Regularized Linear Models

Data Scientist
안건이

목차

- Regularization
- Ridge
- LASSO
- ElasticNet
- 데이터 실습

What is a good model ?

현재 데이터(Training data)를 잘 설명하는 모델

→ Explanatory modeling

미래 데이터(Testing data)에 대한 예측 성능이 좋은 모델

→ Predictive modeling

Good Explanatory Model

현재 데이터(Training data)를 잘 설명하는 모델

= Training Error를 Minimize하는 모델

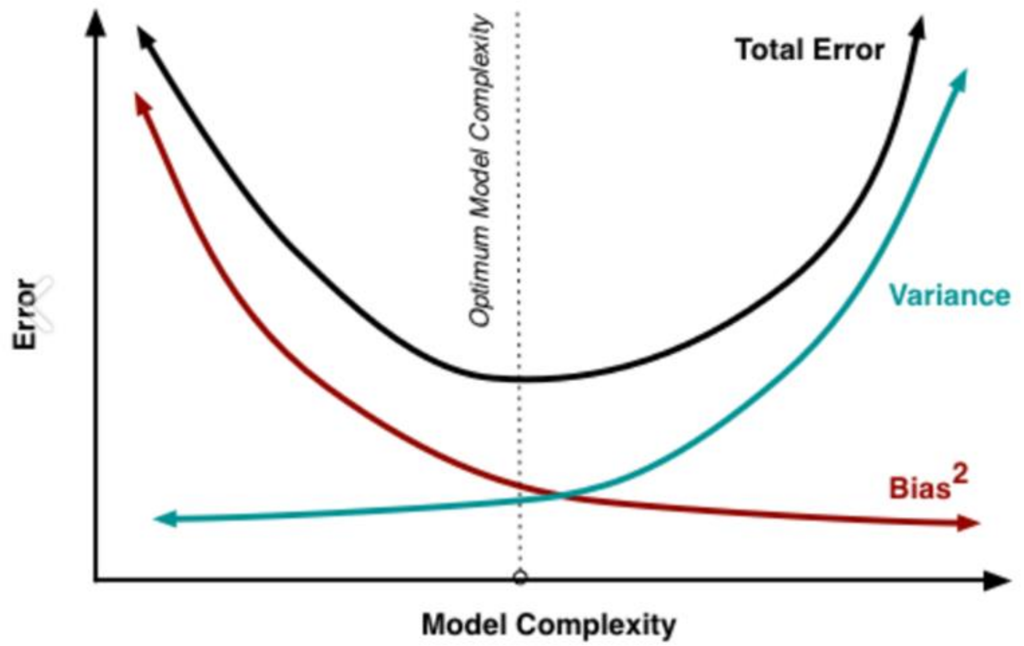
$$MSE = \frac{1}{N}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$$

Good Predictive Model

$$\text{Error}(X) = \underline{\text{Noise}(X)} + \underline{\text{Bias}(X)} + \underline{\text{Variance}(X)}$$

Data Preprocessing　　Model Complexity↑　　Model Complexity↓

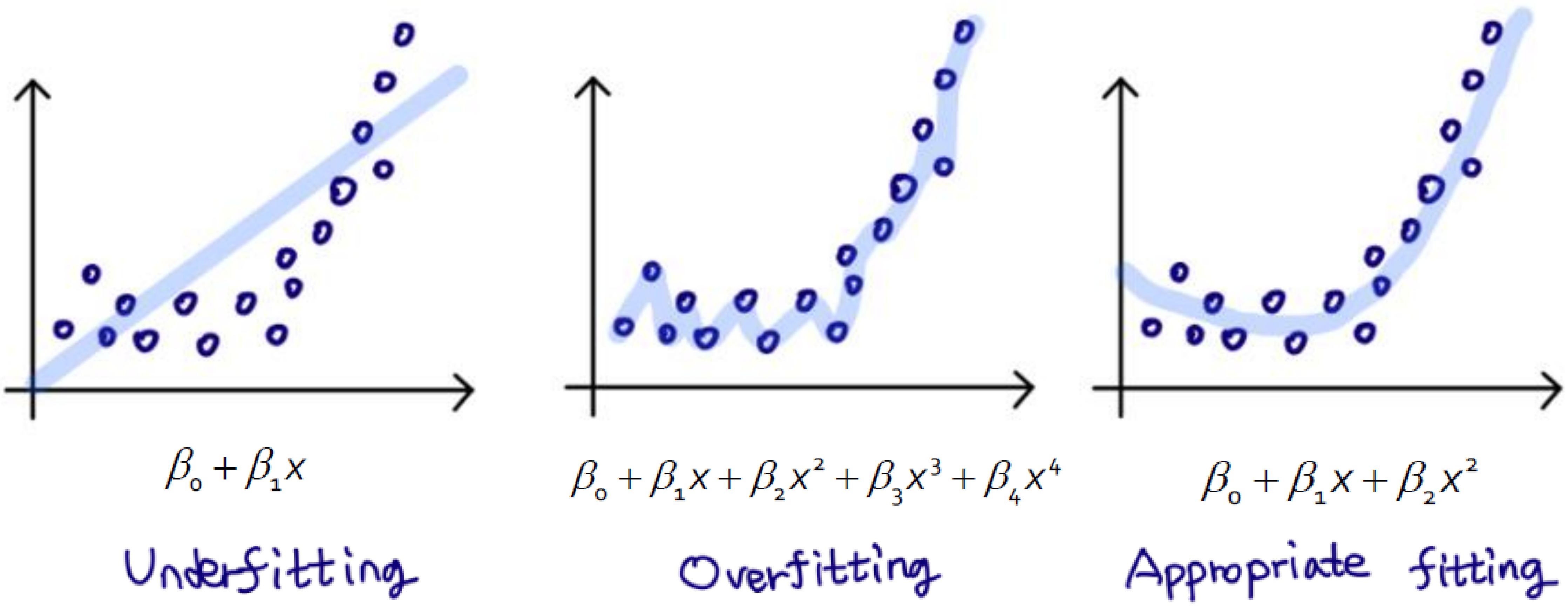
- Bias(X) ➔ Overfitting을 하면 줄어듬
  - 기존 통계학에서는 bias=0 (unbiased)을 중요하게 여김
  - Expected MSE를 줄이려면 variance도 낮춰야함
- Variance(X) ➔ Overfitting을 하면 안됨 ➔ Bias의 희생이 필요함
  - 현대 통계학에서는 variance를 줄이기 위해 bias를 소폭 희생하는 방법론을 연구함
  - 특히, <u>bias를 증가시키더라도 variance 감소폭이 더 크다면 expected MSE는 감소하게 됨</u> (예측 성능 증가)



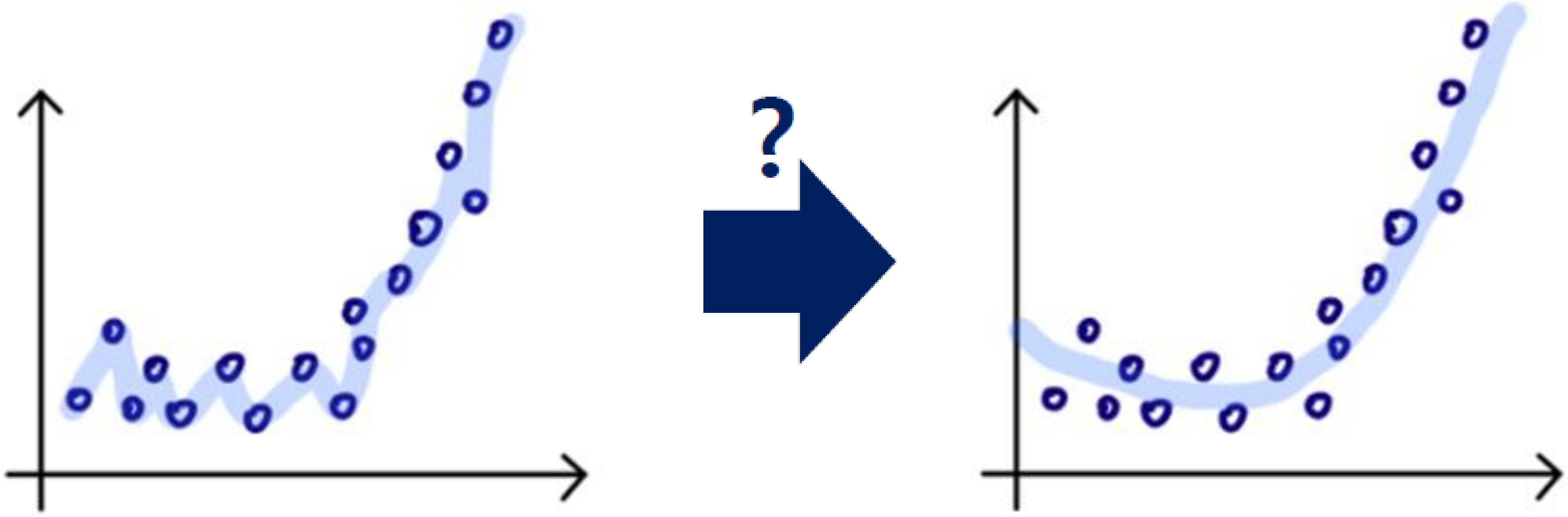
How to Reduce Variance

키와 몸무게 → 변수간 상관관계가 크면 계수가 분산 됨

Regularization Concept



Regularization Concept



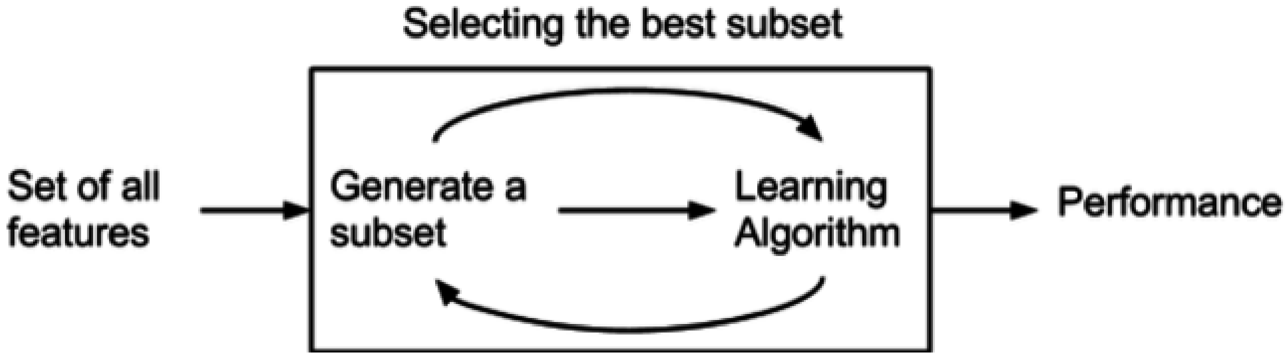
$\beta_0 + \beta_1 X + \beta_2 X^2 + \cancel{\beta_3} X^3 + \cancel{\beta_4} X^4$

$\beta_0 + \beta_1 X + \beta_2 X^2$

Feature Subset Selection

- Subset Selection method
  - 전체 p개의 설명변수(X) 중 일부 k개만을 사용하여 회귀 계수 beta를 추정하는 방법

- 전체 변수 중 일부만을 선택함에 따라 bias가 증가할 수 있지만 variance가 감소함
  - 하지만 변수의 개수가 커지면 커질 수록 경우의 수가 매우 커짐
  - 따라서, 변수의 개수가 적을 때 추천함



Best subset selection

Forward stepwise selection

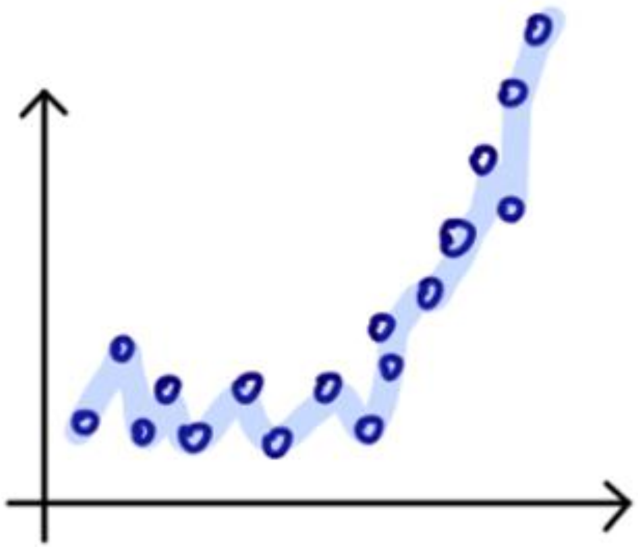
Backward stepwise elimination

Least angle regression

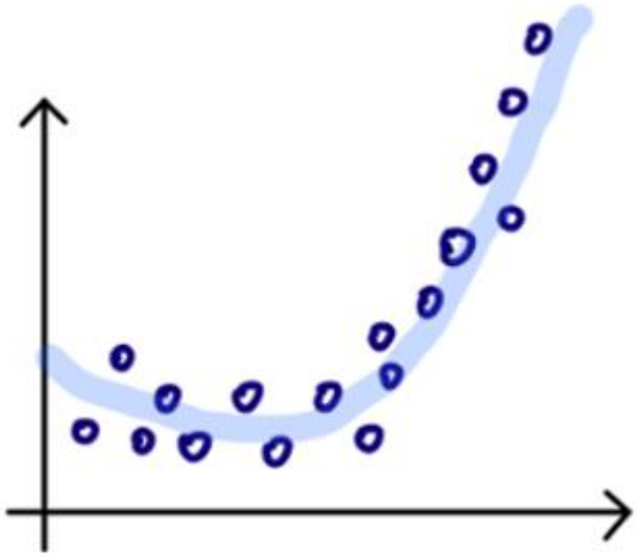
Orthogonal matching pursuit

Embedded Method Feature Selection

- Embedded Regularization Method는 회귀 계수 Beta가 가질 수 있는 값에 **제약조건**을 부여하는 방법
- 제약조건에 의해 bias가 증가할 수 있지만 variance가 감소함



$\beta_0 + \beta_1 x + \beta_2 x^2 + \cancel{\beta_3 x^3} + \cancel{\beta_4 x^4}$



$\beta_0 + \beta_1 x + \beta_2 x^2$

$$\min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + 5000\beta_3^2 + 5000\beta_4^2$$

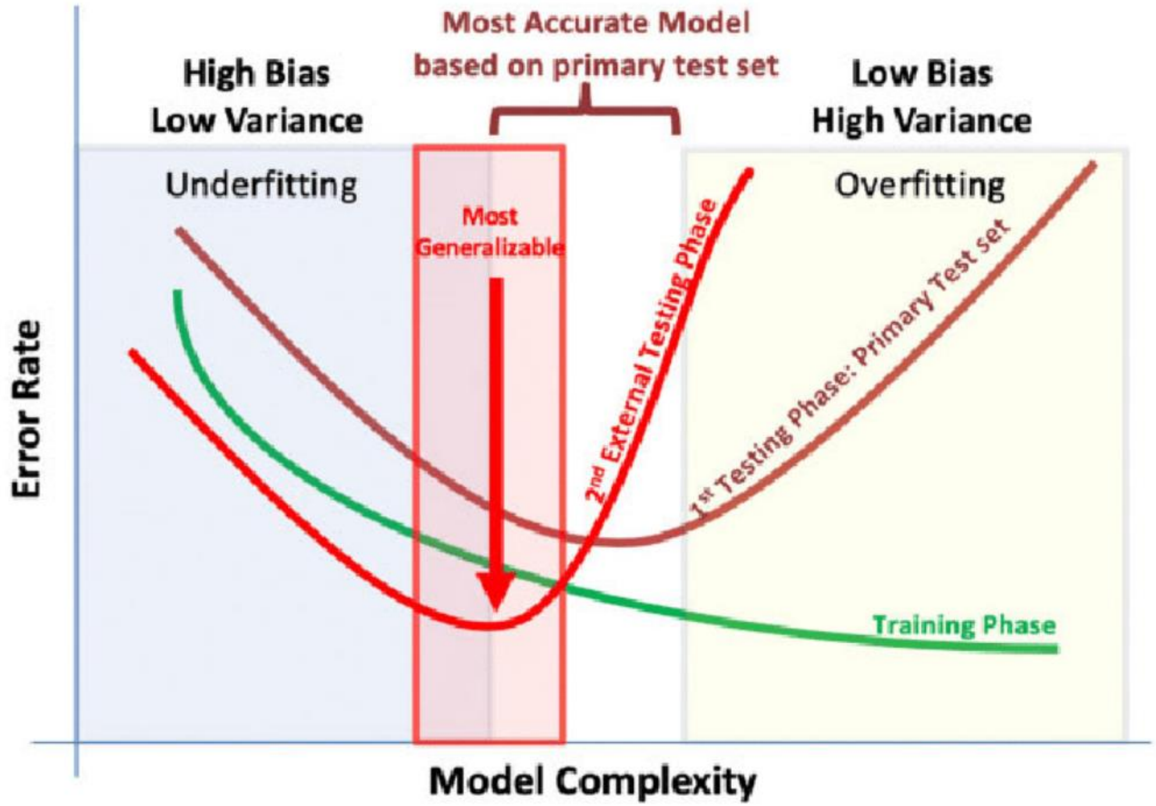
$$\beta_3 \approx 0 \qquad \beta_4 \approx 0$$

Overfitting vs Underfitting

- Train Data vs Test Data
  - Bias와 Variance의 Trade-Off 관계

$$\text{Error}(X) = \underline{\text{Noise}(X)} + \underline{\text{Bias}(X)} + \underline{\text{Variance}(X)}$$

Data Preprocessing　　Model Complexity ↑　　Model Complexity ↓



Embedded Method Feature Selection

- Embedded Regularization Method는 회귀 계수 Beta가 가질 수 있는 값에 **제약조건**을 부여하는 방법
- **Scaling 필수 !!!**

$$\beta_1, \beta_2, \ldots, \beta_p$$

$$L(\beta) = \min_{\beta} \underbrace{\sum_{i=1} (y_i - \hat{y}_i)^2}_{\text{(1) Training accuracy}} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{\text{(2) Generalization accuracy}}$$

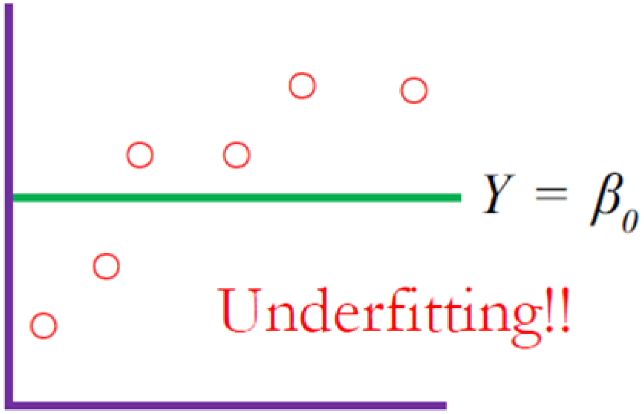
λ: regularization parameter that controls the tradeoff between two objectives

Embedded Method Feature Selection

Hyperparameter

$$C(\beta) = \min_{\beta} \sum_{i=1} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ very big $\rightarrow$ $\beta_1 \approx 0,\ \beta_2 \approx 0, \beta_3 \approx 0, \beta_4 \approx 0$



$$\beta_0 + \cancel{\beta_1 X} + \cancel{\beta_2 X^2} + \cancel{\beta_3 X^3} + \cancel{\beta_4 X^4}$$

λ very small $\rightarrow$

$$\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$$

Embedded Method Feature Selection

(β_1, β_2)	$\beta_1^2+\beta_2^2$	MSE
(4,5)	41	20
(3,5)	34	23
(4,4)	32	25
(2,5)	27	27
(2,4)	18	25
(2,3)	13	29

Embedded Method Feature Selection

$\beta_1^2 + \beta_2^2 \leqslant 30$

(β_1, β_2)	$\beta_1^2+\beta_2^2$	MSE
(4,5)	41	20
(3,5)	34	23
(4,4)	32	25
(2,5)	27	27
(2,4)	18	25
(2,3)	13	29

Ridge Regression

- $L_2 - norm$ Regularization
  - 제곱 오차를 최소화하면서 회귀 계수 beta $L_2 - norm$ 을 제한함

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

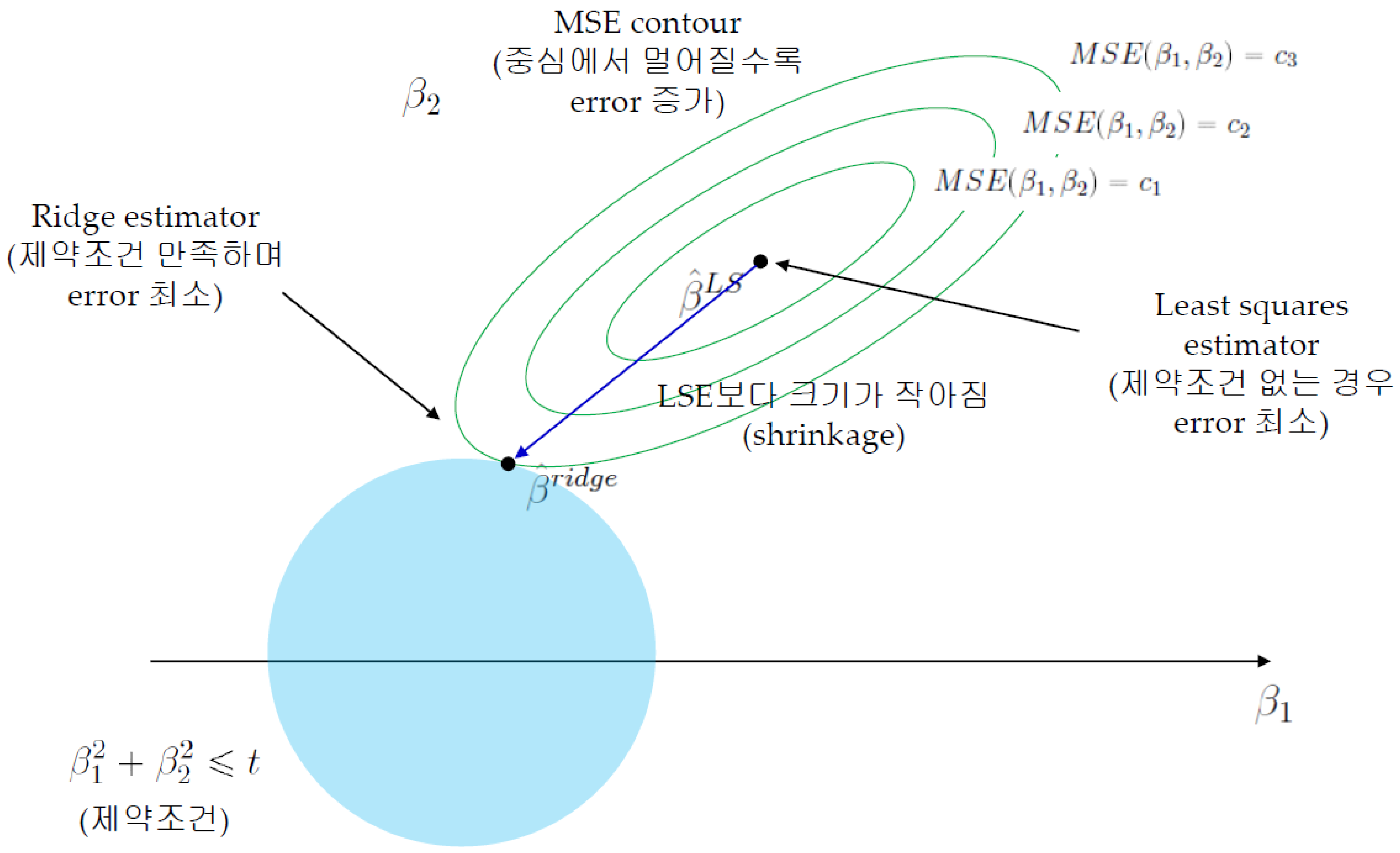
$$subject\ \ to\ \ \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

$\Updownarrow$ Equivalent (Lagrangian multiplier)

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

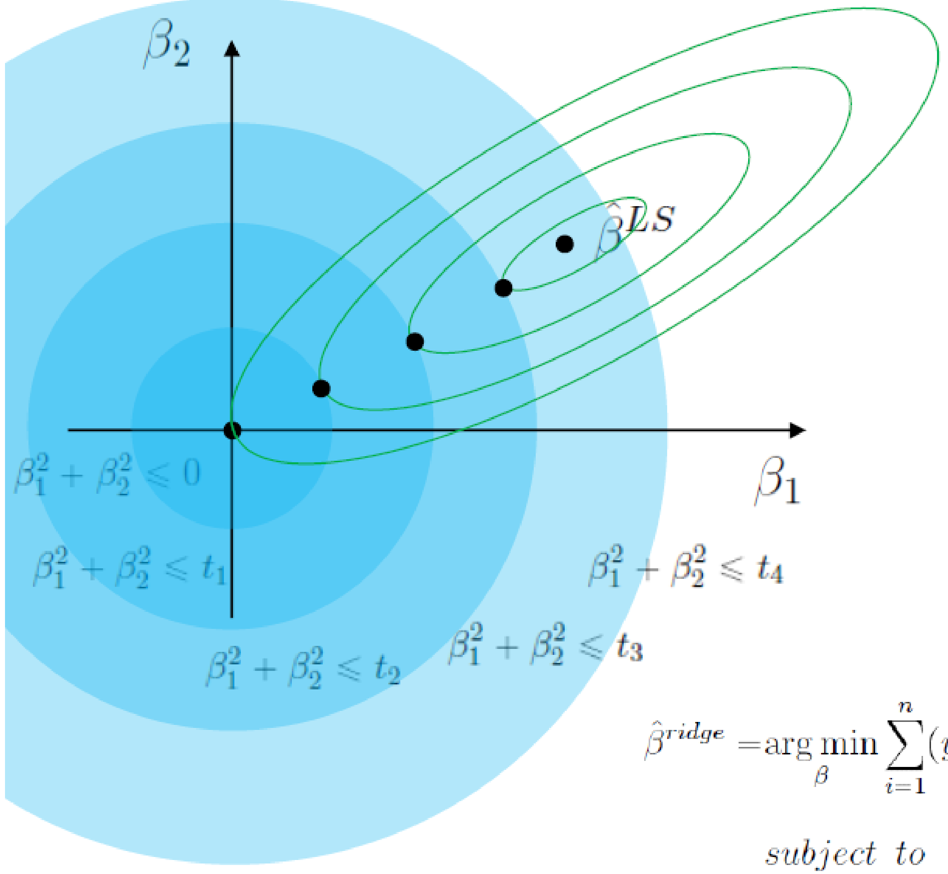
Ridge Regression

- L_2-norm Regularization
  - 제곱 오차를 최소화하면서 회귀 계수 beta L_2-norm 을 제한함



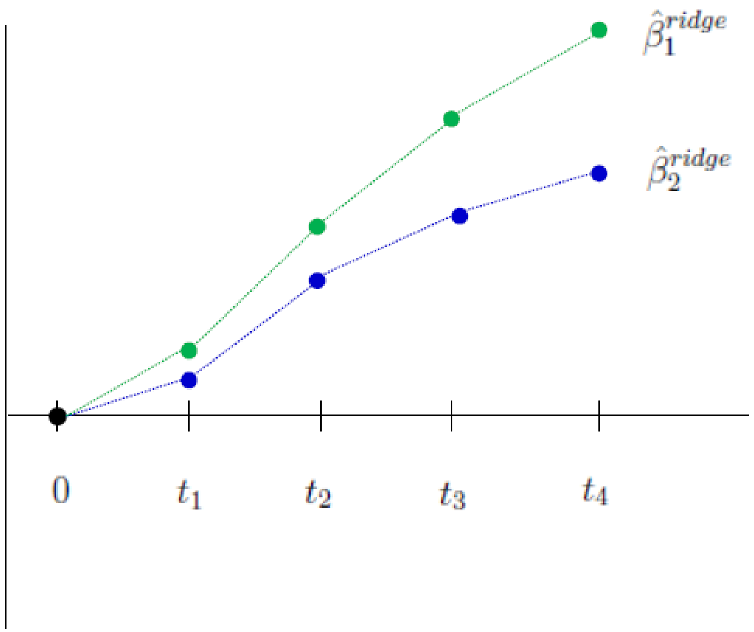
Ridge Regression

Solution path: tuning parameter 값에 따른 $\hat{\beta}^{ridge}$ 의 변화



$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

$$subject\ to\ \sum_{j=1}^{p} \beta_j^2 \leqslant t$$



Ridge Regression

Ridge는 행렬 연산을 통해 closed form solution을 구할 수 있음

$$
\begin{aligned}
Q(\beta) &= (y - X\beta)^T (y - X\beta) + \lambda\beta^T\beta \\
&= y^T y - 2\beta^T X^T y + \beta^T (X^T X + \lambda I_p)\beta
\end{aligned}
$$

$$\frac{\partial}{\partial\beta} Q(\beta) = 2X^T y + 2(X^T X + \lambda I_p)\beta = 0$$

$$\hat{\beta}^{ridge} = (X^T X + \lambda I_p)^{-1} X^T y$$

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T y$$

LASSO

- LASSO : Least Absolute Shrinkage and Selection Operator
  - L_1-norm Regularization : 회귀 계수 beta의 L_1-norm을 제한

$$\hat{\beta}^{lasso} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

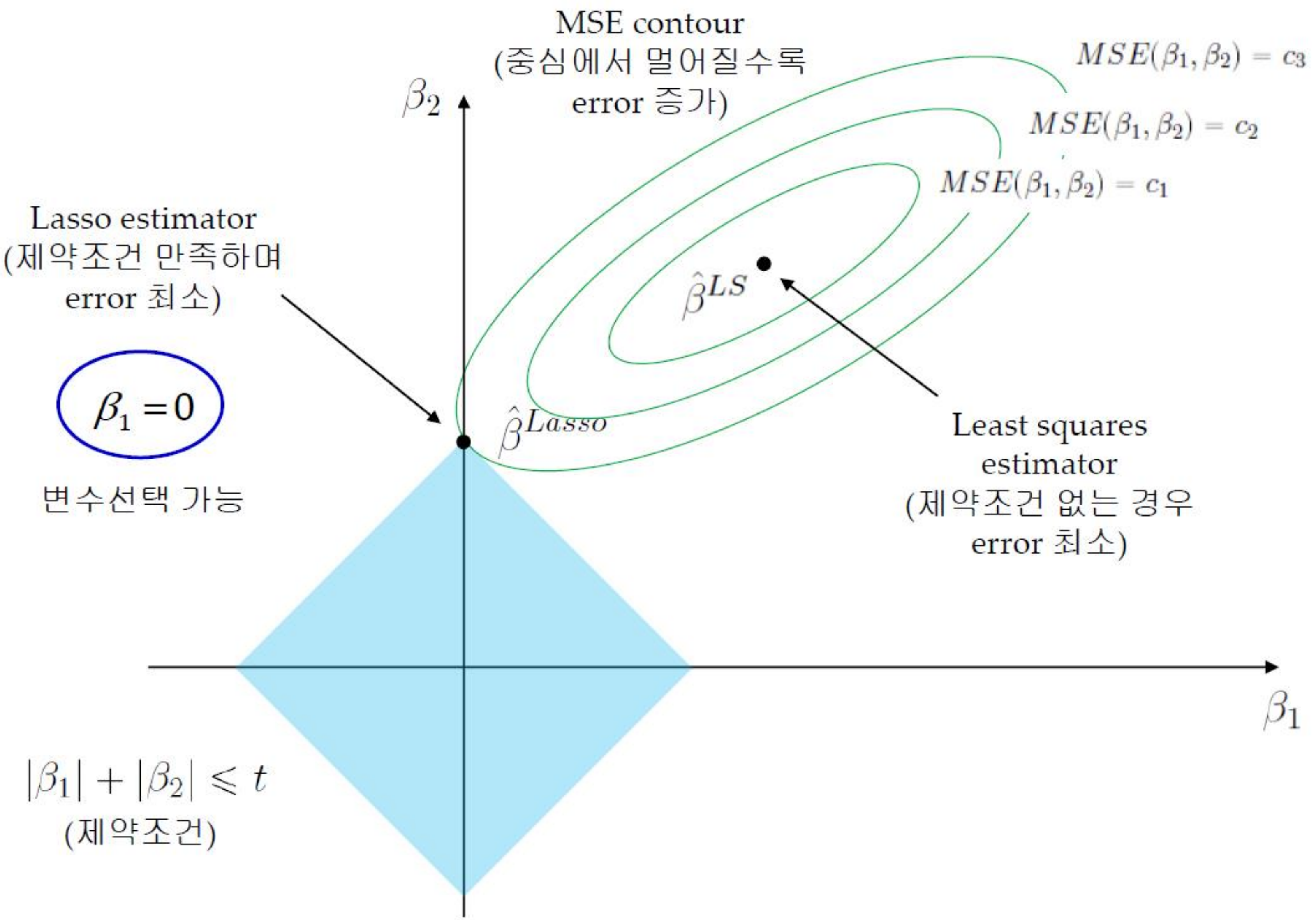
$$subject\ \ to \quad \sum_{j=1}^{p} |\beta_j| \leqslant t$$

$\Updownarrow$ Equivalent

$$\hat{\beta}^{lasso} = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

LASSO

- LASSO : Least Absolute Shrinkage and Selection Operator
  - L_1-norm Regularization : 회귀 계수 beta의 L_1-norm을 제한



LASSO

- LASSO : Least Absolute Shrinkage and Selection Operator
  - $L_1 - norm$ Regularization : 회귀 계수 beta의 $L_1 - norm$을 제한

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \Rightarrow \hat{\beta}^{ridge} = (X^T X + \lambda I_p)^{-1} X^T y$$

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \Rightarrow \hat{\beta}^{Lasso} = ?$$

- Ridge와 달리 Lasso formulation은 closed form solution을 구하는 것이 불가능 (L_1 norm 미분 불가능)

- Numerical optimization methods:
  - Quadratic programming techniques (1996, Tibshirani)
  - LARS algorithm (2004, Efron et al.)
  - Coordinate descent algorithm (2007, Friedman et al.)

LASSO

λ 값을 어떻게 설정할 것인가?

몇 개의 변수를 선택할 것인가?

LASSO

큰 λ 값

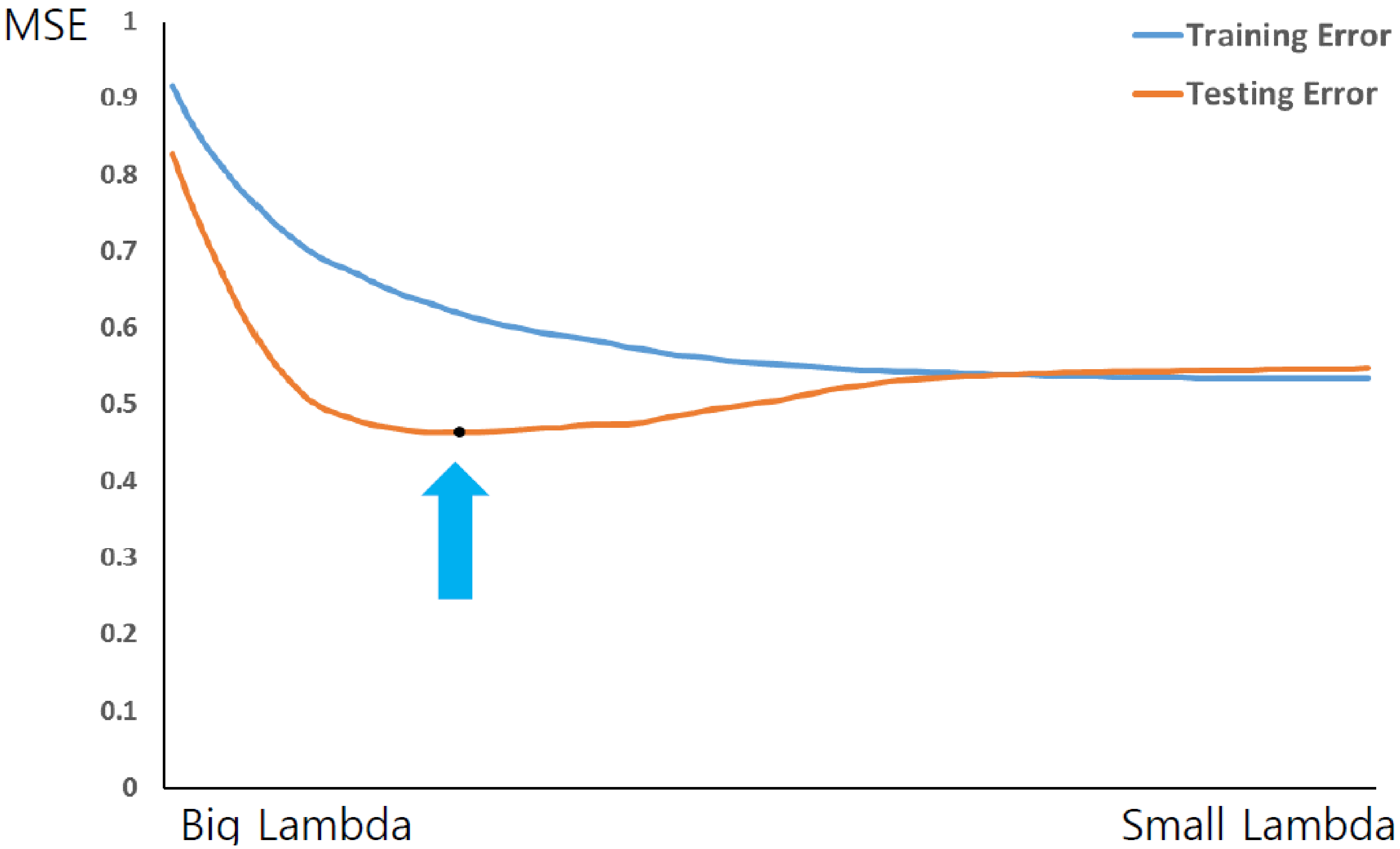
적은 변수
간단한 모델
해석 쉬움
높은 학습 오차
(Underfitting의 위험 증가)

작은 λ 값

많은 변수
복잡한 모델
해석 어려움
낮은 학습 오차

LASSO

- 일정 범위 내로 λ를 조정하여, 가장 좋은 예측 결과를 보이는 λ 값을 선정함

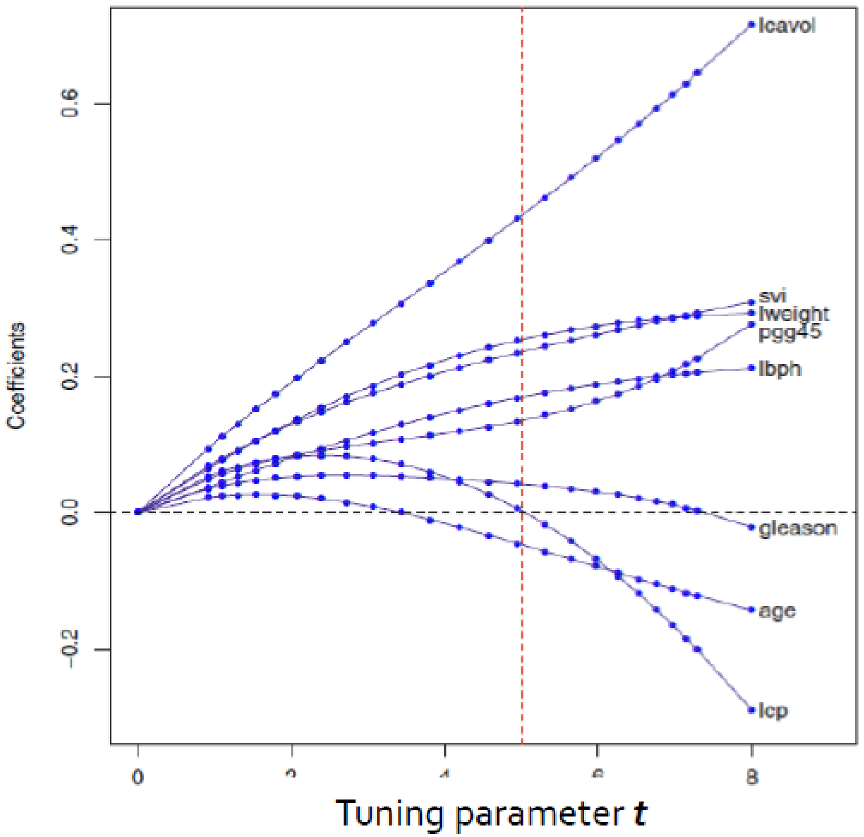


Solution Paths Ridge and LASSO

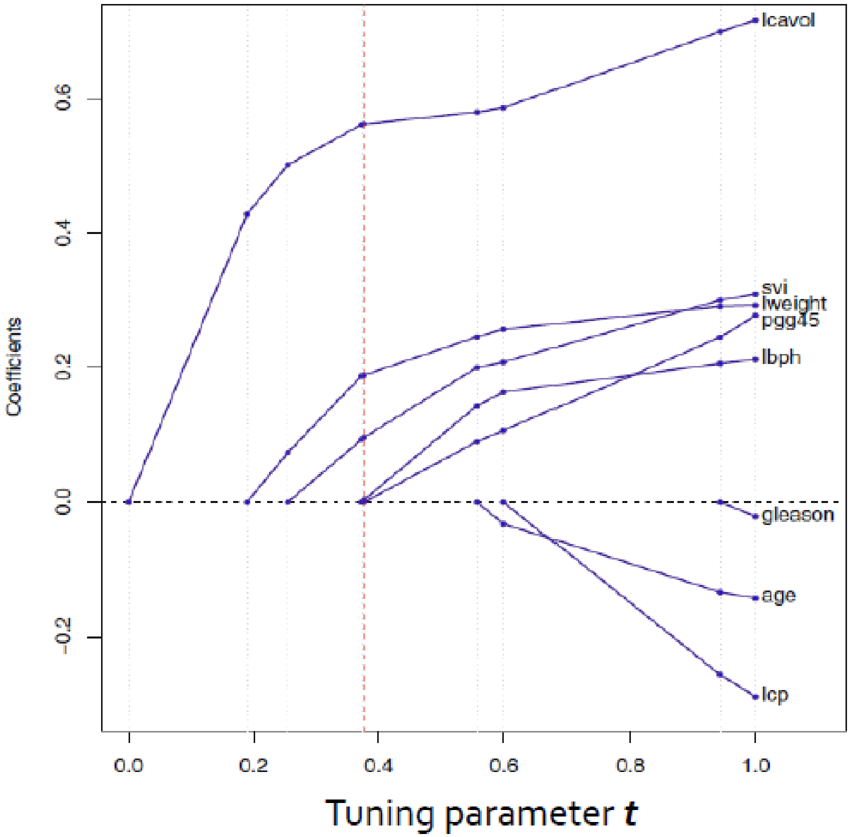
- Ridge와 LASSO 모두 t가 작아짐에 따라 모든 계수의 크기가 감소함
- Ridge : 크기가 큰 변수가 더 빠르게 감소하는 경향을 보임
- LASSO : 예측에 중요하지 않은 변수가 더 빠르게 감소, t가 작아짐에 따라 예측에 중요하지 않은 변수가 0이 됨

Prostate cancer data (Y: 전립선 암 항체, X: 환자 의료 데이터)

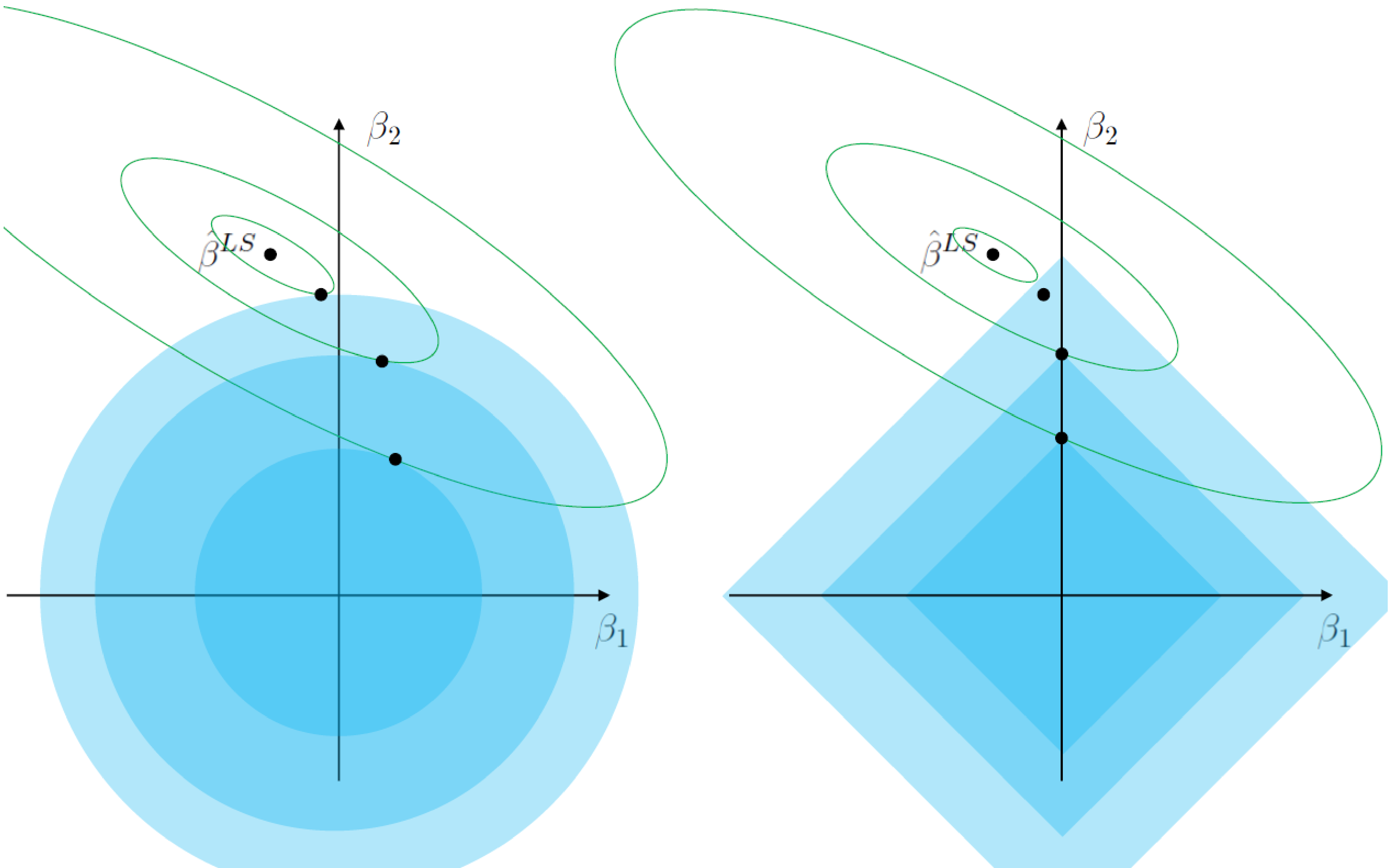
Ridge



Lasso



Solution Paths Ridge and LASSO



Ridge vs LASSO

Ridge	Lasso
L_2 norm regularization	L_1 norm regularization
변수 선택 불가능	변수 선택 가능
Closed form solution 존재 (미분으로 구함)	Closed form solution이 존재하지 않음 (numerical optimization 이용)
변수 간 상관관계가 높은 상황 (collinearity)에서 좋은 예측 성능	변수 간 상관관계가 높은 상황에서 ridge 에 비해 상대적으로 예측 성능이 떨어짐
크기가 큰 변수를 우선적으로 줄이는 경향 이 있음	

Ridge vs LASSO

- 변수들 간 상관관계가 큰 경우
  - 변수 선택 성능 저하
  - 예측 성능 저하

→ 변수 간 상관관계를 반영할 수 있는 방법 필요

그래서 머선 알고리즘을 사용해야 되는 거야?

ElasticNet

- ElasticNet
  - Ridge + LASSO (L1 and L2 Regularization Term)

- ElasticNet은 Correlation이 큰 변수를 동시에 선택/배제하는 특성을 가지고 있음

$$\hat{\beta}^{enet} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} (y_i - x_i\beta)^2$$

$$subject\ \ to\ \ s_1 \sum_{j=1}^{p} |\beta_j| + s_2 \sum_{j=1}^{p} \beta_j^2 \leqslant t$$

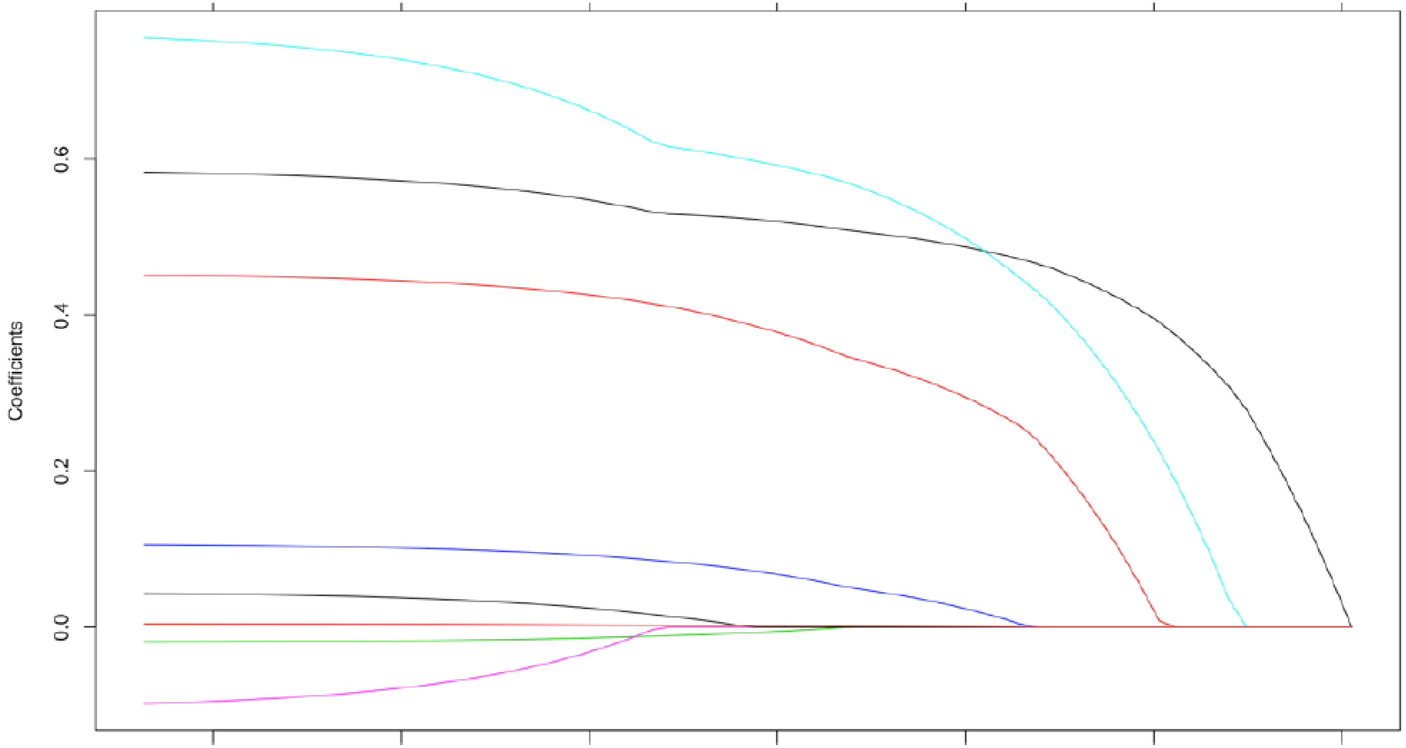
$\Updownarrow$ Equivalent

$$\hat{\beta}^{enet} = \underset{\beta}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\}$$

ElasticNet

- ElasticNet
  - Ridge + LASSO (L1 and L2 Regularization Term)

- ElasticNet은 Correlation이 큰 변수를 동시에 선택/배제하는 특성을 가지고 있음



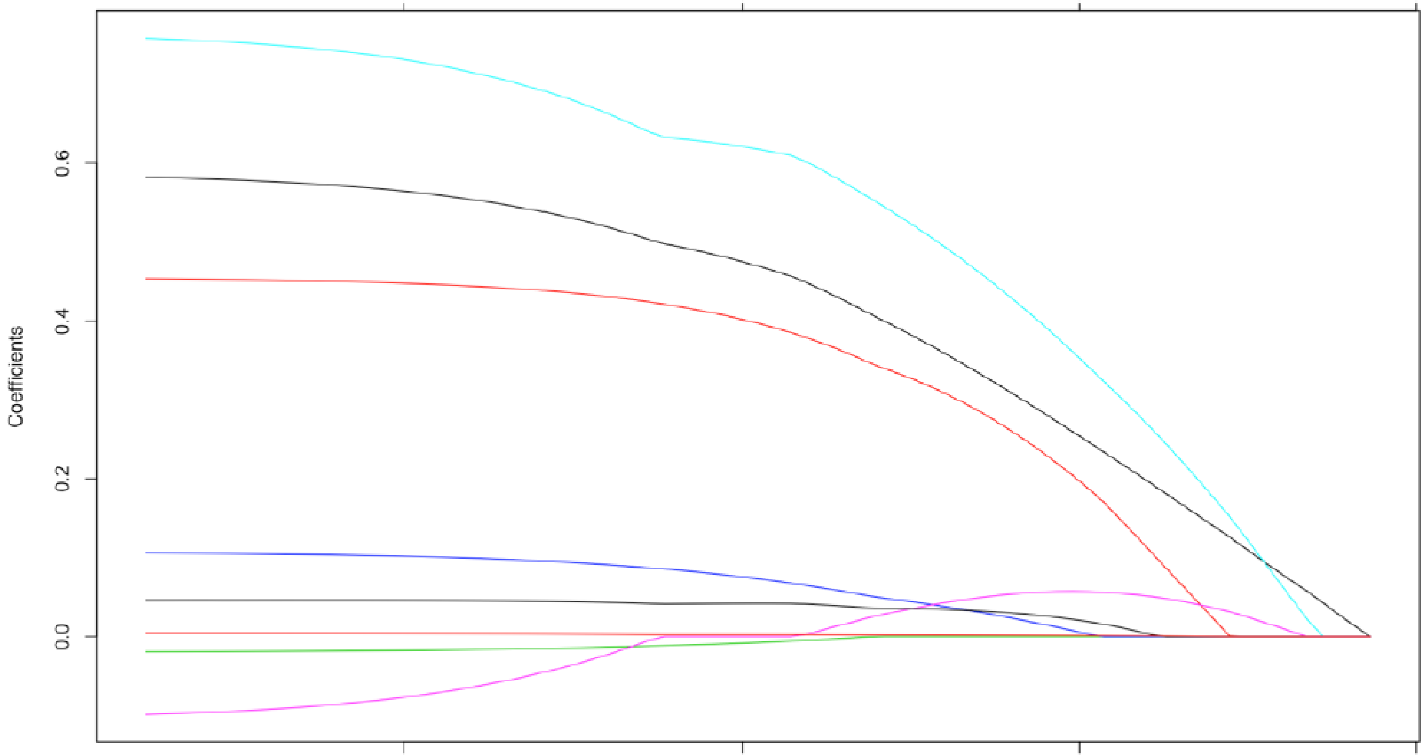
λ_1 값 증가

변수 개수 감소

ElasticNet

- ElasticNet
  - Ridge + LASSO (L1 and L2 Regularization Term)

- ElasticNet은 Correlation이 큰 변수를 동시에 선택/배제하는 특성을 가지고 있음

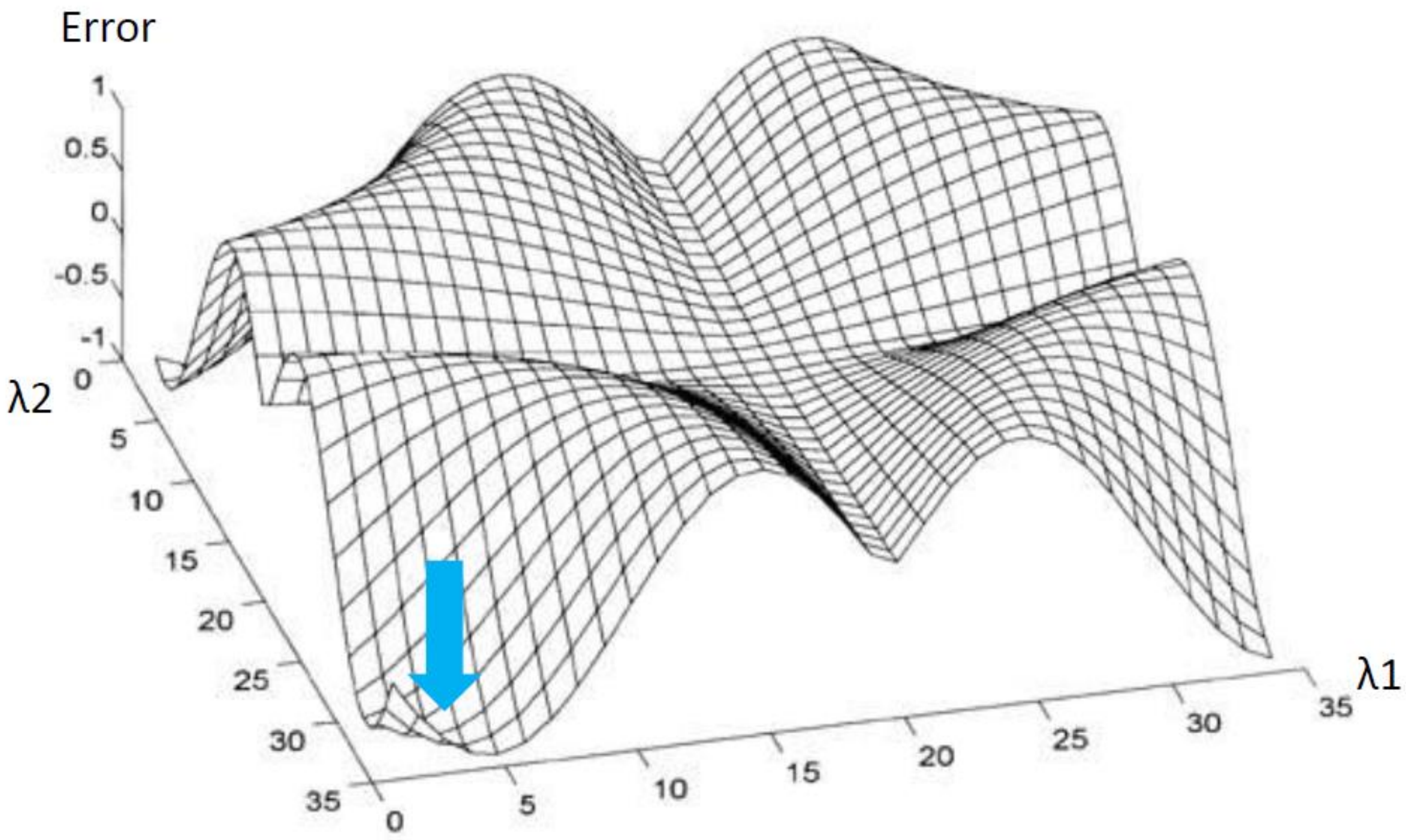


λ_2 값 증가

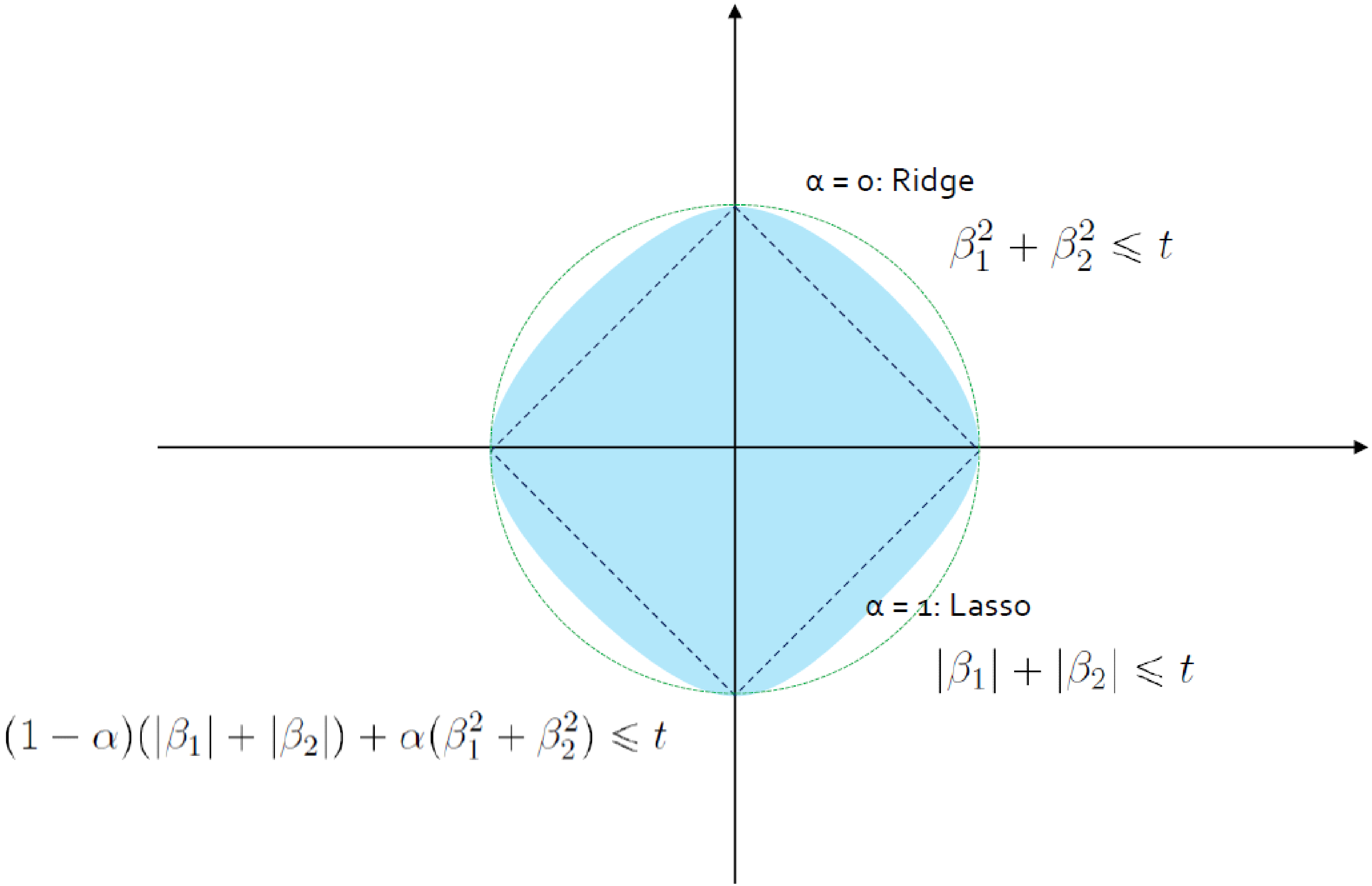
변수 개수 감소효과 거의 없음

ElasticNet Parameters

- 일정 범위 내로 λ1, λ2를 조정하여, 가장 좋은 예측 결과를 보이는 λ1, λ2 값을 선정함

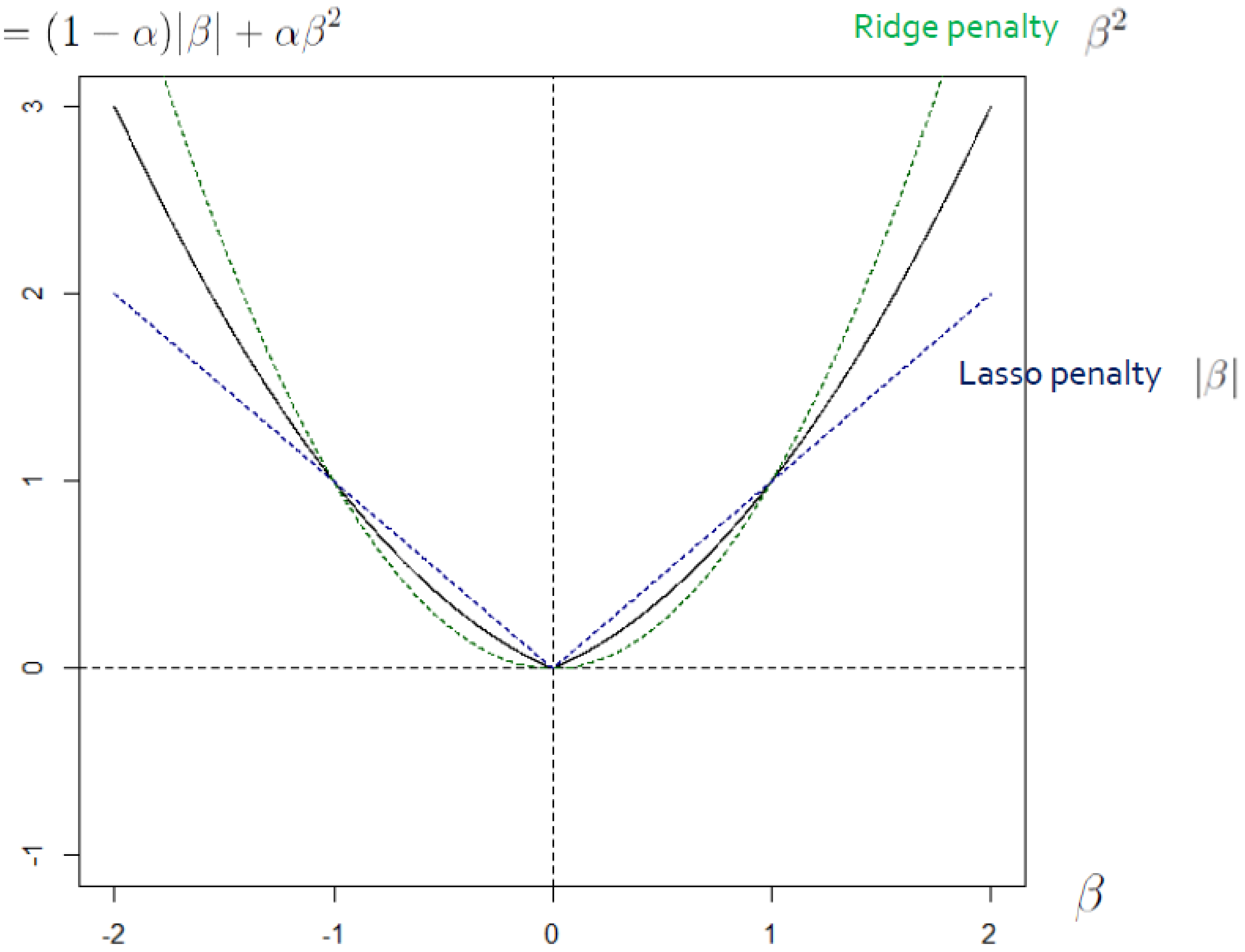


ElasticNet Parameters



ElasticNet Parameters

$$Penalty(\beta) = (1-\alpha)|\beta| + \alpha\beta^2$$



Others

Prior Knowledge	Regularization Method
상관관계 높은 변수들 동시에 선택	Elastic Net
인접한 변수들 동시에 선택	Fused Lasso
사용자가 정의한 그룹 단위로 변수 선택	Group Lasso
사용자가 정의한 그래프의 연결 관계에 따라 변수 선택	Grace

Q & A