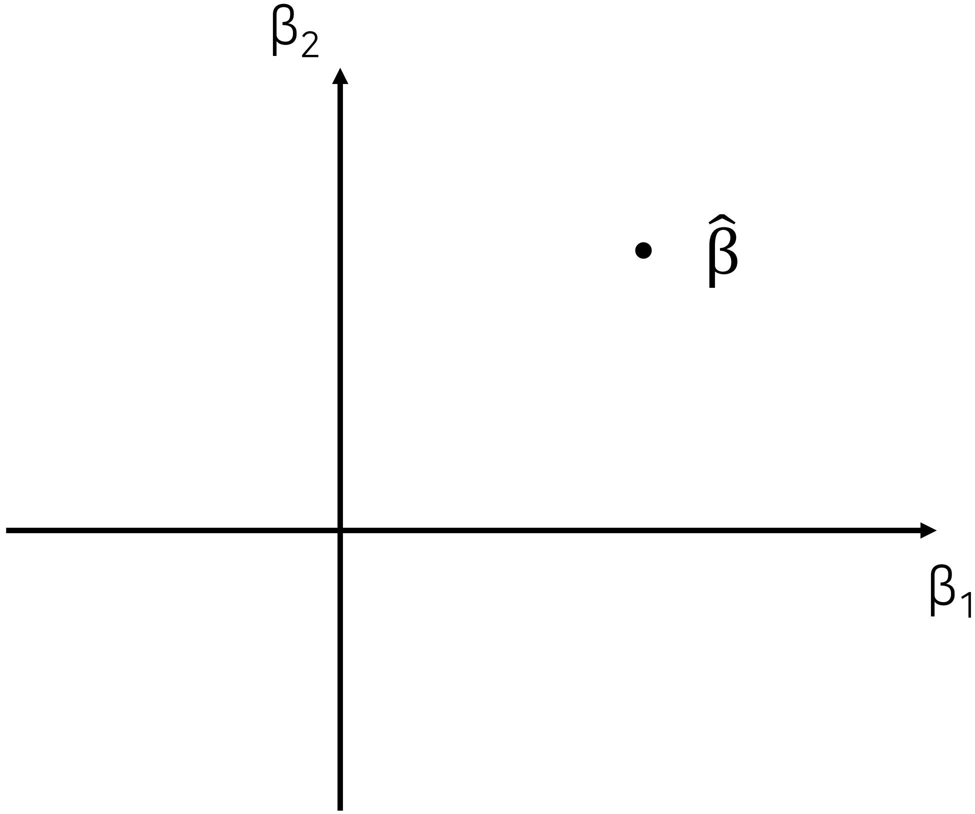
Ridge Regression method

여기서 고를 수 있는 (β_1, β_2)의 조합은 무수히 많다.



Ridge Regression method

어떤 계수가 '좋은' 계수인지 조건이 필요했고,
평균제곱오차(MSE)를 기준으로 사용해 계수를 추정하였다(최소제곱법).
그 계수인 β_1과 β_2 를 점 (β_1, β_2)인 $\hat{\beta}$ 이라고 하자.

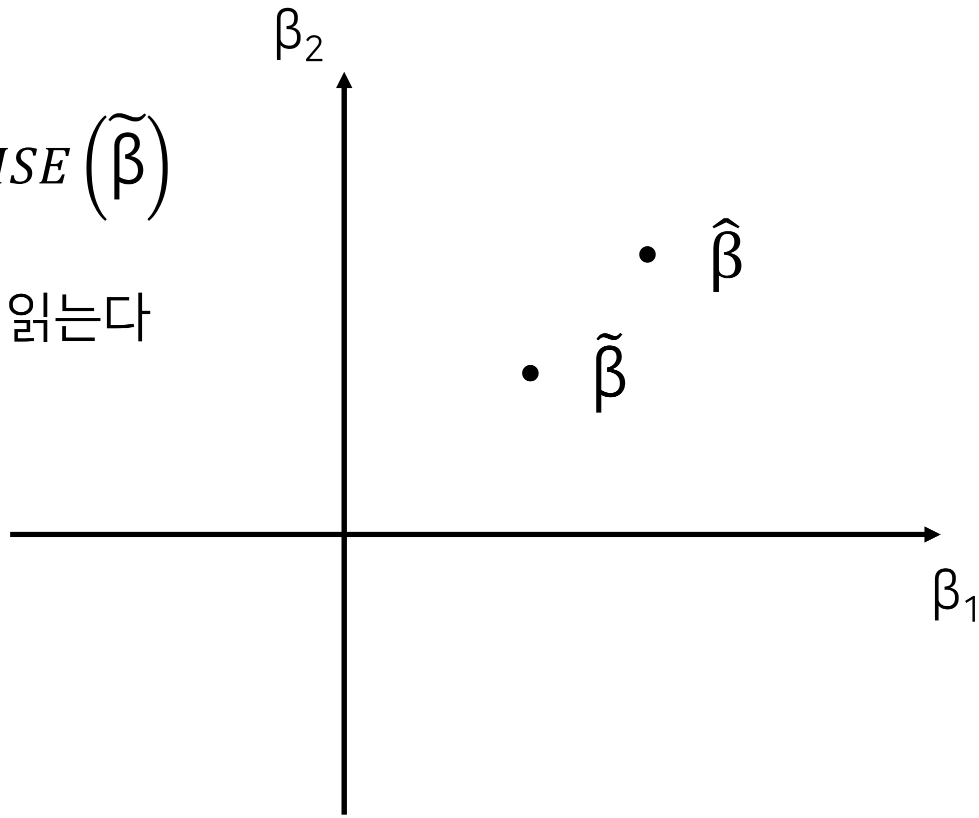


Ridge Regression method

$\hat{\beta}$는 MSE가 가장 작은 점이고, (β_1, β_2)의 조합에 따라 MSE가 달라진다.
$\hat{\beta}$에서 멀어질수록 MSE가 커진다.

$$MSE\left(\hat{\beta}\right) < MSE\left(\tilde{\beta}\right)$$

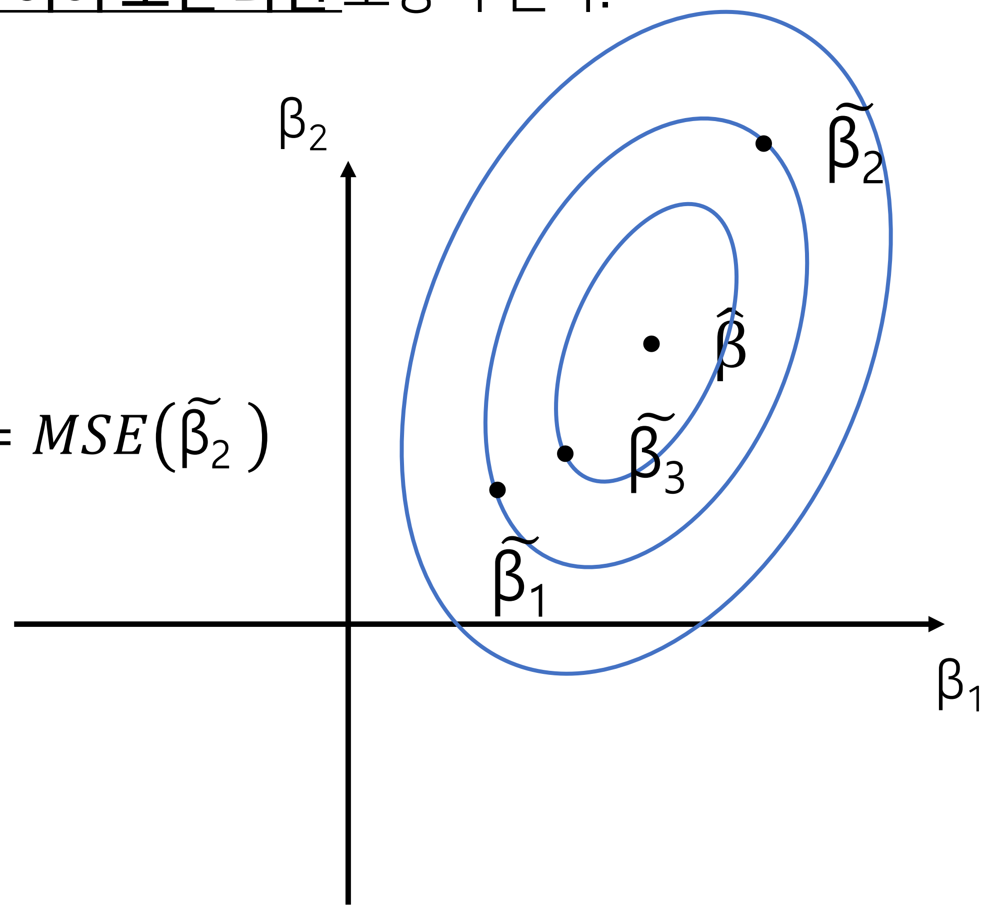
"~"는 tilde라고 읽는다

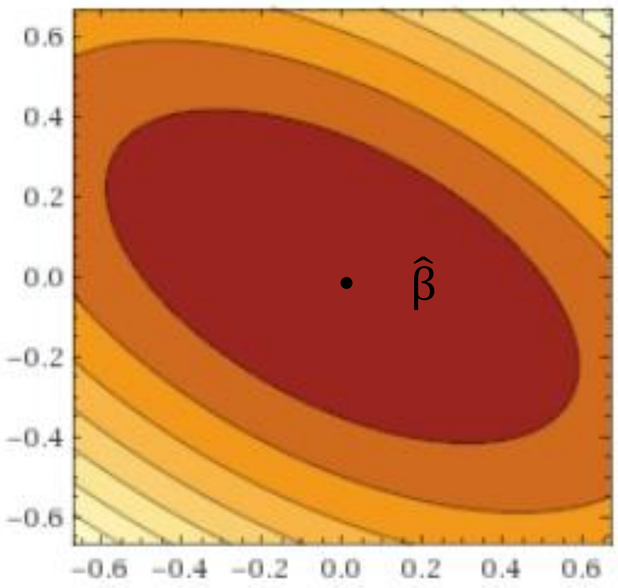


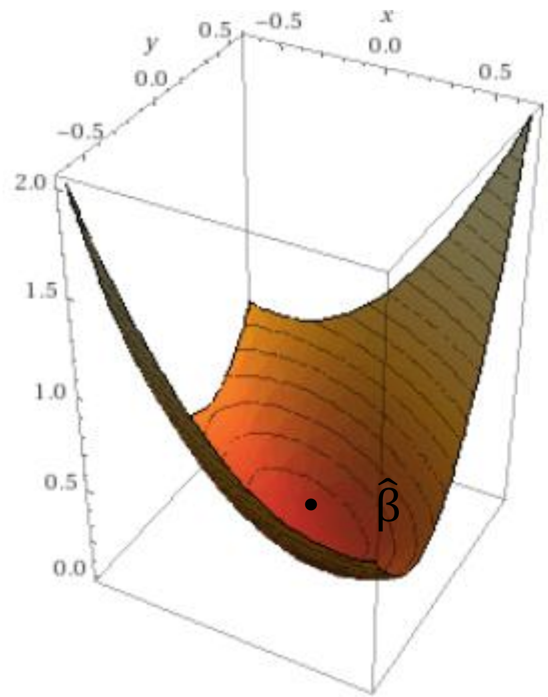
Ridge Regression method

그런데 알려진 사실에 따르면, MSE가 커지는 데는 어떤 '규칙'이 있다.
<u>MSE가 같은 점을 이어 보면 타원</u> 모양이 된다.

$$MSE(\hat{\beta}) < MSE(\tilde{\beta}_3) < MSE(\tilde{\beta}_1) = MSE(\tilde{\beta}_2)$$

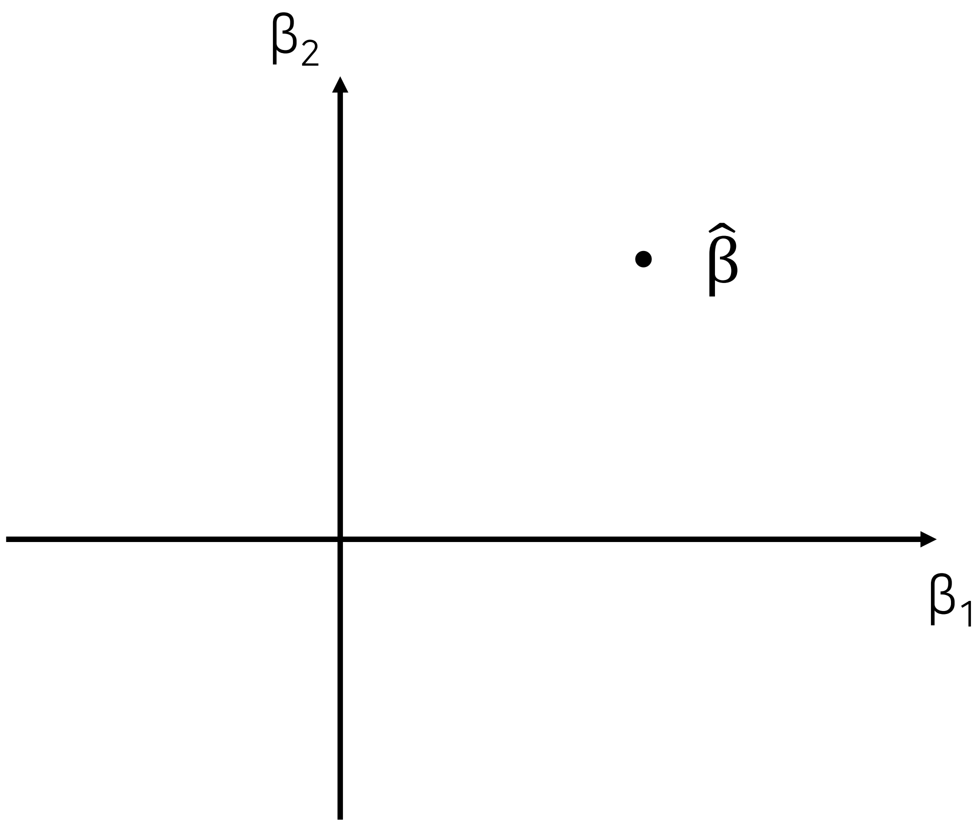



0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
$\hat{\beta}$


x
y
0.5
−0.5
0.0
0.5
0.0
−0.5
2.0
1.5
1.0
0.5
0.0
$\hat{\beta}$

Ridge Regression method

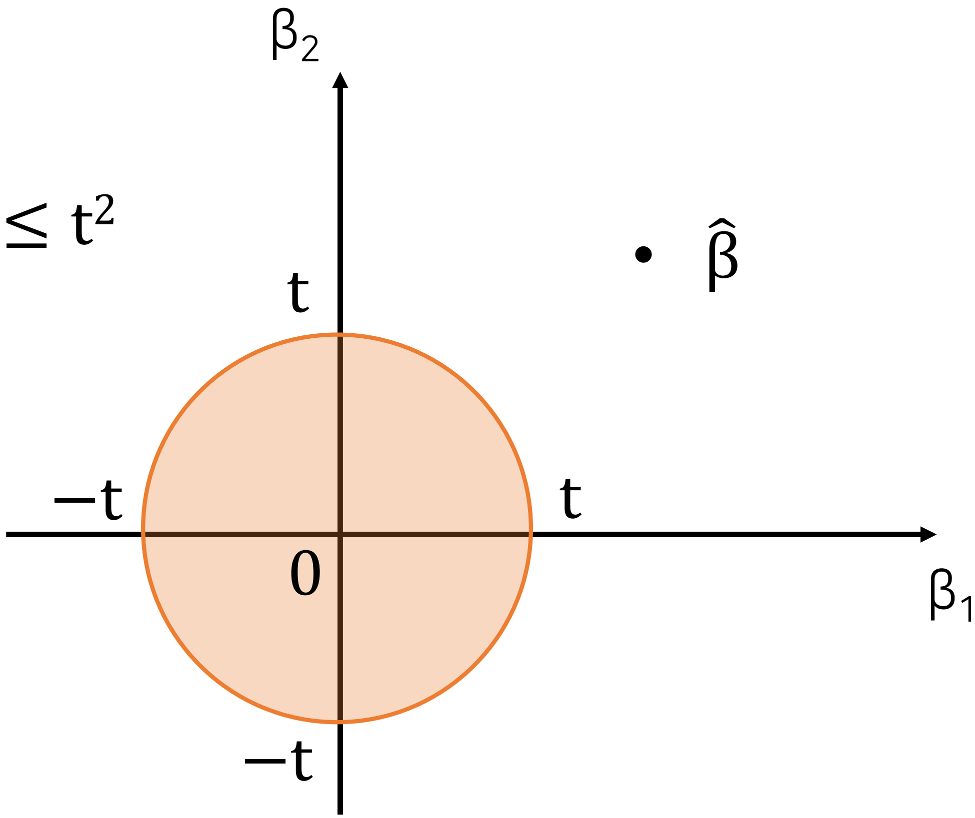
한편 회귀계수의 크기가 너무 크면 과적합(overfitting)의 가능성이 있다.



Ridge Regression method

Ridge regression: 회귀계수의 범위를 원 안으로 제한한다.
단, 평가 기준은 똑같이 MSE를 사용

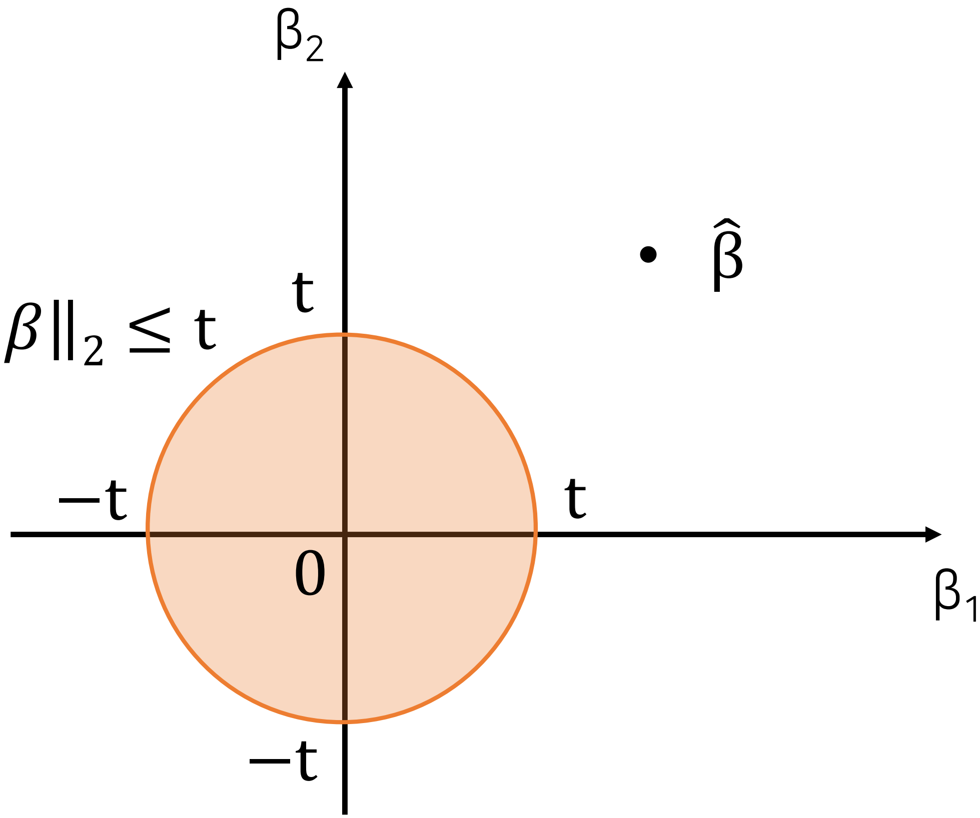
$$\beta_1{}^2 + \beta_2{}^2 \leq t^2$$



Ridge Regression method

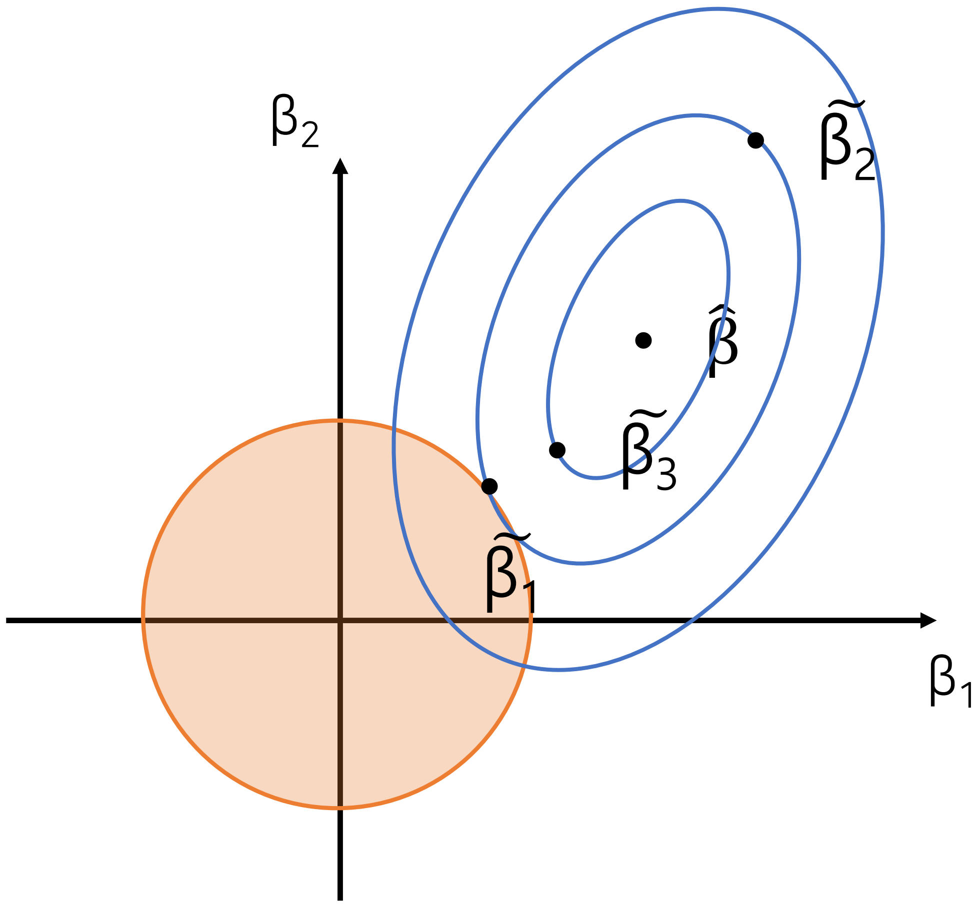
회귀계수의 범위를 원 안으로 제한
= 회귀계수의 L2 크기(=원점과의 L2 거리)를 제한

$$\|\beta - 0\|_2 = \|\beta\|_2 \le t$$



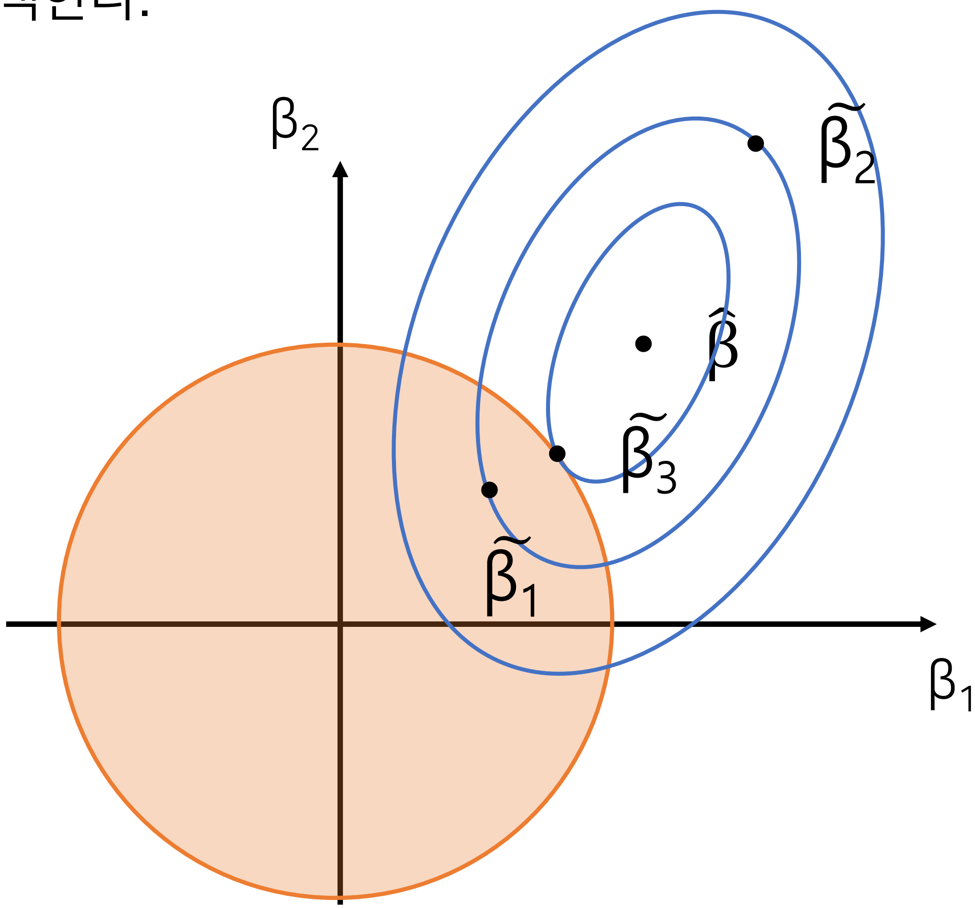
Ridge Regression method

원 안의 회귀계수중 MSE가 제일 작은 점은 $\tilde{\beta}_1$이다.



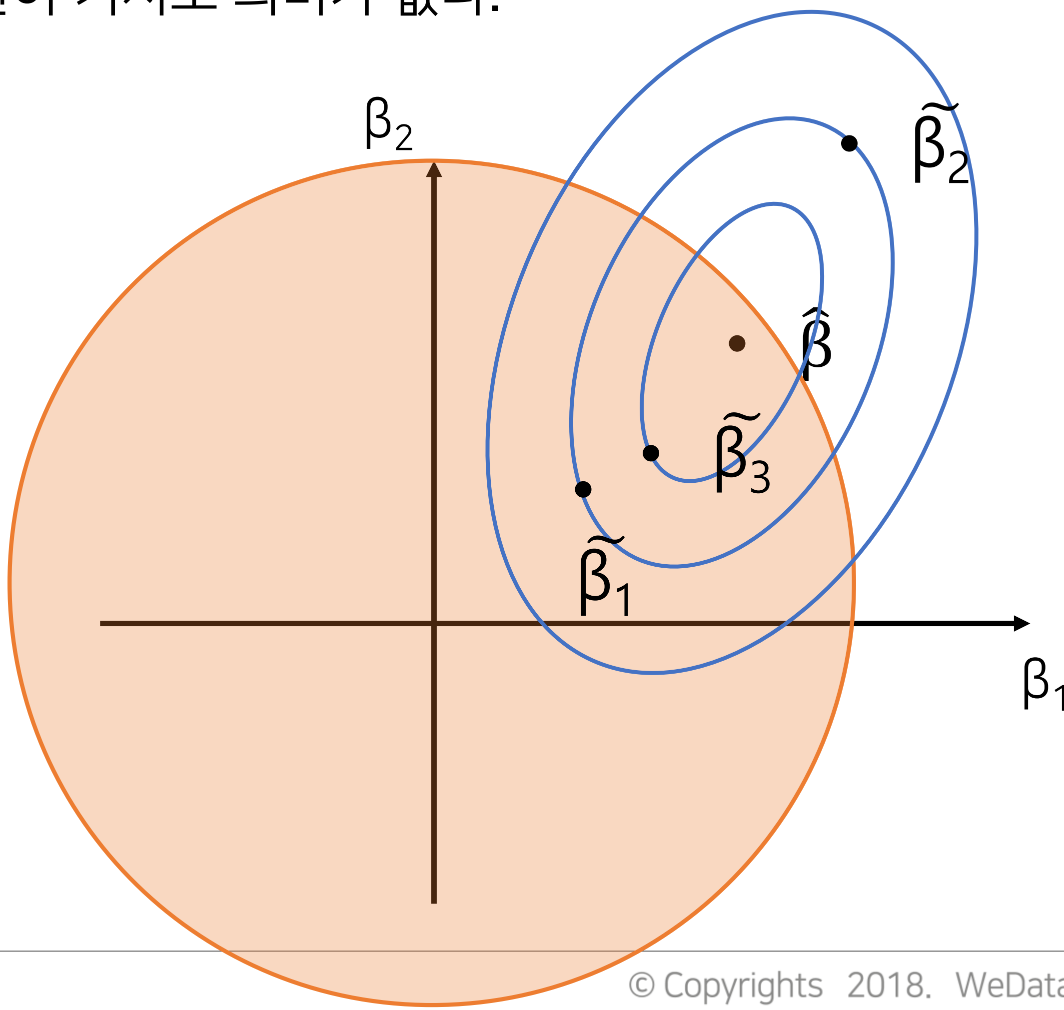
Ridge Regression method

원의 반지름이 커지면 선택되는 회귀계수도 달라진다.
이 경우는 $\tilde{\beta}_3$를 선택한다.

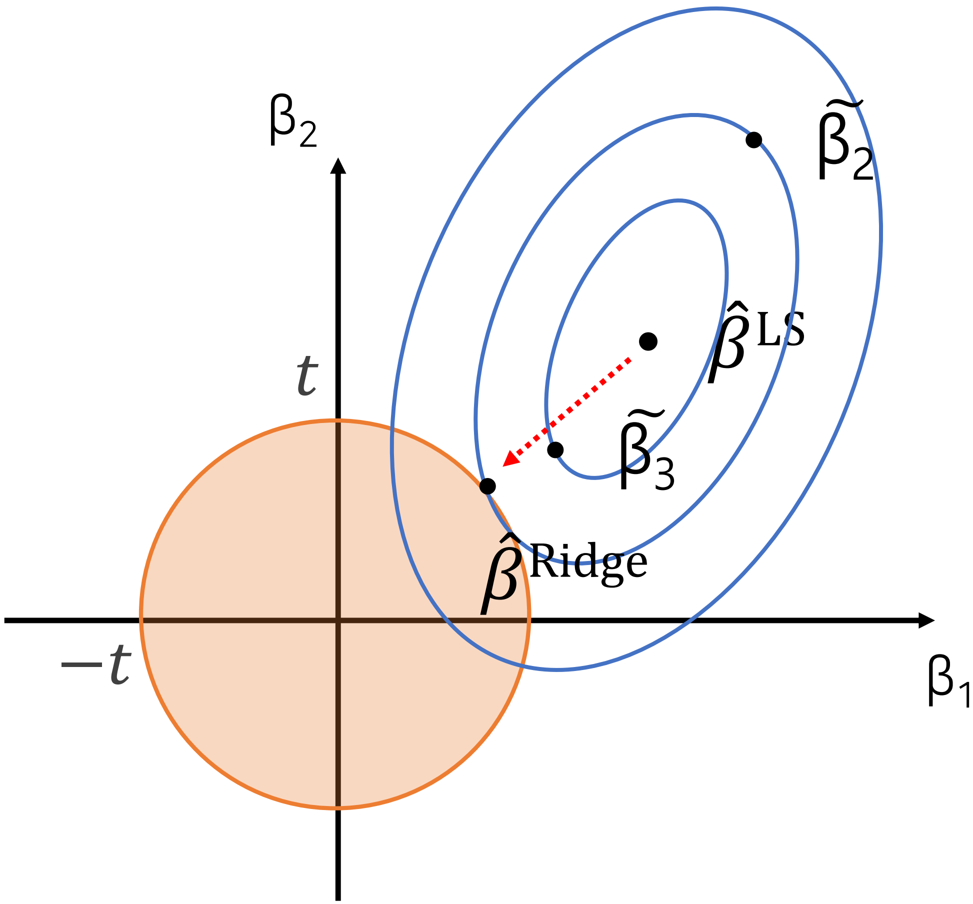


Ridge Regression method

원이 최소제곱법으로 추정한 $\hat{\beta}$을 포함할 정도로 커지면 항상 $\hat{\beta}$을 선택한다.
이때부터는 원이 커져도 의미가 없다.



Ridge Regression method 요약



$t = 0$이면 회귀계수는 0

$t = \infty$이면 회귀계수는 $\hat{\beta}^{\text{LS}}$

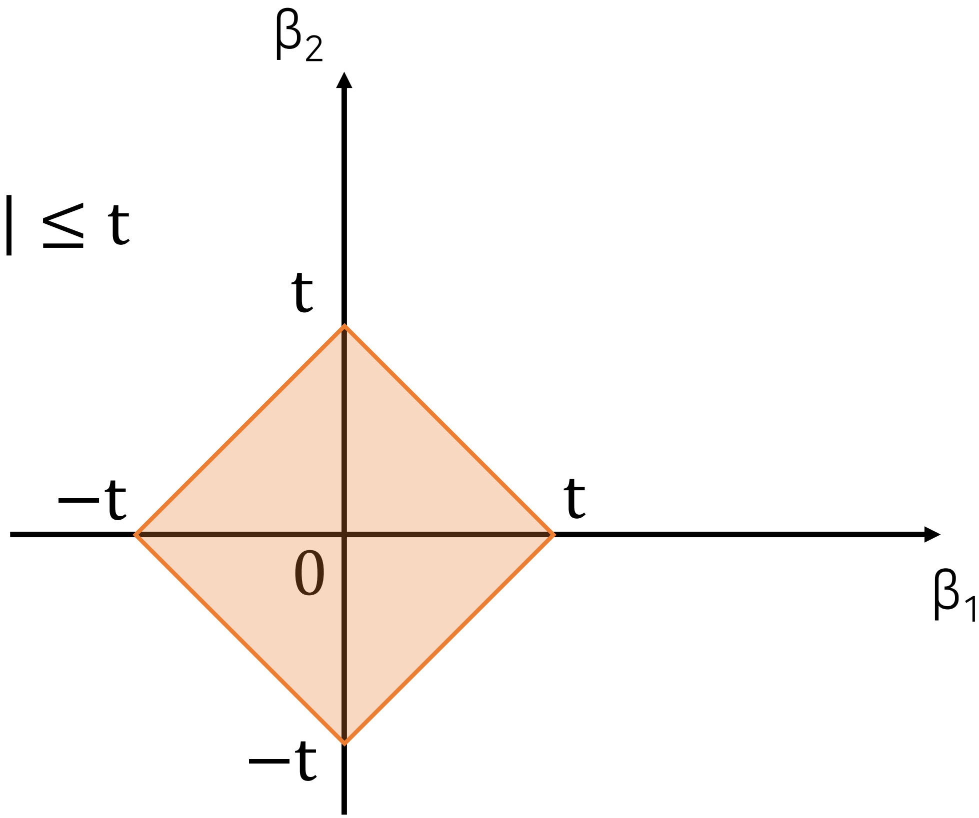
타원과 원의 교점이 $\hat{\beta}^{\text{Ridge}}$

위 과정을 통해 회귀계수의
크기가 작아진다.

LASSO Regression method

LASSO regression: 회귀계수의 범위를 마름모 안으로 제한한다.
단, 평가 기준은 똑같이 MSE를 사용

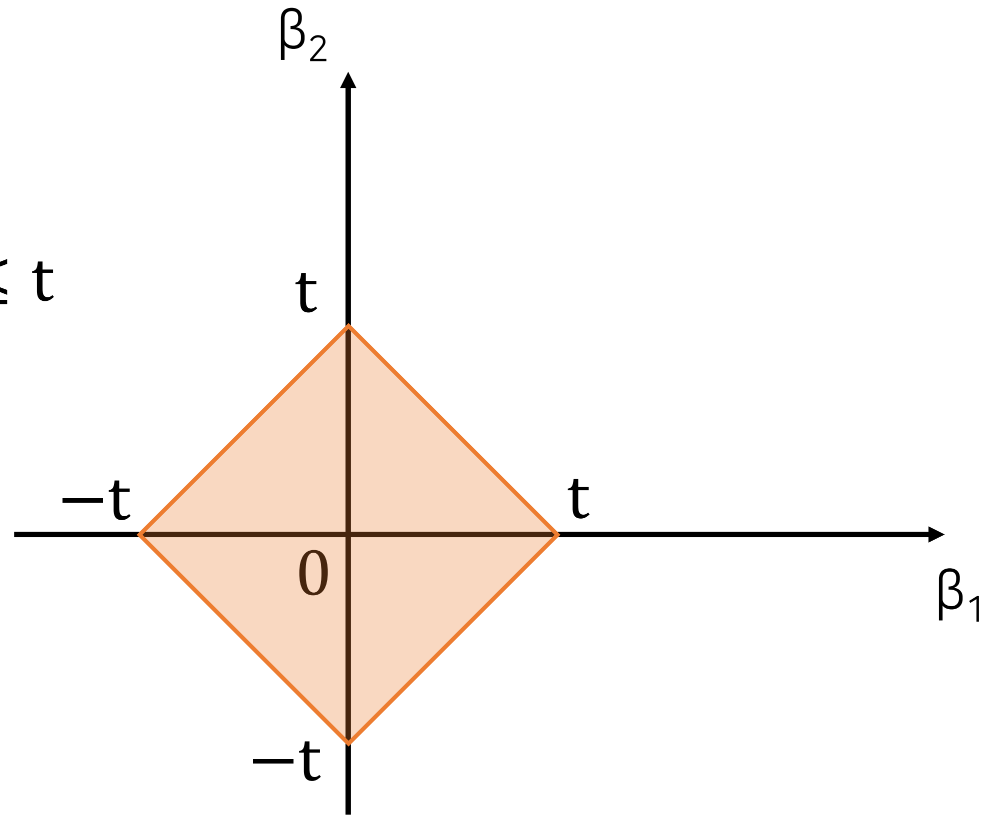
$$|\beta_1| + |\beta_2| \leq t$$



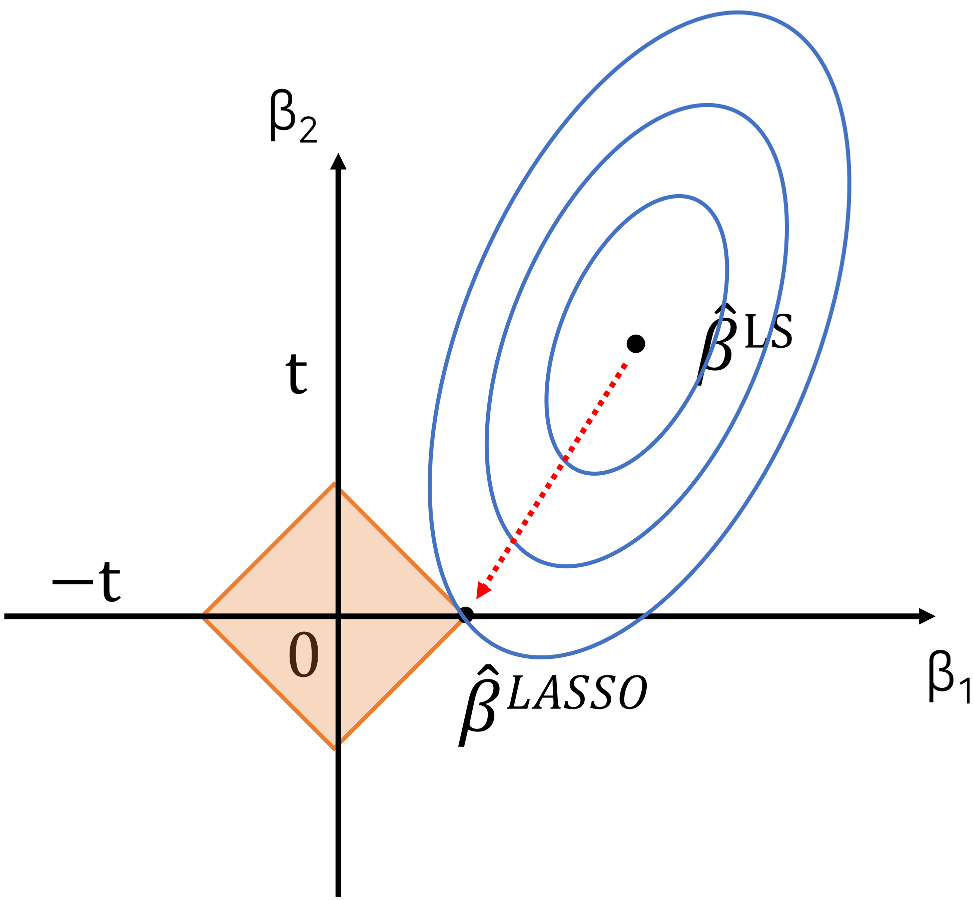
LASSO Regression method

LASSO regression: 회귀계수의 범위를 마름모 안으로 제한한다.
단, 평가 기준은 똑같이 MSE를 사용

$$\|\beta\|_1 \leq t$$



LASSO Regression method 요약

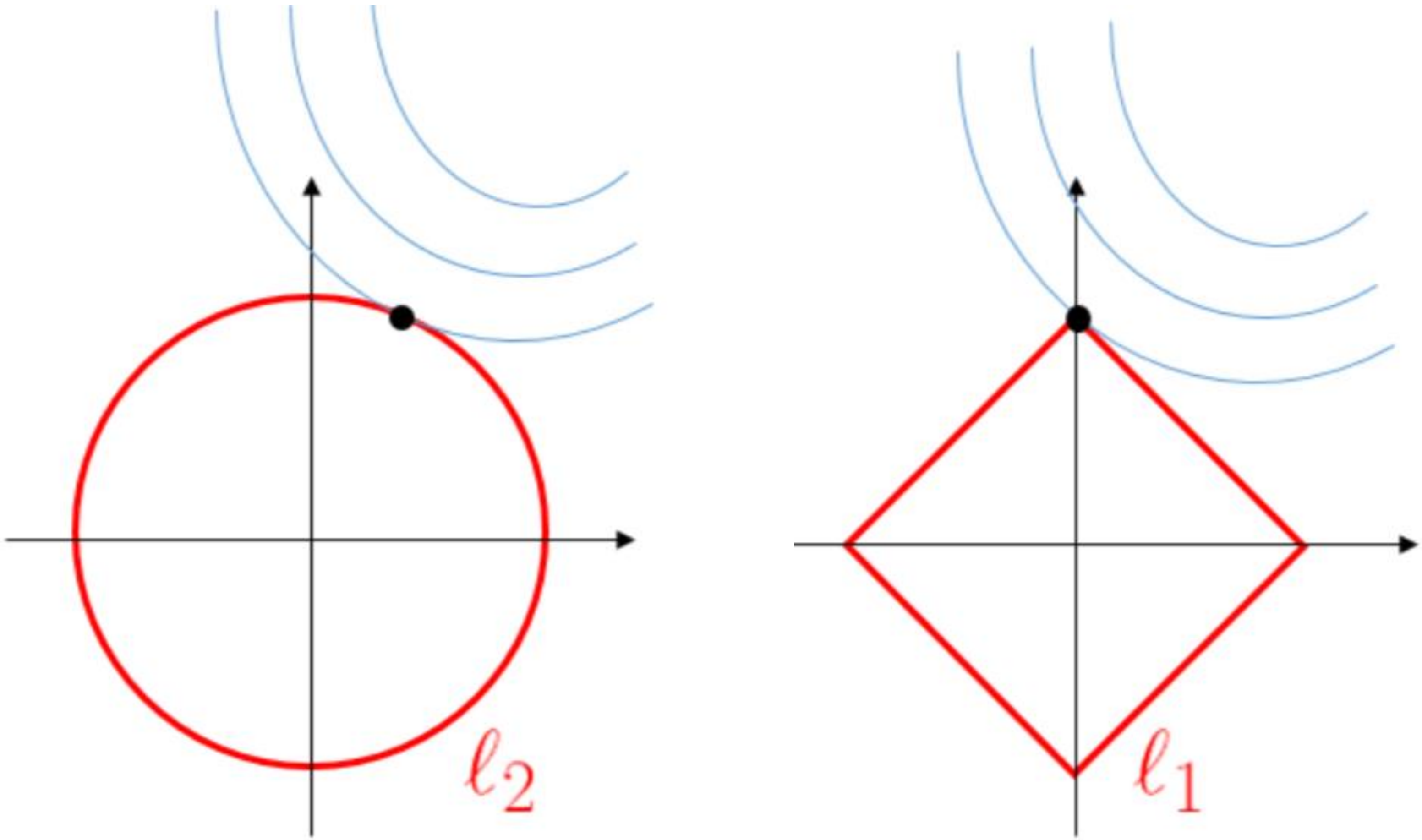


$t = 0$이면 회귀계수는 0
$t = \infty$이면 회귀계수는 $\hat{\beta}^{\mathrm{LS}}$

마름모와 원의 교점이 $\hat{\beta}^{\mathrm{Lasso}}$

위 과정을 통해
회귀계수 대부분이 0이 된다
(sparsity).

Ridge vs LASSO



(참고) LASSO Regression(L1 Regularization)

회귀계수 대부분을 0으로 보내기 때문에 변수 선택의 역할이 가능하다.

Ordinary Least Square method: min $\sum_{i=1}^{n}(y\,i - \hat{y}i)2$

$\Leftrightarrow$ min $(y - X\beta)T(y - X\beta)$

LASSO Regression method: min $\sum_{i=1}^{n}(y\,i - \hat{y}i)2$ s.t. $\sum_{j=1}^{p} |\beta j| \leq t$

$\Leftrightarrow$ min $(y - X\beta)T(y - X\beta) + \boxed{\lambda||\beta||1}$

MSE뿐만 아니라 회귀계수의 (L1)크기까지 최소화
λ 가 커진다 -> 벌점이 커진다 -> 더 작은 β 를 사용해야 한다.

(참고) Ridge Regression(L2 Regularization)

회귀계수를 추정할 때 값이 커지지 않도록 **벌점(L2 penalty)을 부여**한다.

Ordinary Least Square method: $\min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$\Leftrightarrow \min\ (y - X\beta)^T(y - X\beta)$

Ridge Regression method: $\min \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ s.t. $\sum_{j=1}^{p}(\beta_j)^2 \le t^2$

$\Leftrightarrow \min\ (y - X\beta)^T(y - X\beta) + \boxed{\lambda||\beta||^2}$

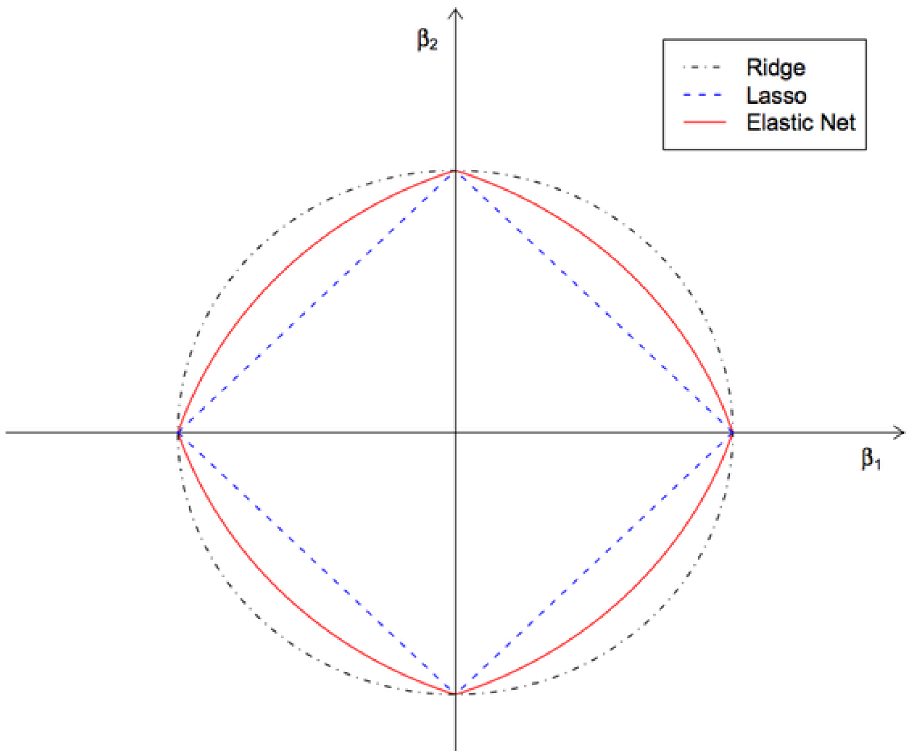
MSE뿐만 아니라 회귀계수의 (L2)크기까지 최소화
λ 가 커진다 -> 벌점이 커진다 -> 더 작은 β 를 사용해야 한다.

Elasticnet: LASSO + Ridge (hybrid method)

min $(y - X\beta)T(y - X\beta)$ $+\lambda 2||\beta||2 + \lambda 1||\beta||1$

$\lambda 2$ =0 이면 LASSO regression, $\lambda 1$ =0 이면 Ridge regression

여기서 $\lambda 1$, $\lambda 2$ 는 모두 사전에 지정해줘야 하는 hyperparameter이다.



(참고) 데이터 표준화

LASSO 회귀분석은 회귀계수의 크기(norm)에 제약조건을 걸어준다.

그런데 회귀계수 $\beta_1, \beta_2, \ldots, \beta_p$ 는 각각 특징 $X_1, X_2, \ldots, X_p$ 와 관련되어 있다.

따라서 회귀계수는 각 특징의 단위(scale)에 의존하고,
표준화를 하지 않고 분석을 시행하면 문제가 생길 수 있다.